AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2023, Vol. 152, No. 11, 3037–3057
https://doi.org/10.1037/xge0001438

Decision Criteria in Signal Detection Model Are Not Based on the Objective Likelihood Ratio

Xiao Hu[1], Chunliang Yang[1, 2], and Liang Luo[2, 3]

[1] Beijing Key Laboratory of Applied Experimental Psychology, National Demonstration Center for Experimental Psychology Education, Faculty of Psychology, Beijing Normal University

[2] Institute of Developmental Psychology, Faculty of Psychology, Beijing Normal University

[3] State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University

How people set decision criteria in signal detection model is an important research question. The likelihood ratio (LR) theory, which is one of the most influential theories about criteria setting, typically assumes that (a) decisions are based on the objective LR of the signal and noise distributions, and (b) LR criteria do not change across tasks with various difficulty levels. However, it is often questioned whether people are really able to know the exact shape of signal and noise distributions, and compute the objective LR accordingly. Here we suggest whether decision criteria are set based on objective LR can be tested in two-condition experiments with different difficulty levels across conditions. We then asked participants in three empirical experiments to perform two-condition perceptual or memory tasks, and give their answer using confidence rating scale. Results revealed that the two assumptions of LR theory contradicted with each other: if we assumed decision criteria were based on objective LR, then the estimated LR criteria differed across difficulty levels, and fanned out as task difficulty decreased. We suggest people might inaccurately estimate the LR in signal detection tasks, and several possible explanations for the distortion of LR are discussed.

Public Significance Statement

This study demonstrates that the three regularities, including the mirror effect, variance effect and *z*-transformed receiver operating characteristic length effect, hold even when the objective likelihood ratio corresponding to the decision criteria differs between experimental conditions. This study indicates that the estimated decision criteria on the axis of objective likelihood ratio differ across difficulty levels, and fan out as task difficulty decreases. This study suggests that people might inaccurately estimate the likelihood ratio in signal detection tasks.

Keywords: signal detection theory, decision criteria, likelihood ratio, evidence strength, three regularities

Supplemental materials: https://doi.org/10.1037/xge0001438.supp

Signal detection theory (SDT) is one of the most widely applied computational theories in experimental psychology and has been used to model people's decision process in tasks across various cognitive domains such as perception, memory, action and reasoning (for reviews, see Maloney & Zhang, 2010; Wixted, 2020). In a signal detection task, participants are presented with a series of stimuli which belong to either signal or noise category, and need to decide whether each stimulus is signal or noise. According to SDT, there are four possible outcomes in the task, including hit (when a stimulus is signal and participants respond "signal"), miss (when a stimulus is signal and participants respond "noise"), false alarm (when a stimulus is noise and participants respond "signal"), and correct rejection (when a stimulus is noise and participants respond "noise"; Wickens, 2002).

One important theoretical question in the framework of SDT is how people make decisions in signal detection tasks. The basic assumption of SDT is that people's decisions rely on the comparison between the value of a decision variable for each stimulus and a criterion on the decision axis (Green & Swets, 1966; Wickens, 2002).

This article was published Online First June 1, 2023.

This study was supported by the Natural Science Foundation of China (32200841, 32171045, 32000742), China Postdoctoral Science Foundation (2022M720485), and the Fundamental Research Funds for the Central Universities (2022NTSS36).

We have no conflicts of interest to disclose.

Data and scripts for the current study are available online at OSF (https://osf.io/grcsj/).

Xiao Hu served as lead for formal analysis, investigation, methodology, visualization, and writing–original draft, contributed equally to conceptualization, and served in a supporting role for funding acquisition and writing–review and editing. Chunliang Yang served in a supporting role for funding acquisition and writing–review and editing. Liang Luo served as lead for conceptualization, funding acquisition, supervision, and writing–review and editing and served in a supporting role for methodology.

Correspondence concerning this article should be addressed to Liang Luo, Institute of Developmental Psychology and State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China. Email: luoliang@bnu.edu.cn

However, since the early days when SDT was introduced to the field of experimental psychology, there has been no consensus in the literature regarding whether the decision variable directly represents the evidence strength (such as physical intensity in perceptual decision tasks, or memory strength in recognition memory tests) associated with each stimulus or reflects an internal quantity that is computed based on the evidence strength but not the evidence strength per se (Balakrishnan & Ratcliff, 1996; Glanzer & Adams, 1985; Glanzer et al., 2009, 2019; Hintzman, 1994; Semmler et al., 2018; Stretch & Wixted, 1998a). Investigating the basis of the decision variable in SDT, including both its mathematical form and the underlying cognitive processes, can greatly foster our understanding of how people discriminate stimuli from different categories in tasks with uncertainty (Lynn et al., 2015).

Researchers have proposed two different theories concerning the decision process in the SDT framework: the strength theory suggests that the identification of signal and noise is directly based on evidence strength of each stimulus (Balakrishnan & Ratcliff, 1996), while the likelihood ratio (LR) theory indicates that people first compute the ratio of the likelihood that each stimulus belongs to signal or noise category given the evidence strength, which then forms the basis of the decision process (Glanzer et al., 2009, 2019; Semmler et al., 2018). Recently, Glanzer et al. (2009, 2019) propose three regularities derived from the LR theory, including the mirror effect, variance effect, and z-transformed receiver operating characteristic *z*ROC length effect (see below for details). The three regularities have been obtained across a wide range of previous studies, suggesting that people's decisions in signal detection tasks may be based on LR (Glanzer et al., 2009, 2019; Hilford et al., 2015, 2019). However, other researchers have questioned whether people are really able to compute objective LR based on the true distributions of signal and noise (Balakrishnan & Ratcliff, 1996; Criss & McClelland, 2006).

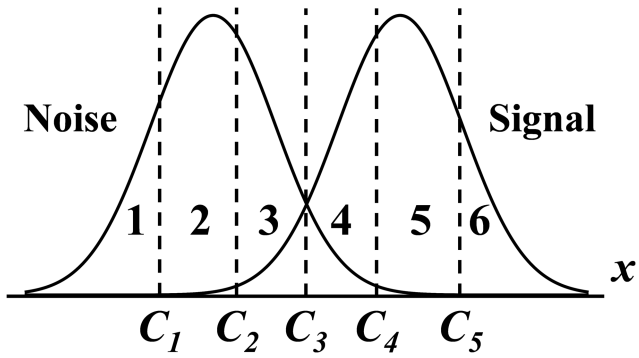
Below we first introduce the basic ideas of the strength theory and LR theory. Next, we describe the three regularities based on the LR theory. We then discuss the question about the LR theory concerning the ability to compute objective LR. Finally, we introduce the rationale of the current study.

Two Theories About Setting Decision Criteria

Researchers have proposed two different theories to explain how people set decision criteria in signal detection tasks. The first one is the strength theory, which assumes that decisions are directly based on the comparison between evidence strength of each stimulus and a criterion C on the strength axis. A "signal" response will occur when evidence strength exceeds C, and a "noise" response will be made when evidence strength is lower than C (e.g., Balakrishnan & Ratcliff, 1996; Bruno et al., 2009; Verde & Rotello, 2007). Sometimes participants are asked to not only identify whether each stimulus is signal or noise, but also to rate confidence about their identification. For example, people may give their response on a scale from 1 to 6 with 1 for *sure that the stimulus is noise* and 6 for *sure that the stimulus is signal* (e.g., Hilford et al., 2015; Stretch & Wixted, 1998a). The strength theory suggests that response on an n-point confidence rating scale can be obtained by setting $(n-1)$ criteria on the axis of evidence strength (Balakrishnan & Ratcliff, 1996; Wickens, 2002). People give a rating of 1 when stimulus strength is lower than the first criterion, 2 when stimulus strength is between the first and second criteria, and so on (see Figure 1).

Figure 1

Illustration of SDT With Multiple Decision Criteria on the Axis of Evidence Strength x



Note. A 6-point confidence rating scale is used as an example. SDT= signal detection theory.

The second theory about criteria setting is the LR theory, which assumes that decisions in signal detection tasks rely on the ratio of the likelihood that each stimulus is signal or noise, which is the same as the ratio of the probability density function in signal and noise distributions (Glanzer et al., 2009, 2019; Osth et al., 2017; Semmler et al., 2018). The LR theory is a special case of the Bayesian decision theory (BDT) applied to signal detection tasks, which suggests that people make decisions according to the ratio of the posterior probability that each stimulus belongs to signal S or noise N (Green & Swets, 1966; Maloney & Zhang, 2010; Wickens, 2002):

$$\frac{P(S|x)}{P(N|x)} = \frac{f(x|S)P(S)}{f(x|N)P(N)}.$$

In the equation above, $P(S)$ and $P(N)$ represent the prior probability for signal and noise, respectively. $f(x|S)$ and $f(x|N)$ refer to the likelihood (i.e., probability density function) of evidence strength x in signal and noise distributions. Furthermore, $P(S|x)$ and $P(N|x)$ represent the posterior probability. BDT assumes that people's decisions are based on the maximum-a-posteriori rule (Burgess, 1985). That is, people classify the current stimulus as signal when the posterior probability is higher for signal than for noise and judge it as noise when the posterior probability for signal is lower. This is mathematically equivalent to setting a criterion β on the ratio of likelihood, and people should judge the stimulus as signal when LR is higher than β (and vice versa):

$$\text{Respond "signal"} \Leftrightarrow \frac{f(x|S)}{f(x|N)} > \beta = \frac{P(N)}{P(S)}.$$

If we take the logarithm of both sides of the equation above, we can get the mathematically equivalent rule that people set a criterion on the log LR (denoted by λ) which can be seen as a function of the evidence strength x:

$$\text{Respond "signal"} \Leftrightarrow \lambda(x) = \log\frac{f(x|S)}{f(x|N)} > \log\beta = \log\frac{P(N)}{P(S)}.$$

The decision principle above has not considered the gains and losses associated with people's response. When the expected reward

(or penalty) is different between signal and noise responses, the LR decision criteria β is affected by both prior probabilities and payoffs (Green & Swets, 1966):

$$\beta = \frac{[G(a = N, w = N) + G(a = S, w = N)]P(N)}{[G(a = N, w = S) + G(a = S, w = S)]P(S)}.$$

in which $G(a, w)$ refers to the gain function, and represents the gains or losses given participants' response a and true stimulus category w (Glanzer et al., 2019; Maloney & Zhang, 2010).

When participants respond on an n-point confidence rating scale, the LR theory suggests that they set $(n - 1)$ criteria $(\beta_1, \beta_2, \ldots, \beta_{n-1})$ on the axis of LR. Participants rate their confidence by first computing LR based on evidence strength x and then comparing LR with each of the β criteria (Glanzer et al., 2019).

It is relatively difficult to distinguish the strength theory and LR theory in a certain signal detection task, because the two theories have similar (or even the same) mathematical forms. For example, in normative equal variance SDT where both signal and noise are drawn from normal distributions with equal variance but different means, $\lambda(x)$ is a monotonic increasing function, and every criterion on log LR axis corresponds to a specific criterion on the strength axis. Thus, the two theories are mathematically equivalent in equal variance SDT, and it is impossible to completely determine whether the setting of decision criteria on the strength axis is directly based on evidence strength or based on LR (Glanzer et al., 2019). Although the two theories make slightly different predictions about people's response in unequal variance SDT (where signal and noise distributions have different variances), this difference is very subtle and noisy in a small sample of empirical data, and thus hard to detect (Balakrishnan & Ratcliff, 1996; Macmillan & Creelman, 2004).[1]

Three Regularities Derived From the LR Theory

To test whether people's decisions rely on evidence strength or LR, Glanzer et al. (2009, 2019) develop the two-condition experimental design, in which participants need to perform two signal detection tasks with different levels of difficulty (easy vs. difficult; see Figure 2A). For example, participants may be asked to complete two recognition memory tests with different types of stimuli (e.g., high vs. low frequency words) or study conditions (e.g., short vs. long study time; Glanzer et al., 2009). Although it is hard to dissociate the strength theory and LR theory in either of the experimental conditions alone, Glanzer et al. suggest that if the LR theory holds, the LR criteria βs should not change across the two conditions because βs only depend on participants' estimation of prior probabilities and gain functions, which should be unaffected by task difficulty (Glanzer et al., 2019). Consequently, participants should adjust the decision criteria on the evidence strength axis across experimental conditions to maintain constant LR criteria (which is called the constant LR account; Semmler et al., 2018; see Figure 2B). Thus, if the estimated strength criteria in empirical data change across conditions in the same way as the constant LR account predicts, then the LR theory is supported.

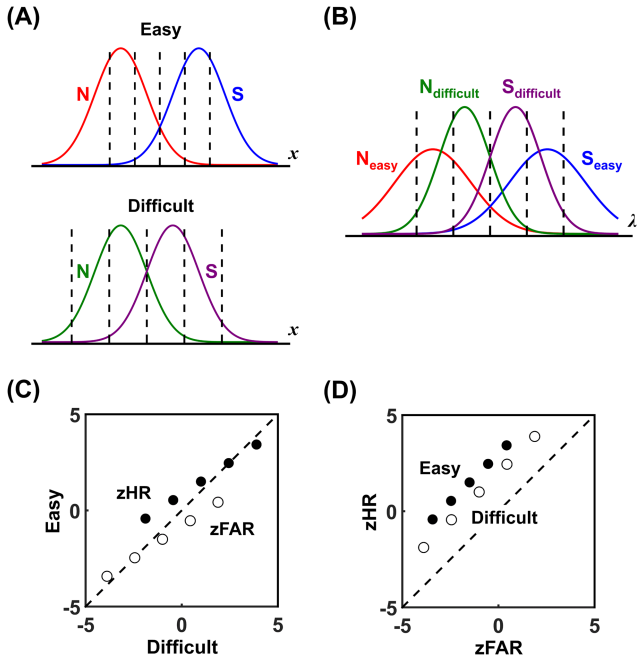
Glanzer et al. (2009, 2019) indicate that the constant LR account predicts three regularities in two-condition experiments, including the mirror effect, variance effect and zROC length effect. They suggest the constant LR account is supported only when the three regularities are obtained in empirical data. The mirror effect has been found in many previous studies (e.g., DeCarlo, 2007; Glanzer & Adams, 1985; Higham et al., 2009; Hilford et al., 2019), and suggests that hit rate (HR, the probability of responding "signal" in signal trials) and false alarm rate (FAR, the probability of responding "signal" in noise trials) in easy and difficult conditions satisfy the following relationship:

$$\mathrm{FAR}_{\mathrm{easy}} < \mathrm{FAR}_{\mathrm{difficult}} < \mathrm{HR}_{\mathrm{difficult}} < \mathrm{HR}_{\mathrm{easy}}.$$

Previous studies have demonstrated that when there is a bias toward either signal or noise in the decision process, the mirror effect may be obscured, and the usual indicator of the mirror effect (i.e., the pattern of HR and FAR) may not always give an accurate picture of the order of the underlying distribution means on the axis of LR

Figure 2

Illustration of the Three Regularities Derived From the LR Theory



Note. (A) The signal (S) and noise (N) distributions in easy and difficult conditions on the axis of evidence strength x. According to the zROC length effect, evidence criteria (the dashed lines) fan out in difficult condition. (B) The signal and noise distributions replotted on the axis of log LR λ. Based on the constant LR account, decision criteria (the dashed lines) on the λ axis remain the same across conditions. (C) The $z\mathrm{HR}_{\mathrm{easy}}/z\mathrm{HR}_{\mathrm{difficult}}$ and $z\mathrm{FAR}_{\mathrm{easy}}/z\mathrm{FAR}_{\mathrm{difficult}}$ plots showing the mirror effect and variance effect. (D) The zHR/zFAR plots for easy and difficult conditions showing the zROC length effect. LR = likelihood ratio; ROC = receiver operating characteristic; HR = hit rate; FAR = false alarm rate. See the online article for the color version of this figure.

[1] The LR theory indicates that in unequal variance SDT, hit rate (HR) should always be higher than false alarm rate (FAR) when the mean of signal distribution is higher than that of noise distribution. In contrast, the strength theory suggests that HR is lower than FAR with extremely liberal or conservative decision criterion, although the difference between HR and FAR is very small in this condition (Balakrishnan & Ratcliff, 1996; Macmillan & Creelman, 2004; Wickens, 2002).

(Glanzer et al., 2009; Hilford et al., 2015). To address this issue, Hilford et al. (2015) suggest that a better measure of the mirror effect is the distance between the means of noise distributions in easy and difficult conditions, and between the means of signal distributions across the two conditions, after the signal and noise distributions on the axis of evidence strength x (see Figure 2A) are transformed to the distributions on the axis of log LR λ (see Figure 2B): the distribution mean should be higher for signal distribution in easy than difficult condition and lower for noise distribution in easy condition. Previous studies have used the zROC plot to demonstrate the distance between distribution means (e.g., Glanzer et al., 2019; Hilford et al., 2015). The receiver operating characteristic (ROC) curve is a plot of HR versus FAR across all possible decision criteria. The zROC is a variant of the ROC curve, in which both HR and FAR are transformed into z-scores (i.e., zHR and zFAR). Each probability (here denoted by P) is transformed to $z = \Phi^{-1}(P)$, in which Φ^{-1} refers to the inverse of the cumulative distribution function of standard normal distribution. In a classical zROC where zHR is plotted against zFAR, the distance between the means of signal and noise distributions can be represented by the distance index d_e (Wickens, 2002):

$$d_e = \frac{2h}{1+s}$$

in which h and s are the intercept and slope of a linear model fit to the zROC, respectively. Hilford et al. (2015) plot two novel zROCs in which $z\text{HR}_{\text{easy}}$ is plotted against $z\text{HR}_{\text{difficult}}$, and $z\text{FAR}_{\text{easy}}$ is plotted against $z\text{FAR}_{\text{difficult}}$ (see Figure 2C). If the mirror effect holds, then the distance index for $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ plot (denoted by d_{SS}) should be higher than 0, and the distance index for $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plot (denoted by d_{NN}) should be lower than 0.

The variance effect can be illustrated on the axis of log LR (λ). Specifically, the variance of the signal and noise distributions on the λ axis in easy condition should be higher than those in difficult condition (Glanzer et al., 2009, 2019; see Figure 2B). This difference in variance can be quantitatively represented by the slope for the $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ plot (s_{SS}) and the $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plot (s_{NN}):

$$s_{SS} = \frac{\sigma_{S(\text{difficult})}}{\sigma_{S(\text{easy})}},$$

$$s_{NN} = \frac{\sigma_{N(\text{difficult})}}{\sigma_{N(\text{easy})}}.$$

In the equations above, σ_S is the standard deviation for signal distribution, and σ_N is the standard deviation for noise distribution. If the variance effect holds, then both s_{SS} and s_{NN} should be lower than 1 (Glanzer et al., 2009, 2019; see Figure 2C).

The zROC length effect is first reported by Stretch and Wixted (1998a), who found that response criteria for confidence ratings on the strength axis tended to fan out in difficult tasks (see Figure 2A). This effect can be represented on the zROC where zHR is plotted against zFAR separately in each experimental condition, which indicates that the distance between the two end points of zROC is longer in difficult than easy condition (Glanzer et al., 2009, 2019; see Figure 2D). The zROC length effect is naturally predicted by the constant LR account: the confidence criteria on the strength axis have to contract when the distance between signal and noise distributions increases (i.e., the task becomes easier) to maintain constant LR corresponding to each evidence strength criterion (Glanzer et al., 2009, 2019; Semmler et al., 2018)

Glanzer et al. (2009, 2019) have analyzed data from both previously published studies and new experiments using tasks from various cognitive domains, such as long-term memory, visual working memory, perception, reasoning, and mental rotation. The three regularities are obtained in most of these experiments, suggesting that people change decision criteria on the strength axis across levels of task difficulty in such a way that the LR theory predicts. In contrast, it is relatively difficult for the strength theory to explain the difference in evidence strength criteria between experimental conditions as shown in empirical datasets (Glanzer et al., 2019). Thus, people are likely to make decisions in signal detection tasks based on LR.

Previous studies have tried to explain one of the regularities, the mirror effect, on the basis of the strength theory with additional assumptions. Some researchers suggest the inclusion of a new noise distribution and hypothesize that manipulating the difficulty in detecting signals may affect the evidence strength in not only the signal distribution, but also the noise distribution with the absence of signals (e.g., Koop et al., 2019; McClelland & Chappell, 1998). Another explanation includes a criterion-shift process, which assumes that people may directly set decision criteria on the axis of evidence strength, and the criteria shift across tasks with different levels of difficulty (Hirshman, 1995; Singer & Wixted, 2006; Stretch & Wixted, 1998b). In addition, the criterion-shift account suggests that during forced-choice tasks in which people should choose the correct answer (i.e., signal) out of two or more options, the shift of decision criteria on the strength axis may be more complicated: separate evidence strength criterion, which shifts across various difficulty levels, would be required for each option, and people may need to separately evaluate the distance of each option's evidence strength from its corresponding criterion (Gillund & Shiffrin, 1984). On the other hand, the LR theory offers a more straightforward explanation for the mirror effect: people only need to make decisions based on LR rather than raw evidence strength. No additional noise distribution is required in the model, and decision criteria on the LR axis need not to be changed across tasks with different difficulty levels (Glanzer et al., 2009, 2019).

Objective LR Versus Subjective LR

Glanzer et al. (2009, 2019) derived the three regularities described above based on the assumption that people set decision criteria on the ratio of objective likelihood in the true signal and noise distributions (see also Semmler et al., 2018). Because the three regularities have been widely found in previous studies, Glanzer et al. suggest that people should have knowledge about the exact shape of the signal and noise distributions. However, other researchers have questioned whether people are really able to compute the objective LR given the information they receive in a certain signal detection task (Balakrishnan & Ratcliff, 1996; Criss & McClelland, 2006; Hintzman, 1994; McClelland & Chappell, 1998). In fact, according to BDT, people may incorrectly estimate the shape (e.g., mean or variance) of the distributions, leading to suboptimal decision criteria and bias in decision making (Lau, 2007; Wickens, 2002).

In response to the question above, Wixted and Gaitan (2002) suggest that people's ability to compute LR comes from everyday

learning based on feedback, and they can learn the probability that their response is correct given specific evidence strength (see also Turner et al., 2011). For example, people should be able to learn in their everyday life that they need stronger evidence to confidently discriminate signal and noise when the task is more difficult, leading to the fact that confidence criteria fan out on the axis of evidence strength as task difficulty increases (Semmler et al., 2018). However, few previous studies have conducted empirical experiments to directly investigate whether people's subjectively estimated LR is the same as or systematically different from the objective LR.

To our knowledge, there is one empirical study examining people's estimation of the shape of the underlying distributions in signal detection task (Kubovy, 1977). Specifically, in Kubovy's experiment, participants were presented with a number in each trial and were told that the number represented the height of either a woman or man. Unknown to participants, the mean of the distribution for the height of men was higher than that of women, and both distributions had the same variance. In the first several sessions, participants were asked to judge whether each number came from the distribution of women or men and received feedback about correctness. Then in the following sessions, participants were given another series of numbers, and required to estimate the probability (from 0 to 100) that each number had been sampled from the distribution of men (i.e., the posterior probability in BDT). Results showed that participants reliably overestimated the posterior probability of men when the number was high and underestimated the probability when the number was low. Because the prior probability of sampling the height of women and men was equal (i.e., 0.5) and clearly told to participants before the experiment, Kubovy suggested that the incorrect estimation of posterior probability should result from a misconception of the likelihood in the distributions. Based on Kubovy's study, it is also possible that when making decisions based on the LR criteria in signal detection tasks, people's estimation of LR may deviate from the true LR.

However, according to BDT, directly obtaining people's estimation of LR and comparing subjective and objective LR based on empirical data in a certain signal detection task is difficult due to parameter redundancy (Lau, 2007; Wickens, 2002). For example, when people set a conservative criterion on the axis of objective LR, it is often unlikely to distinguish whether they are indeed willing to give "noise" response (based on the evaluation of prior probabilities and gain functions), or they actually set unbiased criterion on subjective LR but overestimate the mean of signal distribution (Lau, 2007).

Here we suggest that whether people are able to accurately estimate objective LR can be tested in two-condition experiments. The LR theory indicates that each decision criterion on the strength axis can be mapped to a certain value of LR (Glanzer et al., 2009, 2019).[2] Although it is difficult to directly obtain the subjectively estimated LR for evidence strength criteria in empirical experiments, the objective ratio of likelihood in the true signal and noise distributions corresponding to each criterion on the strength axis can be computed. According to the constant LR account, the decision criteria on the axis of LR should remain the same across experimental conditions with various difficulty levels (Glanzer et al., 2009, 2019; Semmler et al., 2018). If the computed objective LR for the evidence strength criteria is equal across levels of difficulty, we can conclude that people set decision criteria based on objective LR. However, if the computed objective LR differs between conditions, then the seemingly violation of the constant LR account may suggest that people's decision criteria are not based on objective LR, and thus the subjective and objective LR may be different.

The Current Study

The current study is organized as follows. First, we demonstrate that the three regularities proposed by Glanzer et al. (2009, 2019) hold even when the objective LR corresponding to the evidence strength criteria differs across various levels of task difficulty. Thus, observing the three regularities in empirical experiments is not in itself sufficient to conclude that people set decision criteria based on objective LR. We then describe three new experiments in which participants were asked to perform either perceptual or memory tasks, and the tasks in each experiment contained two difficulty levels. We estimated the decision criteria based on objective LR, and compared the objective LR criteria between the two experimental conditions (easy vs. difficult). To foreshadow, our results revealed that the objective LR criteria reliably differed across the conditions and fanned out as task difficulty decreased, indicating participants' estimation of LR might be inaccurate. Finally, we propose several possible explanations for the current results in the General Discussion section.

Reexamining the Three Regularities

In this section, we will show that the three regularities, which are originally derived based on the assumption that people maintain constant decision criteria on the axis of objective LR across experimental conditions with various difficulty levels (Glanzer et al., 2009, 2019), actually hold even when the objective LR corresponding to the decision criteria differs between conditions. Let us first consider the variance effect and *z*ROC length effect. The variance effect is represented by the $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ and $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plots with slope lower than 1. In the $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ plot, the difference in *z*HR between the two end points of the plot should be smaller in easy than difficult condition when the slope is lower than 1. Similarly, the range of *z*FAR should also be smaller for easy condition to observe a slope lower than 1 in the $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plot (see Figure 2C). When the ranges of both *z*HR and *z*FAR are smaller in easy than difficult task, the Euclidean distance between the end points of the *z*HR/*z*FAR plot must be shorter in easy task, leading to the *z*ROC length effect (see Figure 2D). Thus, the establishment of the variance effect is sufficient for the establishment of the *z*ROC length effect, and both effects hold as long as *z*HR and *z*FAR have a narrower range in easy than difficult task.

How should we narrow the range of *z*HR and *z*FAR by adjusting the decision criteria in signal detection tasks? For mathematical simplicity, here we use equal variance SDT as an example, in which the means of signal and noise distributions are denoted by

[2] In equal variance SDT, each evidence strength is mapped to exactly one value of LR and vice versa. In unequal variance SDT, two different criteria on the strength axis can be mapped to the same LR criteria (Glanzer et al., 2009; Wickens, 2002).

μ_S and μ_N, respectively, and both distributions have a variance of 1 (Wickens, 2002). In this case, the zHR and zFAR can be calculated as (supposing the decision criterion on the axis of evidence strength x is C):

$$\mathrm{HR} = P(x > C|\mu = \mu_S) = 1 - \Phi(C - \mu_S) = \Phi(\mu_S - C)$$

$$\mathrm{FAR} = P(x > C|\mu = \mu_N) = 1 - \Phi(C - \mu_N) = \Phi(\mu_N - C)$$

$$z\mathrm{HR} = \Phi^{-1}(\mathrm{HR}) = \mu_S - C$$

$$z\mathrm{FAR} = \Phi^{-1}(\mathrm{HR}) = \mu_N - C$$

in which Φ refers to the cumulative distribution function of standard normal distribution. From the equations above, we can conclude that the ranges of both zHR and zFAR in equal variance SDT are equal to the range of evidence strength criteria C. Thus, zHR and zFAR should have a narrower range in easy condition as long as the evidence strength criteria contract in easy (compared with difficult) condition, resulting in the variance effect and zROC length effect.

Let us then examine the mirror effect. According to previous studies (Glanzer et al., 2019; Hilford et al., 2015), the mirror effect is indicated by the distance mirror index d_e in the $z\mathrm{HR}_{\mathrm{easy}}/z\mathrm{HR}_{\mathrm{difficult}}$ and $z\mathrm{FAR}_{\mathrm{easy}}/z\mathrm{FAR}_{\mathrm{difficult}}$ plots, which is calculated based on both the slope and intercept of the plots. Because the slope in these two plots should always be higher than 0, we can conclude that the mirror effect holds when the intercept is higher than 0 for the $z\mathrm{HR}_{\mathrm{easy}}/z\mathrm{HR}_{\mathrm{difficult}}$ plot, and lower than 0 for the $z\mathrm{FAR}_{\mathrm{easy}}/z\mathrm{FAR}_{\mathrm{difficult}}$ plot. In Section S1 in the online supplemental materials, we provide several ways to adjust decision criteria on the strength axis to obtain the mirror effect in equal variance SDT. Specifically, the mirror effect exists when the evidence strength criteria either contract, expand, or simply shift in the easy (compared with difficult) condition. However, the variance effect and zROC length effect cannot be observed in the latter two cases (see Figure S1 in the online supplemental materials).

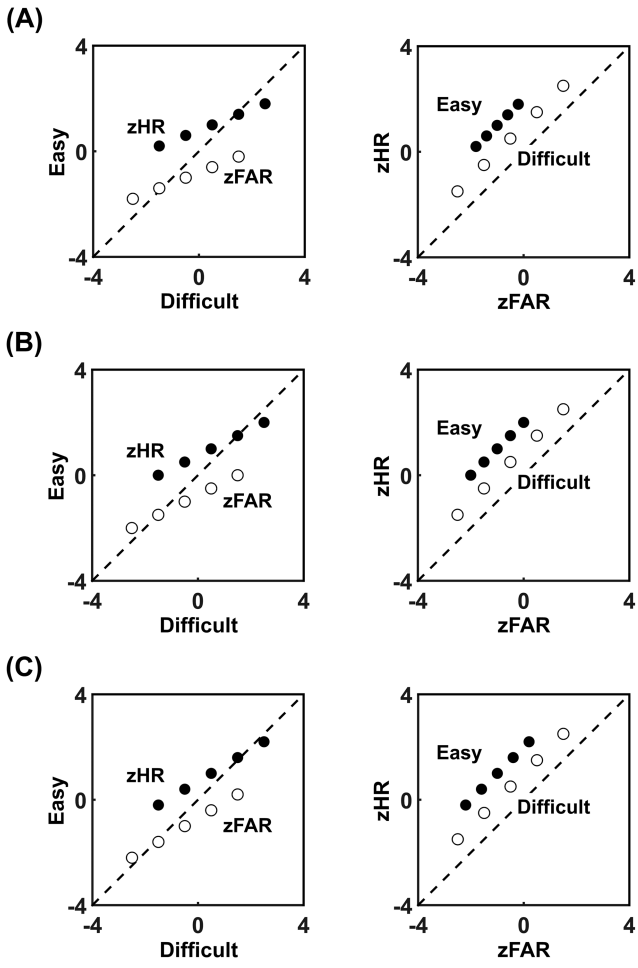
In brief, to obtain the three regularities at the same time in equal variance SDT, we only need to contract the evidence strength criteria when task difficulty decreases. Any theory that accounts for the contraction of evidence strength criteria in the easy condition can explain the regularities found in previous empirical studies. Maintaining constant objective LR is not the only possible prerequisite for the strength criteria contraction in easy tasks, which can actually be observed even when the objective LR corresponding to the strength criteria differs across levels of task difficulty. In Section S2 in the online supplemental materials, we mathematically prove that the decision criteria on the strength axis can contract in the easy condition when the criteria on the axis of objective LR either contract, expand, or remain constant. This is consistent with the results from the study conducted by Stretch and Wixted (1998a), who indicate that although the confidence criteria on the strength axis fan out as task difficulty increases, the adjustment of confidence criteria is not quantitatively consistent with the hypothesis that people maintain constant objective LR criteria across conditions. Specifically, as the distance between signal and noise distributions approaches zero, the evidence strength criteria based on constant objective LR should fan out infinitely far. However, the actual strength criteria did not fan out to such an extreme degree in Stretch and Wixted's study.

Several examples are shown in Figure 3 to further illustrate the relationship between the three regularities and the change of decision criteria on the objective log LR (λ) axis across difficulty levels. Here, we assume the distance between the means of the signal and noise distributions ($\mu_S - \mu_N$, which is often denoted by d') is 1 for difficult condition and 2 for easy condition. We also set five decision criteria on the objective λ axis in the difficult condition ($\lambda_{\mathrm{difficult}}$), including -2, -1, 0, 1, and 2. Furthermore, the decision criteria on the objective λ axis in the easy condition (λ_{easy}) are calculated as $\lambda_{\mathrm{easy}} = b \times \lambda_{\mathrm{difficult}}$, in which b is equal to 0.8, 1, and 1.2 in Figure 3A–C, respectively. The three regularities can be observed in all of the three subfigures in Figure 3, even when b is unequal to 1 (i.e., decision criteria on the λ axis are different between easy and difficulty conditions).

All of the conclusions above are derived based on equal variance SDT. In unequal variance SDT, the relationship between zHR/zFAR

Figure 3

The $z\mathrm{HR}_{easy}/z\mathrm{HR}_{difficult}$ and $z\mathrm{FAR}_{easy}/z\mathrm{FAR}_{difficult}$ Plots (Left) and zHR/zFAR Plots (Right) for Equal Variance SDT When b *Is Equal to 0.8 (A), 1 (B), or 1.2 (C)*



Note. See the main text for details. HR = hit rate; FAR = false alarm rate; SDT = signal detection theory.

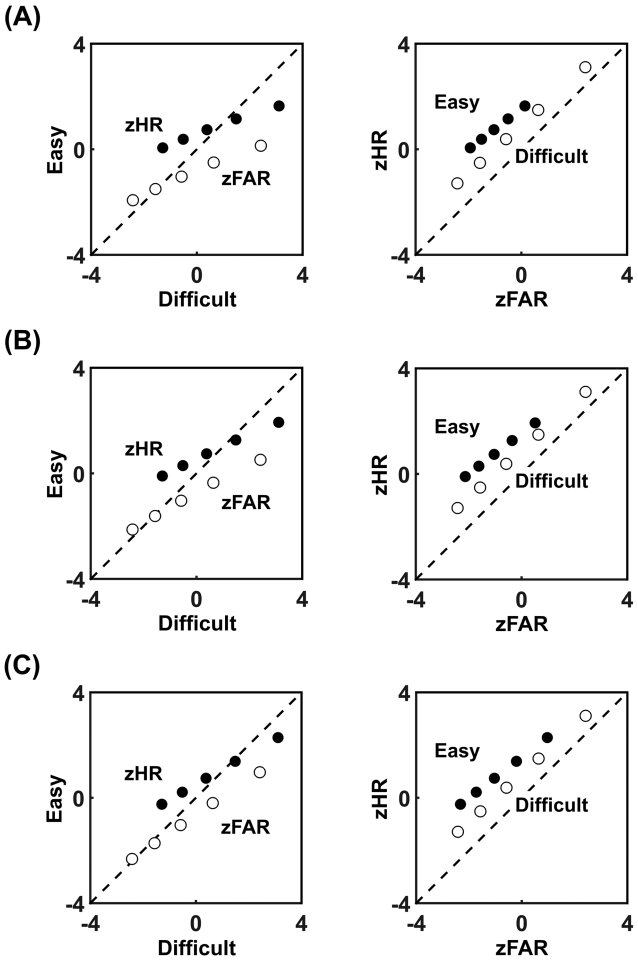
and decision criteria on the LR axis becomes more complex, and it is relatively difficult to mathematically demonstrate how to adjust the LR criteria to obtain the three regularities. However, we are still able to provide examples in which the objective LR corresponding to the decision criteria differs across difficulty levels but the three regularities still hold (see Figure 4). The setting of parameter values in Figure 4 is very similar to that in Figure 3, except that the standard deviation of the signal distribution is 1.3 for easy condition and 1.1 for difficult condition, while the noise distribution has a standard deviation of 1 in both easy and difficult conditions. The three regularities exist in all of the three subfigures in Figure 4, even when *b* is unequal to 1.

In summary, we illustrate that the three regularities derived by Glanzer et al. (2009, 2019) can be obtained when the objective LR for the decision criteria is different across levels of task difficulty. Thus, observing the three regularities in empirical experiments is insufficient to conclude that the objective LR criteria remain constant across conditions, and that people set decision criteria based on objective LR. We argue that to examine whether people's decision criteria are based on the same objective LR across difficulty levels, it should be better to directly estimate the value of objective LR corresponding to the decision criteria from empirical data separately in easy and difficult task, and compare the estimated objective LR criteria in the two conditions.

To address this issue, we conducted three experiments in which participants were asked to perform either two-alternative-forced-choice (2AFC) perceptual decision tasks (Experiment 1), 2AFC recognition memory tests (Experiment 2), or Old/New recognition memory tests (Experiment 3). Each experiment contained two different levels of task difficulty (easy vs. difficult). We separately fit the SDT model to data in each experimental condition and investigated whether there was reliable difference between the computed objective LR criteria in the two conditions. To foreshadow, our results indicated that the objective LR criteria differed across conditions and fanned out as task difficulty decreased. Thus, it should be highly possible that the subjectively estimated LR, on which participants' decisions were based, differed from objective LR, because relying on objective LR to make decisions would violate the constant LR account.

Figure 4

The $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ Plots (Left) and zHR/zFAR Plots (Right) for Unequal Variance SDT When b *Is Equal to 0.8 (A), 1 (B), or 1.2 (C)*



Note. See the main text for details. HR = hit rate; FAR = false alarm rate; SDT = signal detection theory.

Experiment 1

In Experiment 1, participants were asked to perform two 2AFC perceptual decision tasks, including a color discrimination task and a dot discrimination task, and give their response using a 6-point confidence rating scale. The two tasks were taken from a previous study by Hu et al. (2023), in which the performance was reliably higher for dot discrimination task than color discrimination task (i.e., the dot discrimination task is easier). Here we used 2AFC tasks because participants' response in a 2AFC task can be characterized by SDT with two distributions that have equal variance, leading to simplicity in mathematics. Specifically, SDT assumes that participants make decisions in a 2AFC task based on the difference in evidence strength between the two stimuli (a correct and an incorrect stimulus) presented in each trial, and the strength difference in the two types of trials (e.g., whether the correct stimulus is presented on the left or right side of the screen) has equal variance even if the variances of the evidence strength for correct and incorrect stimuli are unequal (Wickens, 2002).

According to the constant LR account, if participants set decision criteria based on objective LR, then the objective LR criteria should be the same in easy and difficult tasks. However, if the value of objective LR reliably differed between conditions, then participants' decisions should depend on the subjectively estimated LR that was systematically different from objective LR.

Method

Transparency and Openness

The experimental materials, data, and scripts for all experiments can be accessed at the Open Science Framework (OSF) project page (https://osf.io/grcsj/; Hu, 2023). We obtained ethical approval to conduct this research from the Ethics Committee at Beijing Normal University. This study was not preregistered.

Participants

The sample size in Experiment 1 was determined based on the results from Glanzer et al. (2019), who observed Cohen's ds > 0.41 for the three regularities in their Experiments 1 and 2. A power analysis conducted by G*power revealed that at least 49 participants were needed to replicate the three regularities (at a significance level of .05 in a two-tailed test) with a power of .80 (Faul et al., 2007). Accordingly, 50 participants (25 female and 25 male; age: $M =$ 28.60 years, $SD = 6.07$) were recruited online via the Prolific website (https://prolific.ac/). Participants signed informed consent before the experiment and received monetary compensation (£4) and a bonus (up to £0.5, dependent on task performance) after the experiment. All participants spoke English as the first language and reported normal or corrected-to-normal vision. All procedures were approved by the Ethics Committee at Beijing Normal University.

Materials and Procedure

Participants were asked to perform two perceptual decision tasks, including a color discrimination task and a dot discrimination task. There were 100 trials in each task, and task order was randomly assigned for each participant. Both tasks were programmed using Gorilla (Anwyl-Irvine et al., 2020).

In the color discrimination task (see Figure 5A), a fixation was first presented for 500 ms in each trial, after which two rectangles were presented on the screen for 2,000 ms. Each rectangle was split into two areas filled by either orange or blue color. For one of the two rectangles, the sizes of blue and orange areas were equal (the incorrect stimulus). For the other one (the correct stimulus), the blue area was larger than the orange area (the difference between blue and orange area was between 2.5% and 16% of the total rectangle area, randomly defined in each trial). The positions (left or right) for the two rectangles were randomly determined in each trial. After stimuli presentation, participants needed to press a number key from 1 to 6 to decide which rectangle contained larger blue area, and rate confidence about their decision. The six points on the confidence rating scale were 1 = *sure left*, 2 = *probably left*, 3 = *maybe left*, 4 = *maybe right*, 5 = *probably right*, and 6 = *sure right*. There was no time pressure on the confidence rating. The chosen number would then be highlighted for 500 ms.

In the dot discrimination task (see Figure 5B), after presenting a fixation for 500 ms in each trial, two circles were shown on the screen for 700 ms. One of the circles always contained 50 dots (the incorrect stimulus), and the number of dots in the other circle varied between 51 and 75, randomly defined in each trial (the correct stimulus). Participants needed to press a number key from 1 to 6 to decide which circle contained more dots and rate their confidence. The chosen number would then be highlighted for 500 ms, as in the color discrimination task.

Data Analysis

Model Fitting. We fit equal variance SDT model to data of confidence ratings. The model assumed that participants made decisions based on objective LR, and should give higher confidence ratings when objective LR was higher. There are six free parameters in the model, including the distance (d') between the means of the distributions for two types of trials (whether the correct stimulus was presented on the left or right side of the screen; here we arbitrarily defined the trials with correct stimulus on the right side as "signal" and trials with correct stimulus on the left side as "noise" in the equal variance SDT model), and five decision criteria on the objective log LR (λ) axis (λ_1, λ_2, λ_3, λ_4, and λ_5 from left to right). Based on the model assumptions, we computed the likelihood of observing the current data (confidence ratings) given the value of each parameter (see Section S3 in the online supplemental materials for details). Then, we fit the model to data separately in each task for each participant and found the parameter values that maximized the likelihood function.

Some participants did not use all of the confidence rating categories, which might lead to error in the model fitting procedure and parameter estimation. To solve this issue, here we performed a padding correction before fitting the SDT model (Hautus, 1995; Maniscalco & Lau, 2012).[3] Specifically, for each participant, we first counted the number of trials with each point on the confidence rating scale separately for signal and noise trials in each task and then added $1/n$ to the count for each point regardless of whether zeros were present. The n represents the total number of points on the confidence scale (6 in the current study). All of the data analyses below were conducted after padding correction.

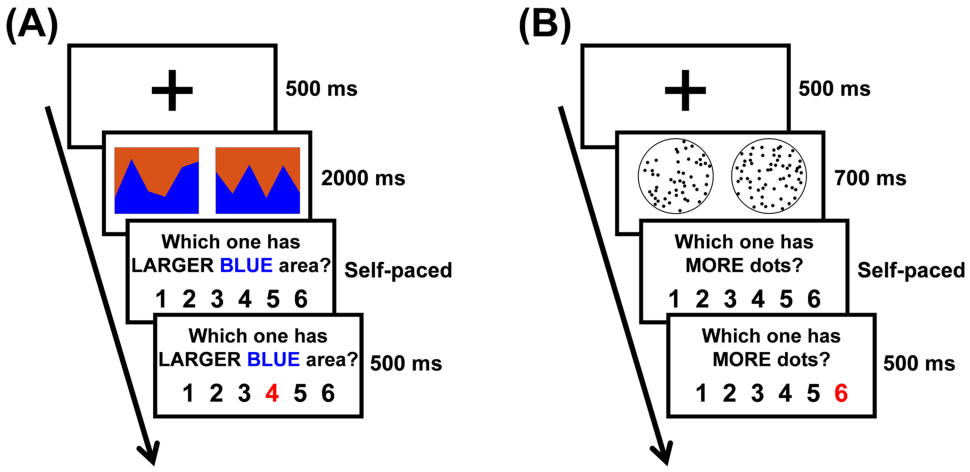
Comparing the Performance in Two Tasks. After finishing model fitting, we first compared the performance in color discrimination and dot discrimination tasks using two different measures of task performance. The first measure is the estimated value of d' from the SDT model. The second measure is A_{trap} which is a model-free measure of the area under the ROC curve based on trapezoidal rule (Wickens, 2002). To compute A_{trap}, we need to connect the neighboring observed HR/FAR points on the ROC curve, as well as the (0, 0) and (1, 1) points, with straight lines. Then A_{trap} is calculated as the area under the straight lines. We performed paired-sample t tests to examine whether the difference in the two measures of performance between color and dot discrimination tasks was statistically detectable. To foreshadow, our results revealed that task performance was reliably better in the dot discrimination task than color discrimination task. Thus, in the following data analyses, we define the dot discrimination task as easy condition and color discrimination task as difficult condition.

Examining the Three Regularities. Next, we examined whether the three regularities proposed by Glanzer et al. (2009, 2019) could also be observed in the current experiment. For each participant, the zHR/zFAR plots separately in easy and difficult conditions, as well as the $z\text{HR}_{easy}/z\text{HR}_{difficult}$ and $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$ plots, were drawn based on confidence rating data. To examine the mirror effect, we fit the linear model separately to the $z\text{HR}_{easy}/z\text{HR}_{difficult}$ and $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$ plots and computed the distance mirror index (d_{SS} and d_{NN}) based on the slope and intercept of the two plots. One-sample t tests were conducted to investigate whether d_{SS} and d_{NN} were significantly different from 0. To examine the variance effect, we performed one-sample t tests to investigate whether the slope of the $z\text{HR}_{easy}/z\text{HR}_{difficult}$ and $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$ plots (s_{SS} and s_{NN}) differed from 1. To examine the zROC length effect, we conducted paired-sample t test to compare

[3] Previous study suggests that padding correction may bias parameter estimation when the trial number is low (Fleming, 2017). However, the trial number in the current study is relatively large (100 trials in each experimental condition), and the effect of padding correction on parameter estimation should be small.

Figure 5

Procedure for the Perceptual Tasks in Experiment 1, Including the Color Discrimination Task (A) and Dot Discrimination Task (B)



Note. Each rectangle presented in the color discrimination task has an orange area (the top part) and a blue area (the bottom part). See the online article for the color version of this figure.

the Euclidean distance between the end points of the *z*HR/*z*FAR plot in easy and difficult conditions.

Comparing the Objective Log LR Criteria in Two Tasks. Finally, we explored the potential difference between decision criteria on the λ axis (λ_1, λ_2, λ_3, λ_4, and λ_5) in easy and difficult conditions. Specifically, we conducted a 2 (condition: easy vs. difficult) × 5 (position: 1, 2, 3, 4, and 5) repeated measures analysis of variance (ANOVA) on the estimated value of λ criteria from the SDT model. We were mostly interested in the main effect of condition and the interaction effect between condition and position, which could indicate whether the objective LR corresponding to decision criteria was affected by task difficulty.

Glanzer et al. (2019) suggest that estimating the λ criteria used by participants is often difficult when the true value of λ is extreme (i.e., when λ lies far from 0). In this case, the value of λ estimated from empirical confidence rating data is determined by a small proportion of trials, and thus very noisy. Furthermore, estimation of decision criteria in the tails of distributions largely relies on assumptions about the distribution shape, and a subtle change in the tails of the signal and noise distributions leads to large change in the estimation of their ratio (Maloney & Thomas, 1991). To mitigate this concern, we also conducted a 2 (condition: easy vs. difficult) × 3 (position: 2, 3, and 4) repeated measures ANOVA on the middle three λ criteria to rule out the effect of the two extreme criteria (λ_1 and λ_5) on statistical results.

We conducted both frequentist and Bayesian hypothesis tests when performing the *t* tests and ANOVA described above. All Bayesian hypothesis tests were performed via JASP (http://www.jasp-stats.org). For frequentist repeated measures ANOVA, Greenhouse–Geisser correction was used to adjust for violation of the sphericity assumption.

Results

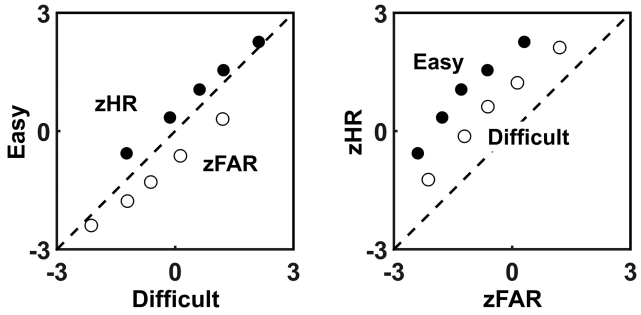
The estimated value of d' was reliably greater in the dot discrimination task ($M = 2.23$, $SD = 0.35$) than color discrimination task ($M = 1.12$, $SD = 0.33$), $t(49) = 18.20$, $p < .001$, Cohen's $d = 2.57$, $BF_{10} > 100$. Similarly, the dot discrimination task ($M = 0.92$, $SD = 0.04$) also had a higher value of A_{trap} than the color discrimination task ($M = 0.76$, $SD = 0.06$), $t(49) = 16.43$, $p < .001$, Cohen's $d = 2.32$, $BF_{10} > 100$. These results indicated that the dot discrimination task was easier than the color discrimination task, consistent with the findings observed in previous study (Hu et al., 2023).

The $z\text{HR}_{easy}/z\text{HR}_{difficult}$, $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$, and *z*HR/*z*FAR plots averaged across participants are shown in Figure 6, in which all of the three regularities could be obtained. Specifically, the distance mirror index was reliably higher than 0 for the $z\text{HR}_{easy}/z\text{HR}_{difficult}$ plot (d_{SS}), $t(49) = 7.86$, $p < .001$, Cohen's $d = 1.11$, $BF_{10} > 100$, and lower than 0 for the $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$ plot (d_{NN}), $t(49) = -11.87$, $p < .001$, Cohen's $d = -1.68$, $BF_{10} > 100$, indicating the mirror effect (see Figure 7A). In addition, the slope of the $z\text{FAR}_{easy}/z\text{FAR}_{difficult}$ plot (s_{NN}) was reliably lower than 1, $t(49) = -3.14$, $p = .003$, Cohen's $d = -0.44$, $BF_{10} = 11.27$. However, the difference between the slope of the $z\text{HR}_{easy}/z\text{HR}_{difficult}$ plot (s_{SS}) and 1 was only marginally significant, and the Bayes factor was inconclusive, $t(49) = -1.93$, $p = .059$, Cohen's $d = -0.27$, $BF_{10} = 0.86$, which might be due to the presence of an extreme value ($s_{SS} = 2.59$). To reduce the effect of the extreme value on statistical analysis, we conducted one-sample Wilcoxon signed rank test showing that s_{SS} was significantly lower than 1, $Z = -2.74$, $p = .006$, $BF_{10} = 4.53$. These results suggest that the variance effect existed in the current experiment (see Figure 7B). Furthermore, the distance between end points of the *z*HR/*z*FAR plot was longer in difficult (color discrimination task) than easy (dot discrimination task) condition, and this difference was statistically detectable, $t(49) = 4.25$, $p < .001$, Cohen's $d = 0.60$, $BF_{10} > 100$, revealing the *z*ROC length effect (see Figure 7C).

Finally, we performed a 2 (condition) × 5 (position) ANOVA on the estimated value of the five λ criteria. The main effect of position was statistically significant, $F(1.30, 63.75) = 154.09$, $p < .001$, $\eta_p^2 = .76$, $BF_{incl} > 100$ (Greenhouse–Geisser corrected), indicating

Figure 6

The $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ Plots (Left) and zHR/zFAR Plots (Right) Averaged Across Participants in Experiment 1



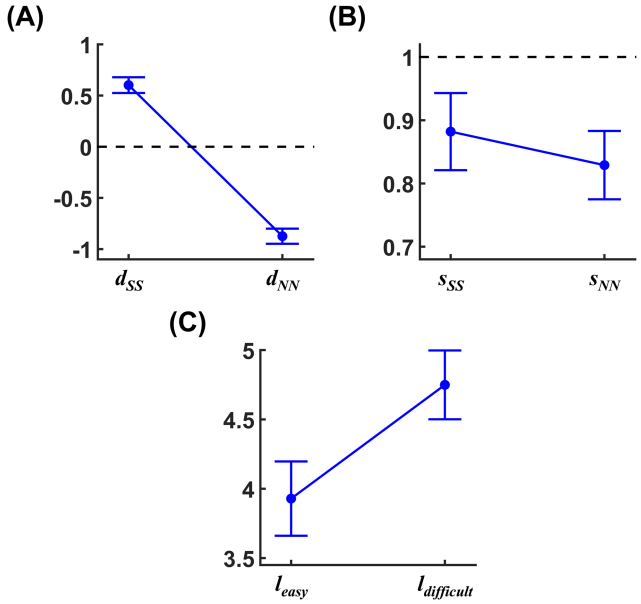
Note. HR = hit rate; FAR = false alarm rate.

that the value of λ for the decision criteria increased along the axis from left to right. The main effect of condition was also statistically detectable based on p value, although the Bayes factor was inconclusive, $F(1, 49) = 12.33$, $p < .001$, $\eta_p^2 = .20$, $BF_{incl} = 1.07$, suggesting that the value of λ criteria might be overall higher in easy than difficult condition. Most importantly, there was a reliable interaction effect between condition and position, $F(1.28, 62.71) = 28.99$, $p < .001$, $\eta_p^2 = .37$, $BF_{incl} > 100$ (Greenhouse–Geisser corrected). We then conducted a 2 (condition) × 3 (position) ANOVA on the middle three λ criteria. Similarly, we found that the main effect of condition was statistically detectable based on p value but the Bayes factor was inconclusive, $F(1, 49) = 12.57$, $p < .001$, $\eta_p^2 = .20$, $BF_{incl} = 2.43$. In addition, there were reliable main effect of position and interaction effect between condition and position, Fs > 14.31, ps < .001, $\eta_p^2 > .22$, $BF_{incl} > 100$.

Further analyses revealed that λ was reliably lower in easy than difficult condition at Position 1, $t(49) = 4.64$, $p < .001$, Cohen's $d = 0.66$, $BF_{10} > 100$. Although λ was also lower in easy condition at Position 2, this difference did not reach statistical significance, $t(49) = 1.87$, $p = .068$, Cohen's $d = 0.26$, $BF_{10} = 0.76$. Furthermore, λ was significantly higher in easy than difficult condition at Positions 3–5, ts > 3.02, ps < .005, Cohen's $d > 0.42$, $BF_{10} > 8.43$ (see Figure 8). These results indicate that compared with the difficult condition, decision criteria on the axis of objective log LR fanned out in the easy condition.

Figure 7

The Three Regularities in Experiment 1



Note. (A) The distance mirror index for the $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ plots (d_{SS} and d_{NN}). (B) The slope of the $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ plots (s_{SS} and s_{NN}). (C) The distance between end points of the zHR/zFAR plot in easy and difficult condition (l_{easy} and $l_{difficult}$). Error bars represent standard errors. HR = hit rate; FAR = false alarm rate. See the online article for the color version of this figure.

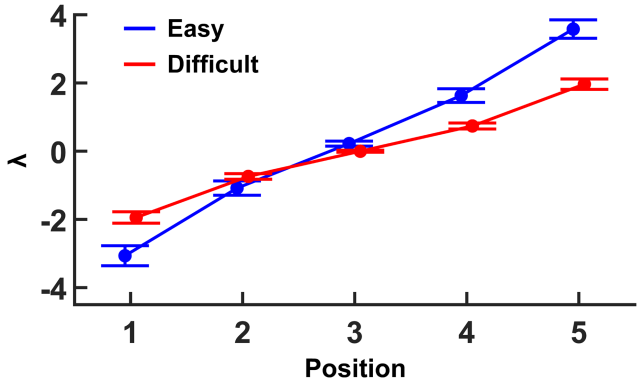
Discussion

Consistent with previous studies by Glanzer et al. (2009, 2019), all of the three regularities were observed in Experiment 1. However, we also found that the objective log LR (λ) corresponding to the decision criteria reliably differed between easy and difficult conditions, and λ criteria fanned out in the easy condition. Our results indicate that the three regularities can occur even when the value of objective LR for decision criteria varies across levels of difficulty. Furthermore, participants in Experiment 1 might not set decision criteria based on the objective LR in the true distributions, because relying on objective LR to make decisions violated the constant LR account.

The expansion of objective LR criteria in the easy condition could explain the results from the study by Stretch and Wixted (1998a), who found that the adjustment of decision criteria on the axis of evidence strength was not quantitatively consistent with the hypothesis that people maintained constant objective LR criteria across conditions. When the task was very difficult (i.e., d' was close to zero), the evidence strength criteria based on constant objective LR should fan out infinitely far. However, the actual strength criteria in Stretch and Wixted's study did not fan out to such an extreme extent. This

Figure 8

The λ Criteria as a Function of Experimental Condition (Easy vs. Difficult) and Position (From 1 to 5) in Experiment 1



Note. Error bars represent standard errors. See the online article for the color version of this figure.

result could be accounted for by the fact that the objective LR for decision criteria contracted as task difficulty increased, and thus the corresponding criteria on the strength axis were adjusted to a less degree in the difficult condition than those predicted by the hypothesis of maintaining constant objective LR.

One previous study conducted by Semmler et al. (2018) used another method, which we name as the model-comparison approach, to examine whether participants based their decision on objective LR. Specifically, Semmler et al. quantitatively compared the fit of two models to empirical data. In the first model, the decision criteria on the evidence strength axis could be freely estimated in both easy and difficult conditions. In contrast, the evidence strength criteria across the two conditions were constrained such that they had the same objective LR in the second model. Semmler et al. revealed that the second model could better fit the empirical data, according to which they suggest participants' decisions might depend on objective LR. This method is different from the parameter-estimation approach in the current study, in which we separately estimated the decision criteria on the objective λ axis in the easy and difficult condition, and then compared the λ criteria across the two conditions based on ANOVA. In Section S4 in the online supplemental materials, we used data simulation to illustrate that the parameter-estimation approach can accurately detect whether there is reliable difference in objective λ criteria across experimental conditions. However, the model-comparison approach lacks statistical power, and is unable to reveal the difference in objective λ criteria across conditions even when this difference exists. Thus, the result derived from the model-comparison approach that the objective λ for decision criteria does not differ between difficulty levels may be false negative, and we advocate the parameter-estimation approach used in the current study.

A limitation of Experiment 1 is that the two perceptual tasks differed not only in the level of difficulty, but also in the task requirements (discriminating two color areas or the number of dots). This is inconsistent with the original two-condition experimental design, in which participants are asked to perform the same task in two conditions and the only difference between conditions is the task difficulty (Glanzer et al., 2019). In addition, Experiment 1 only examined the difference in the objective LR criteria across difficulty levels in perceptual domain, and whether similar results could be observed in tasks from other cognitive domains (e.g., memory) needs further investigation. Thus, in Experiment 2, we asked participants to perform the same 2AFC recognition memory test in both easy and difficult conditions, and manipulated task difficulty using study time.

Experiment 2

In Experiment 2, participants learned a series of words and performed 2AFC recognition memory test in four blocks. The time for learning the words varied across blocks (30 vs. 90 s), which affected task difficulty. As in Experiment 1, participants gave their response in recognition test using a 6-point confidence rating scale. We aimed to replicate the main results in Experiment 1 and extend these findings to memory domain. That is, we expected that the value of objective LR corresponding to the decision criteria should fan out in the easy condition (with study time as 90 s) compared with the difficult condition (with study time as 30 s).

Method

Participants

Participants in Experiment 2 were recruited online via the Prolific website. We collected data from 50 participants (28 female and 22 male; age: $M = 30.72$ years, $SD = 6.91$), the same sample size as in Experiment 1. Data from another two participants were excluded because their area under the ROC curve, measured by A_{trap}, was lower than 0.5 in either easy or difficult condition, suggesting their performance was worse than chance level. Participants signed informed consent before the experiment and received monetary compensation (£4) and a bonus (up to £0.5, dependent on task performance) after the experiment. All participants spoke English as the first language and reported normal or corrected-to-normal vision. All procedures were approved by the Ethics Committee at Beijing Normal University.

Materials and Procedure

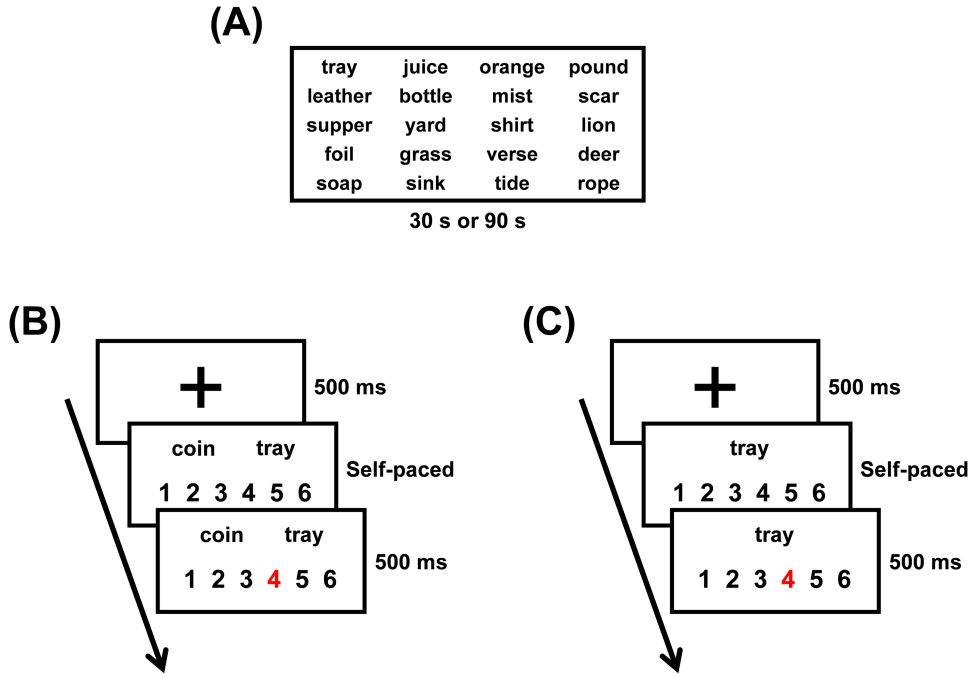
The experimental materials for the 2AFC recognition memory test in Experiment 2 were 400 English words selected from the MRC Psycholinguistic Database (Coltheart, 1981). All words contained 4–8 letters and 1–3 syllables. The Kucera and Francis word frequency for all words was between 8 and 90, and the ratings of familiarity, concreteness, and imageability were between 450 and 650. The words were randomly divided into two halves for each participant, in which 200 words were used as study materials in the learning phase, with the other 200 words serving as lures in the memory test.

The procedure for the recognition memory test was similar to that in previous studies (e.g., Fleming et al., 2014; Hu et al., 2023; McCurdy et al., 2013). The task contained four blocks, and in each block participants were first presented with 50 words simultaneously on the screen (arranged in 10 rows and five columns) and asked to memorize as many words as possible (see Figure 9A). The exposure duration of the words was 90 s for two of the four blocks (easy condition), and 30 s for the other two blocks (difficult condition). All of the words were randomly divided into the four blocks, and the order of the blocks was randomized across participants. When there were 10 s left of the learning phase, participants were notified by a countdown clock presented below the 50 words. Following the learning phase in each block, participants completed 50 trials in the test phase (see Figure 9B). After presenting the fixation for 500 ms, a learned word and a new word were simultaneously shown on the left and right side of the screen, respectively. The position of the two words was randomly determined in each trial. Participants needed to decide which word had been learned. As in Experiment 1, participants pressed a number key from 1 to 6 to make decision and rate their confidence. The chosen number would then be highlighted for 500 ms. The task was programmed using Gorilla.

Data Analysis

Data analysis in Experiment 2 was very similar to that in Experiment 1, except that we analyzed confidence rating data from two different conditions (study time: 30 vs. 90 s) in the same recognition memory test in Experiment 2 rather than from two perceptual tasks in Experiment 1. Here we define the blocks with 90 s study

Figure 9
Procedure for the Recognition Memory Tests in Experiments 2 and 3



Note. (A) The learning phase. (B) The test phase in Experiment 2. (C) The test phase in Experiment 3. See the online article for the color version of this figure.

time as easy condition and blocks with 30 s study time as difficult condition.

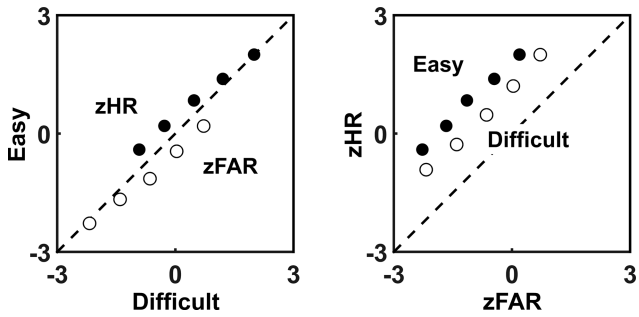
Results

The estimated value of d' was reliably greater when study time was 90 s ($M = 1.87$, $SD = 0.92$) than 30 s ($M = 1.14$, $SD = 0.53$), $t(49) = 7.23$, $p < .001$, Cohen's $d = 1.02$, $BF_{10} > 100$. Similarly, A_{trap} was also higher with 90 s ($M = 0.85$, $SD = 0.12$) than 30 s ($M = 0.76$, $SD = 0.10$) study time, $t(49) = 7.23$, $p < .001$, Cohen's $d = 1.02$, $BF_{10} > 100$. These results indicated that the difficulty level of recognition memory test reduced with increasing study time.

The $zHR_{easy}/zHR_{difficult}$, $zFAR_{easy}/zFAR_{difficult}$, and $zHR/zFAR$ plots averaged across participants are shown in Figure 10, in which all of the three regularities could be obtained. Specifically, the distance mirror index was reliably higher than 0 for the $zHR_{easy}/zHR_{difficult}$ plot (d_{SS}), $t(49) = 5.74$, $p < .001$, Cohen's $d = 0.81$, $BF_{10} > 100$, and lower than 0 for the $zFAR_{easy}/zFAR_{difficult}$ plot (d_{NN}), $t(49) = -5.77$, $p < .001$, Cohen's $d = -0.82$, $BF_{10} > 100$, indicating the mirror effect (see Figure 11A). In addition, the slope of the $zHR_{easy}/zHR_{difficult}$ plot (s_{SS}) was reliably lower than 1, $t(49) = -3.69$, $p < .001$, Cohen's $d = -0.52$, $BF_{10} = 47.45$. However, the difference between the slope of the $zFAR_{easy}/zFAR_{difficult}$ plot (s_{NN}) and 1 did not reach statistical significance, $t(49) = -0.96$, $p = .341$, Cohen's $d = -0.14$, $BF_{10} = 0.24$, which might be due to the presence of an extreme value ($s_{NN} = 3.71$). To reduce the effect of the extreme value on statistical analysis, we conducted one-sample Wilcoxon signed rank test showing that s_{NN} was significantly lower than 1, $Z = -3.11$, $p = .002$, $BF_{10} = 8.36$. These results suggest that the variance effect existed in the current experiment (see Figure 11B). Furthermore, the distance between end points of the $zHR/zFAR$ plot was longer in difficult (30 s study time) than easy (90 s study time) condition, and this difference was statistically detectable, $t(49) = 4.97$, $p < .001$, Cohen's $d = 0.70$, $BF_{10} > 100$, revealing the zROC length effect (see Figure 11C).

Finally, we performed a 2 (condition) × 5 (position) ANOVA on the estimated value of the five λ criteria. The main effect of position

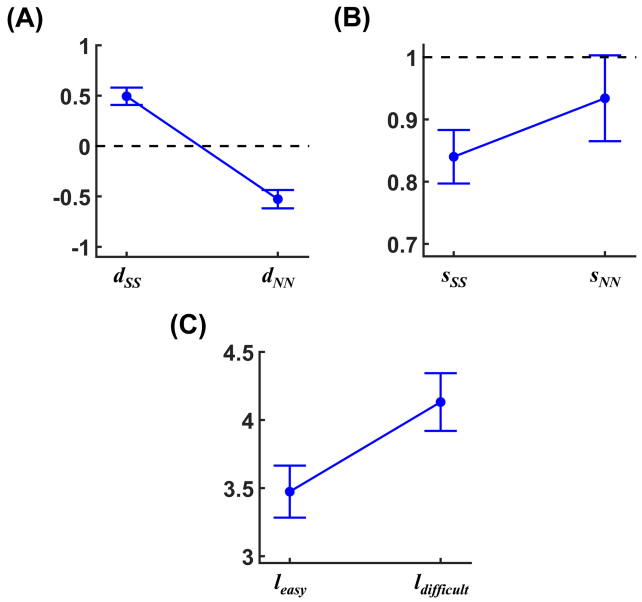
Figure 10
The $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ Plots (Left) and $zHR/zFAR$ Plots (Right) Averaged Across Participants in Experiment 2



Note. HR = hit rate; FAR = false alarm rate.

Figure 11

The Three Regularities in Experiment 2



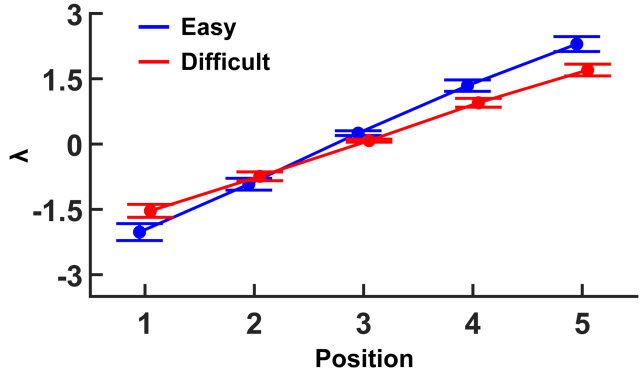
Note. (A) The distance mirror index for the $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ and $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plots (d_{SS} and d_{NN}). (B) The slope of the $z\text{HR}_{\text{easy}}/z\text{HR}_{\text{difficult}}$ and $z\text{FAR}_{\text{easy}}/z\text{FAR}_{\text{difficult}}$ plots (s_{SS} and s_{NN}). (C) The distance between end points of the zHR/zFAR plot in easy and difficult condition (l_{easy} and $l_{\text{difficult}}$). Error bars represent standard errors. HR = hit rate; FAR = false alarm rate. See the online article for the color version of this figure.

was statistically significant, $F(1.17,\ 57.21) = 137.66$, $p < .001$, $\eta_p^2 = .74$, $\text{BF}_{\text{incl}} > 100$ (Greenhouse–Geisser corrected), indicating that the value of λ for the decision criteria increased along the axis from left to right. The main effect of condition was also statistically detectable based on p value, although the Bayes factor was inconclusive, $F(1,\ 49) = 4.97$, $p = .030$, $\eta_p^2 = .09$, $\text{BF}_{\text{incl}} = 0.40$, suggesting that the value of λ criteria might be overall higher in easy than difficult condition. Most importantly, there was a reliable interaction effect between condition and position, $F(1.38,\ 67.69) = 15.96$, $p < .001$, $\eta_p^2 = .25$, $\text{BF}_{\text{incl}} > 100$ (Greenhouse–Geisser corrected). We then conducted a 2 (condition) × 3 (position) ANOVA on the middle three λ criteria. We found that the main effect of condition was statistically detectable based on p value but the Bayes factor was inconclusive, $F(1,\ 49) = 6.14$, $p = .017$, $\eta_p^2 = .11$, $\text{BF}_{\text{incl}} = 1.49$. In addition, there were reliable main effect of position and interaction effect between condition and position, Fs > 13.92, ps < .001, $\eta_p^2 > .22$, $\text{BF}_{\text{incl}} > 100$.

Further analyses revealed that λ was reliably lower in easy than difficult condition at Position 1, $t(49) = 3.51$, $p < .001$, Cohen's $d = 0.50$, $\text{BF}_{10} = 29.20$. Although λ was also lower in easy condition at Position 2, this difference did not reach statistical significance, $t(49) = 1.85$, $p = .071$, Cohen's $d = 0.26$, $\text{BF}_{10} = 0.74$. Furthermore, λ was significantly higher in easy than difficult condition at Positions 3–5, ts > 2.79, ps < .01, Cohen's $d > 0.39$, $\text{BF}_{10} > 4.82$ (see Figure 12). These results indicate that compared with the difficult condition, decision criteria on the axis of objective log LR fanned out in the easy condition.

Figure 12

The λ Criteria as a Function of Experimental Condition (Easy vs. Difficult) and Position (From 1 to 5) in Experiment 2



Note. Error bars represent standard errors. See the online article for the color version of this figure.

Discussion

Experiment 2 replicated the findings of Experiment 1. Specifically, all of the three regularities were observed in Experiment 2. Furthermore, decision criteria on the axis of objective log LR reliably differed across levels of difficulty, and fanned out in the easy condition. These results suggest that the expansion of decision criteria on the objective LR axis with decreasing task difficulty should be general across different cognitive domains (perception and memory).

In the current experiment, a learned word and a new word were simultaneously presented in each trial during memory test. Glanzer and colleagues have used another special experimental paradigm containing the null target condition in which two old words were presented in the same trial, and the null lure condition in which two new words were presented. In addition, the two words in each trial had different levels of task difficulty (e.g., high or low word frequency). Participants were forced to choose which one of the two words was seen earlier (Glanzer et al., 1991; Glanzer & Bowles, 1976). Results revealed that participants were more likely to choose easy (rather than difficult) word when they tried to distinguish two old words, while they preferred to pick the difficult word when two new words were presented. Furthermore, the proportion of choosing easy or difficult word in both null target and null lure conditions changed when a delay was introduced between study and test phase (Glanzer et al., 1991). This result is interesting especially because the probability of choosing one of two items that were not learned is affected by the quality of memory for learned words (i.e., immediate vs. delayed test). Glanzer et al. (1991) explained these results using the phenomenon of concentering, which is naturally derived from the LR theory and suggests both the signal and noise distributions on the axis of LR converge toward zero as memory quality gets worse (see also Figure 2B). Future studies could use similar experimental design, and examine the decision criteria on the objective LR axis in null target and null lure conditions.

Experiments 1 and 2 used 2AFC tasks to examine whether the objective LR corresponding to decision criteria differed across difficulty levels in equal variance SDT model. However, not all of the signal detection tasks could be characterized by equal variance

SDT. For example, previous studies indicate that in the Old/New recognition memory test in which participants are required to judge whether each stimulus has been encountered before or not in a particular context (i.e., old vs. new), the variance of signal distribution (old stimuli) is often higher than that of noise distribution (new stimuli; Mickes et al., 2007; Spanton & Berry, 2020; Starns & Ratcliff, 2014). In Experiment 3, we further investigated the effect of task difficulty on objective LR criteria in unequal variance SDT.

Experiment 3

Participants in Experiment 3 performed a recognition memory test with two levels of difficulty manipulated by study time (30 vs. 90 s). However, instead of selecting the learned word from two different words in each trial (as in Experiment 2), participants needed to decide whether each presented word had been learned or not in the study phase (old or new) in Experiment 3. We fit the unequal variance SDT model to the data from this Old/New recognition test. In unequal variance SDT, the correspondence between decision criteria on the strength axis and LR axis becomes complex, and the strength theory and LR theory make different predictions about participants' response in the task (see Footnotes 1 and 2; Wickens, 2002). Here, we assumed that participants made decisions based on LR criteria rather than evidence strength criteria, and examined whether task difficulty affected the decision criteria on objective LR axis.

Method

Participants

Participants in Experiment 3 were recruited online via the Prolific website. We collected data from 50 participants (27 female and 23 male; age: $M = 30.20$ years, $SD = 5.46$), the same sample size as in Experiments 1 and 2. Data from another seven participants were excluded because the value of A_{trap} was lower than 0.5 in either easy (with study time as 90 s) or difficult (with study time as 30 s) condition, suggesting their performance was worse than chance level. Participants signed informed consent before the experiment and received monetary compensation (£4) and a bonus (up to £0.5, dependent on task performance) after the experiment. All participants spoke English as the first language and reported normal or corrected-to-normal vision. All procedures were approved by the Ethics Committee at Beijing Normal University.

Materials and Procedure

The experimental materials for the Old/New recognition memory test in Experiment 3 were 200 English words that were randomly chosen from the materials in Experiment 2. The 200 words were randomly divided into two halves for each participant, in which 100 words were used as study materials in the learning phase, with the other 100 words served as lures in the test phase.

The Old/New recognition test contained two blocks with a learning phase and a test phase in each block. The procedure for the learning phase in Experiment 3 was the same as that in Experiment 2. The study time of the words was 90 s for one block (easy condition) and 30 s for the other block (difficult condition). The block order was randomized across participants. In each trial of the test phase (see Figure 9C), after presenting the fixation for 500 ms, one word was shown on the screen, and participants needed to decide whether the presented word was an old or new word. There were 100 trials in the test phase of each block, including 50 old words and 50 new words. All of the words were randomly divided into the two blocks. Participants were asked to press a number key from 1 to 6 to make decision and rate their confidence. The six points on the confidence rating scale were 1 = *sure new*, 2 = *probably new*, 3 = *maybe new*, 4 = *maybe old*, 5 = *probably old*, and 6 = *sure old*. The chosen number would then be highlighted for 500 ms.

Data Analysis

Model Fitting. The confidence rating data in Experiment 3 was fitted by unequal variance SDT model in which the variance differed between signal (old) and noise (new) distributions. Furthermore, the model assumed that participants used the perceived evidence strength in each trial to compute the value of objective LR, and made decisions based on the decision criteria on the objective LR axis. There are seven free parameters in the model, including the distance between signal and noise distributions (d'), the five decision criteria on the objective log LR (λ) axis (λ_1, λ_2, λ_3, λ_4, and λ_5 from left to right), and the standard deviation of the signal distribution (σ_S). The standard deviation of the noise distribution was fixed to 1 (Glanzer et al., 2009; Wickens, 2002). We fit the model to data separately in each experimental condition for each participant and found the parameter values that maximized the likelihood function (see Section S5 in the online supplemental materials for the calculation of likelihood function in unequal variance SDT). As in Experiments 1 and 2, here we performed a padding correction before model fitting and all of the following data analyses.

Comparing Task Performance Between Two Conditions. In unequal variance SDT, it is inappropriate to directly use the estimated value of d' to reflect task performance due to variation in σ_S across experimental conditions and participants (Wickens, 2002). Thus, in Experiment 3, we used two measures of the area under the ROC curve to represent performance in the recognition memory test, including the parametric measure A_z and the nonparametric measure A_{trap}. The A_{trap} is computed as in Experiments 1 and 2. On the other hand, A_z is approximated based on the estimated parameter values (including d' and σ_S) in the unequal variance SDT model (see Section S5 in the online supplemental materials for details). We performed paired-sample t tests to examine whether the difference in the two measures of performance between easy and difficult conditions was statistically detectable.

Other Data Analyses. All of the other analyses, including the investigation of three regularities, and the comparison of objective λ criteria between easy and difficult conditions, were the same as those in Experiments 1 and 2.

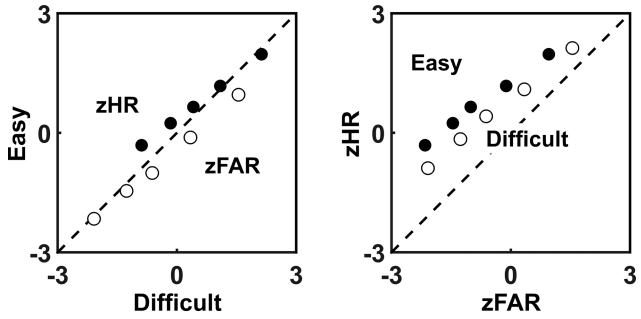
Results and Discussion

The value of A_{trap} was reliably greater when study time was 90 s ($M = 0.80$, $SD = 0.11$) than 30 s ($M = 0.71$, $SD = 0.10$), $t(49) = 5.74$, $p < .001$, Cohen's $d = 0.81$, $BF_{10} > 100$. Similarly, A_z was also higher with 90 s ($M = 0.82$, $SD = 0.11$) than 30 s ($M = 0.73$, $SD = 0.11$) study time, $t(49) = 5.99$, $p < .001$, Cohen's $d = 0.85$, $BF_{10} > 100$. These results indicated that the difficulty manipulation was successful.

The $zHR_{easy}/zHR_{difficult}$, $zFAR_{easy}/zFAR_{difficult}$, and $zHR/zFAR$ plots averaged across participants are shown in Figure 13, in

Figure 13

The $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ Plots (Left) and zHR/zFAR Plots (Right) Averaged Across Participants in Experiment 3

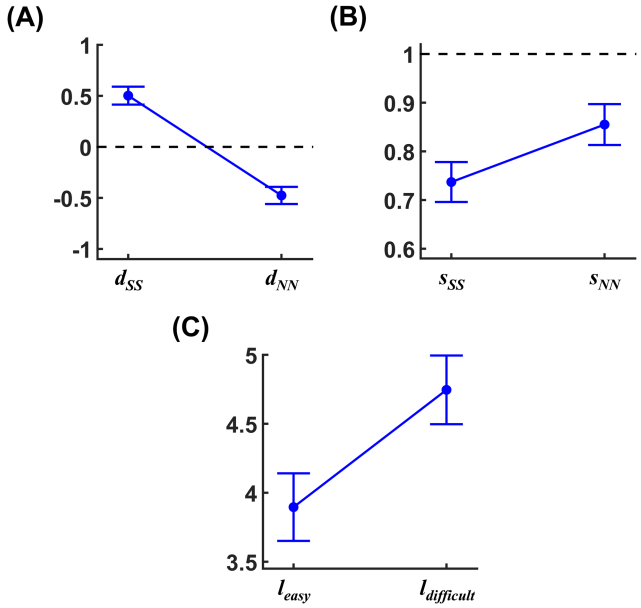


Note. HR = hit rate; FAR = false alarm rate.

which all of the three regularities were replicated. Specifically, the distance mirror index was reliably higher than 0 for the $zHR_{easy}/zHR_{difficult}$ plot (d_{SS}), $t(49) = 5.69$, $p < .001$, Cohen's $d = 0.81$, $BF_{10} > 100$, and lower than 0 for the $zFAR_{easy}/zFAR_{difficult}$ plot (d_{NN}), $t(49) = -5.69$, $p < .001$, Cohen's $d = -0.81$, $BF_{10} > 100$, replicating the mirror effect (see Figure 14A). In addition, the slopes of both the $zHR_{easy}/zHR_{difficult}$ plot (s_{SS}) and $zFAR_{easy}/zFAR_{difficult}$ plot (s_{NN}) were reliably lower than 1, s_{SS}: $t(49) = -6.40$, $p < .001$, Cohen's $d = -0.91$, $BF_{10} > 100$; s_{NN}: $t(49) = -3.49$, $p = .001$, Cohen's $d = -0.49$, $BF_{10} = 27.68$, replicating the variance effect (see Figure 14B). Furthermore, the distance between end points of the zHR/zFAR plot was longer in difficult (30 s study time) than easy (90 s study time) condition, $t(49) = 4.99$, $p < .001$, Cohen's $d = 0.71$, $BF_{10} > 100$, replicating the zROC length effect (see Figure 14C).

Figure 14

The Three Regularities in Experiment 3



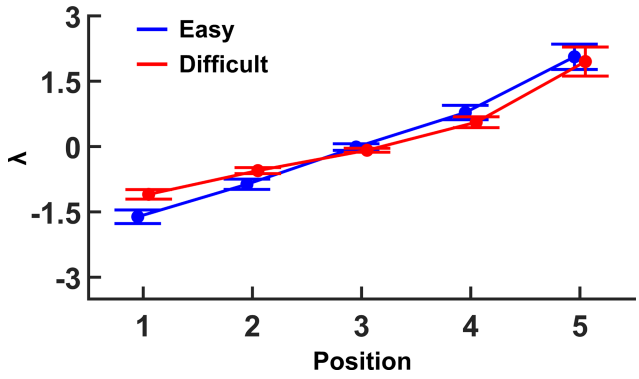
Note. (A) The distance mirror index for the $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ plots (d_{SS} and d_{NN}). (B) The slope of the $zHR_{easy}/zHR_{difficult}$ and $zFAR_{easy}/zFAR_{difficult}$ plots (s_{SS} and s_{NN}). (C) The distance between end points of the zHR/zFAR plot in easy and difficult condition (l_{easy} and $l_{difficult}$). Error bars represent standard errors. HR = hit rate; FAR = false alarm rate. See the online article for the color version of this figure.

Finally, we performed a 2 (condition) × 5 (position) repeated measures ANOVA on the estimated value of the five λ criteria. The main effect of position was statistically significant, $F(1.25, 61.18) = 64.87$, $p < .001$, $\eta_p^2 = .57$, $BF_{incl} > 100$ (Greenhouse–Geisser corrected), indicating that the value of λ for the decision criteria increased along the axis from left to right. However, the main effect of condition did not reach statistical significance, $F(1, 49) = 1.69$, $p = .200$, $\eta_p^2 = .03$, $BF_{incl} = 0.23$, suggesting that the overall value of λ criteria did not differ between easy and difficult conditions. Most importantly, there was a reliable interaction effect between condition and position, $F(2.02, 99.19) = 5.55$, $p = .005$, $\eta_p^2 = .10$, $BF_{incl} > 100$ (Greenhouse–Geisser corrected). We then conducted a 2 (condition) × 3 (position) ANOVA on the middle three λ criteria. Similarly, we found that the main effect of condition did not reach significance, $F(1, 49) = 0.01$, $p = .916$, $\eta_p^2 < .01$, $BF_{incl} = 0.16$. In addition, there were reliable main effect of position and interaction effect between condition and position, $Fs > 12.61$, $ps < .001$, $\eta_p^2 > .20$, $BF_{incl} > 100$.

Further analyses revealed that λ was reliably lower in easy than difficult condition at Positions 1 and 2, $ts > 3.33$, $ps < .005$, Cohen's $d > 0.47$, $BF_{10} > 18.39$. At Positions 3–5, λ was numerically higher in easy than difficult condition. However, this difference did not reach statistical significance at Positions 3 and 5, $ts < 1.08$, $ps > .28$, Cohen's $d < 0.16$, $BF_{10} < 0.27$. Furthermore, λ was significantly higher in easy condition at Position 4 based on p value (although the Bayes factor was inconclusive), $t(49) = 2.46$, $p = .018$, Cohen's $d = 0.35$, $BF_{10} = 2.32$ (see Figure 15). These results suggest that the objective λ criteria tended to fan out in easy (compared with difficult) condition, although to a less extent than that in Experiments 1 and 2 for the λ criteria in the middle position or on the right.

Figure 15

The λ Criteria as a Function of Experimental Condition (Easy vs. Difficult) and Position (From 1 to 5) in Experiment 3



Note. Error bars represent standard errors. See the online article for the color version of this figure.

In brief, as in Experiments 1 and 2, the three regularities were observed in Experiment 3, and the decision criteria on the objective log LR axis fanned out in the easy condition. Thus, Experiment 3 replicated the main results of the first two experiments, and extended these findings to unequal variance SDT.

General Discussion

Many previous studies suggest that people set decision criteria in signal detection tasks based on the ratio of the likelihood that each stimulus is signal or noise, rather than directly on the evidence strength (Glanzer et al., 2009, 2019; Hilford et al., 2015, 2019; Osth et al., 2017; Semmler et al., 2018; Stretch & Wixted, 1998a; Wixted & Gaitan, 2002). Glanzer et al. (2009, 2019) indicate that the three regularities, including the mirror effect, variance effect and *z*ROC length effect, should be observed when the LR theory holds. Although the three regularities do exist in previous published studies from different cognitive domains (Glanzer et al., 2009, 2019), other researchers have questioned whether people are really able to compute the objective LR of the true signal and noise distributions in a certain signal detection task (Balakrishnan & Ratcliff, 1996; Criss & McClelland, 2006; Hintzman, 1994; McClelland & Chappell, 1998), which is the underlying assumption according to which the three regularities are derived by Glanzer and colleagues. Furthermore, few studies have conducted empirical experiments to directly examine whether people's subjectively estimated LR is the same as or different from the objective LR.

To address this issue, the current study suggests that whether people are able to accurately estimate objective LR can be tested in two-condition experiments. According to the constant LR account, people's decision criteria on the axis of LR should only depend on the evaluation of prior probabilities and gain functions, and thus should not be affected by task difficulty (Glanzer et al., 2019). Thus, we can estimate the value of objective LR corresponding to the decision criteria in experimental conditions with various difficulty levels, and compare the objective LR criteria across conditions. If the objective LR criteria remain constant across levels of task difficulty, then we can conclude that people rely on objective LR to make decisions. Otherwise (i.e., if the objective LR criteria differs between conditions), it is highly possible that people's decisions are not based on objective LR.

In the current study, we first demonstrate that the three regularities actually hold even when the objective LR criteria are different across difficulty levels. Thus, observing the three regularities in empirical experiments is insufficient to conclude whether people's decision criteria depend on objective LR, and it should be better to separately estimate and then compare the objective LR criteria in each experimental condition. We then describe three experiments in which participants performed either 2AFC perceptual tasks, 2AFC recognition memory test or Old/New recognition test with two levels of difficulty. Results revealed that the objective LR corresponding to decision criteria reliably differed between easy and difficult conditions in Experiments 1–3, and fanned out as task difficulty decreased. These results suggest that people might set decision criteria based on subjectively estimated LR that was systematically different from objective LR. Then one interesting question is: how and why does subjective LR deviate from objective LR?

In the following sections, we first discuss the relationship between the current results and previous studies about people's distortion of probability, based on which we provide a hypothesis about how the subjective estimation of LR was distorted across levels of task difficulty in this study. Next, we propose several possible theoretical explanations to account for why the distortion of LR occurred. Finally, we describe the limitations of the current study.

Distortion of Probability

Previous studies suggest that people's estimation of the probability or frequency that an event happens is often distorted. For example, many studies show that people overestimate the probability of an event when the true probability is low, and underestimate the probability that is high (e.g., Attneave, 1953; Gonzalez & Wu, 1999; Ren et al., 2018; Ungemach et al., 2009; Varey et al., 1990; Zhang & Maloney, 2012). However, the opposite pattern with underestimation of low probability and overestimation of high probability has also been observed (Gigerenzer et al., 1991; McGraw et al., 2004). Researchers have shown that the relationship between subjectively estimated probability and objective probability can be well fit by the linear in log odds (LLO) function (Gonzalez & Wu, 1999; Zhang & Maloney, 2012). The LLO function suggests that subjective estimation of the logarithm of odds, which refer to the ratio of the probability that an event happens to the probability that it does not happen, is a linear function of the objective log odds. When the slope of the LLO function is lower than 1, people overestimate low probability and underestimate high probability. The opposite pattern occurs when the slope is higher than 1 (Zhang & Maloney, 2012).

Here we assume that the imperfect estimation of objective LR in the current study may also be characterized by the LLO function, because log LR (λ) can be seen as the logarithm of the ratio of two probabilities (that evidence strength occurs in signal and noise distributions), which is similar to log odds. The linear relationship between subjectively estimated log LR (λ_{subj}) and objective log LR (λ_{obj}) can be written as Gonzalez & Wu (1999):

$$\lambda_{subj} = \gamma \cdot \lambda_{obj} + \tau$$

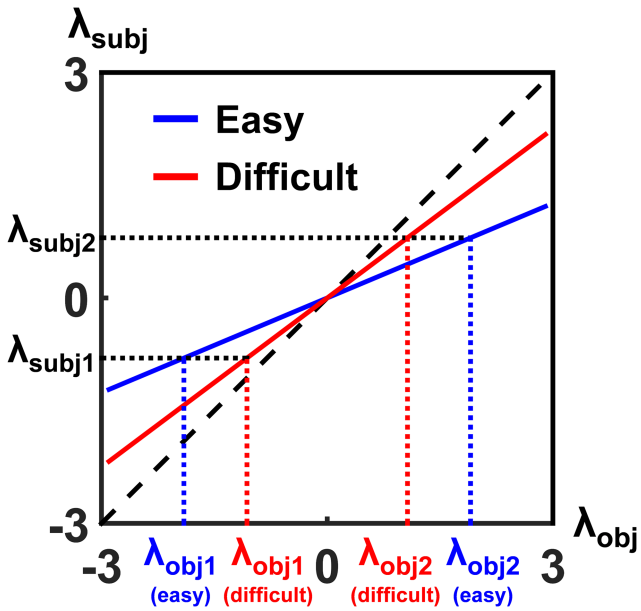
in which γ and τ are the slope and intercept of the LLO function, respectively. When $\gamma < 1$, people overestimate low values of λ_{obj} and underestimate high values of λ_{obj}. The opposite pattern occurs when $\gamma > 1$.

Results from the current study could not indicate whether γ was higher or lower than 1 for experimental conditions with various levels of difficulty. However, the direction of the difference in γ between conditions could be determined. To explain the fact that the objective λ criteria fanned out in easy (compared with difficult) condition, the slope of the LLO function should be smaller for easy than difficult condition. An example is shown in Figure 16, in which people set two decision criteria on the axis of subjectively estimated log LR. When LLO function has a smaller slope in easy than difficult condition, the distance between the values of objective log LR corresponding to the two criteria is longer in easy condition.

To test whether the relationship between objective and subjective log LR in the current study could be characterized by the LLO function, we performed a linear regression analysis on the relationship between the estimated five objective log LR criteria in easy and difficult conditions. Because each criteria represented a certain value of subjective log LR which should be equal across the conditions, the

Figure 16

Using the LLO Function to Illustrate the Expansion of Objective Log LR Criteria in Easy (Compared With Difficult) Condition



Note. LLO = linear in log odds; LR = likelihood ratio. See the online article for the color version of this figure.

relationship between the objective LR corresponding to decision criteria in the two conditions should also follow a linear function:

$$\gamma_{easy} \cdot \lambda_{obj(easy)} + \tau_{easy} = \gamma_{difficult} \cdot \lambda_{obj(difficult)} + \tau_{difficult},$$

$$\lambda_{obj(easy)} = \frac{\gamma_{difficult}}{\gamma_{easy}} \cdot \lambda_{obj(difficult)} + \frac{\tau_{difficult} - \tau_{easy}}{\gamma_{easy}},$$

$$\lambda_{obj(difficult)} = \frac{\gamma_{easy}}{\gamma_{difficult}} \cdot \lambda_{obj(easy)} + \frac{\tau_{easy} - \tau_{difficult}}{\gamma_{difficult}}.$$

.

When the objective log LR in easy condition was regressed on that in difficult condition, the regression slope should be higher than 1, because the slope γ for the LLO function should be smaller in easy condition. In contrast, the slope should be lower than 1 when the objective log LR in difficult condition was regressed on that in easy condition. Furthermore, the R^2 should be high for both regression analyses.

We conducted the two linear regression analyses above separately for each participant in each experiment. Results showed that the mean R^2 across participants was higher than 0.88 in all of the three experiments, suggesting the relationship between the estimated objective log LR criteria in easy and difficult conditions could be characterized by linear function. In addition, one-sample Wilcoxon signed rank tests revealed that when the objective log LR in easy condition was regressed on that in difficult condition, the overall value of slope across participants was reliably higher than 1 in Experiments 1–3, $bs > 1.58$, $Zs > 2.98$, $ps < .005$, $BF_{10} > 60$. When the objective log LR in difficult condition was regressed on that in easy condition, the overall slope was significantly lower than 1 based on p value in the three experiments (although the Bayes factor was inconclusive in Experiment 3), $bs < 0.96$, $Zs < -2.07$, $ps < .05$; $BF_{10} > 10$ in Experiments 1–2, and $BF_{10} = 0.52$ in Experiment 3. These results indicate that the subjective log LR might be a linear function of objective log LR, as suggested by the LLO function. Furthermore, the slope γ of the LLO function should be smaller for easy than difficult condition.

Previous studies have provided several theories concerning why people's estimation of the probability that an event happens is distorted (for a review, see Zhang & Maloney, 2012). Based on these theories, here we propose two possible theoretical explanations to account for the distortion of LR in the current study, including the calibration model and the Bayesian weighting model. In addition, we also provide another possible explanation based on the strength theory.

Explanations for the Distortion of LR

The first explanation for the distortion of LR is based on the calibration model, which suggests that people may inaccurately estimate their ability to discriminate signal and noise stimuli (Smith & Ferrell, 1983). Previous studies have mathematically proved that when the distance between signal and noise distributions (d') is longer, the value of log LR increases faster as evidence strength x increases (Glanzer et al., 2009, 2019; see also Section S2 in the online supplemental materials). Thus, the slope γ for the LLO function is higher than 1 when the subjectively estimated d' is higher than true d'. In contrast, γ is lower than 1 when people underestimate the true value of d' (see also Smith & Ferrell, 1983; Zhang & Maloney, 2012). Thus, γ should be smaller for easy than difficult condition in the current study when participants overestimated the true d' to a higher degree in difficult than easy condition, or when the extent of underestimation of d' was higher in the easy condition.

The calibration model is consistent with Lau's (2007) discussion about BDT, which indicates that when people infer the posterior probability that each stimulus comes from signal or noise category, their subjective estimation of the shape of signal and noise distributions, such as the mean and variance, may differ from the true shape. For example, when people overestimate the mean of the signal distribution, they may set a decision criterion that is more conservative than the optimal criterion (Lau, 2007). Furthermore, the calibration model could explain the setting of decision criteria when signal and noise distributions completely overlap (Snodgrass & Corwin, 1988). If decision criteria are based on the objective LR, then the stimuli with different evidence strength cannot be discriminated when signal and noise distributions are the same, because objective LR should be always equal to 1. In contrast, the subjective LR could change with evidence strength when people's estimation of d' is different from the objective value 0, and thus the decision criteria on the axis of subjective LR could be set. However, one limitation of the calibration model is that it is relatively difficult to further explain why the calibration of d' differs across signal detection tasks with various levels of task difficulty.

The second explanation is named as the Bayesian weighting model. The basic idea of this model is that even when people are able to accurately estimate the objective ratio of the likelihood that the observed evidence strength occurs in the true signal and noise distributions, this estimation may still be associated with uncertainty

due to the existence of random sampling error (Martins, 2006). For example, the observed evidence strength in a signal detection task may be the combination of the true evidence strength and a random noise, which could make the estimation of LR deviate from the true LR even if people have exact knowledge about the shape of the signal and noise distributions (Pouget et al., 2016). To reduce the influence of sampling error on the estimation of LR, people may hold a prior belief about the value of LR, and integrate this prior belief with the observed LR using a Bayesian inference process (Martins, 2006).

In Section S6 in the online supplemental materials, we offer a specific version of the Bayesian weighting model in equal variance SDT, which assumes that the random noise in observed evidence strength comes from a normal distribution with a mean of 0, and people's prior belief about LR is distributed as another normal distribution. It is hypothesized that people first compute the value of LR based on the observed evidence strength, and then infer the posterior distribution of LR, which forms the basis of their decisions in signal detection tasks, through Bayesian inference. In this case, the posterior mean of LR is a weighted average of the observed LR and the mean of prior LR, and thus is a linear function of the observed LR, as suggested by the LLO function. The weight for the observed LR (which is the slope γ for the LLO function) increases as the uncertainty in observed LR decreases (Ma, 2019).

This specific Bayesian weighting model is in accordance with the support theory for the distortion of probability, which suggests that the subjective estimated logarithm of the odds that an event occurs is a weighted average of the objective log odds and a prior (Fox & Rottenstreich, 2003; See et al., 2006). Furthermore, an advantage of this model is that it can naturally explain the difference in the slope γ between difficulty levels: when the variance of random noise in observed evidence strength remains constant in easy and difficult conditions, the variance of the sampling error (i.e., uncertainty) for log LR is higher for easy condition (Glanzer et al., 2009, 2019; see also Figure 2B). Thus, people should rely more on the prior belief about LR to infer the posterior mean in the easy condition, leading to smaller slope in the LLO function (see Section S6 in the online supplemental materials for details). However, one limitation of the Bayesian weighting model is that it suggests γ should be lower than 1, and is not able to account for previous empirical results with $\gamma > 1$ (Zhang & Maloney, 2012). For example, Kubovy's (1977) study about the misconception of likelihood suggests that participants underestimated low log LR and overestimated high log LR, indicating a LLO function with $\gamma > 1$ and thus could not be solely explained by the Bayesian weighting model.

We should note that the calibration model and the Bayesian weighting model are not mutually exclusive. That is, people may both incorrectly estimate the distance between the signal and noise distributions, and base their decisions on the posterior mean of log LR which is a combination of the prior belief about LR and the computed (and perhaps inaccurate) LR using the observed evidence strength.

Both the calibration model and the Bayesian weighting model are on the basis of the LR theory, which suggests that people's decisions in signal detection tasks should depend on the (although distorted) computation of LR. In contrast, another explanation for the current results that cannot be fully ruled out is based on the strength theory, which indicates that people might directly set decision criteria on the axis of evidence strength without computing LR (Balakrishnan & Ratcliff, 1996). As we have illustrated above, the three regularities can be obtained as long as the decision criteria on the evidence strength axis fan out as task difficulty increases. In fact, if we estimate the decision criteria on the strength axis in Experiments 1–3 rather than on the objective LR axis, we could find that the evidence strength criteria tended to fan out in the difficult (compared with easy) condition (see Section S7 in the online supplemental materials). Thus, it is possible that participants only needed to know they should be more cautious when giving high-confidence ratings in more difficult tasks to avoid high-confidence erroneous response, and set more conservative confidence criteria on the strength axis accordingly in the difficult condition without computing the likelihood that the evidence strength occurred in the signal and noise distributions (Mickes et al., 2011; Stretch & Wixted, 1998a).

However, a limitation of the strength theory is that it is relatively difficult to further explain why the evidence strength criteria fanned out in the difficult condition just to the extent that the corresponding objective LR criteria contracted compared with the easy condition. In the current study, we have demonstrated that evidence strength criteria can fan out in the difficult condition when the corresponding criteria on the objective LR axis either expand, contract or remain constant. However, the fact that objective LR criteria contracted in the difficult condition was consistently observed in the three experiments. If participants were motivated to avoid high-confidence error, as suggested by previous studies (Mickes et al., 2011; Stretch & Wixted, 1998a), then it is also possible that they might set very conservative decision criteria on the strength axis in the difficult condition, which may be more conservative (rather than liberal) than those with constant objective LR. This hypothesis was not confirmed in the current study. Furthermore, the strength theory also has difficulty in explaining why the change of evidence strength criteria across difficulty levels leads to a linear relationship between the decision criteria on the objective log LR axis in easy and difficult conditions, as revealed by our data analysis above.

In summary, none of the theoretical explanations described above could perfectly account for the distortion of LR across various levels of task difficulty. Future studies are required to further distinguish these possible theories and examine whether the imperfect estimation of LR is explained by the combination of more than one theory. Besides the perceptual and memory tasks implemented in this study, future studies are recommended to design new psychophysical tasks in which the true evidence strength could be better measured. Process models could also be developed to interpret the psychological content underlying the subjective estimation of LR (Osth & Dennis, 2015; Turner et al., 2011). For example, Osth and Dennis (2015) propose a quantitative model of recognition memory, in which memory strength is decomposed into a number of relevant features such as interference among items and contexts, and people are assumed to use a LR transformation of memory strength. It will be of interest to further explore how subjective LR is formed based on people's evaluation of features in memory storage. In addition, we encourage future studies to investigate whether the results in this study could also be explained in the framework of strength theory.

The current study is not the first to point out that individuals' behavior in signal detection tasks does not exactly match the LR model assuming decisions depend on objective LR (Stretch & Wixted, 1998a; Wixted & Gaitan, 2002). However, we are among the first to indicate that whether people base their decisions on

objective or subjective LR is unrelated to the occurrence of the three regularities. Previous studies on the LR theory often derive the three regularities on the basis of the assumption that people's decisions rely on objective LR, and then reversely infer that observing the regularities in empirical experiments support the utilization of objective LR in decision process (Glanzer et al., 2009, 2019). Our results suggest that this inference is inappropriate. Furthermore, we provide several possible explanations about why subjective LR deviated from the objective LR in a manner that could be characterized by the LLO function. We encourage future studies to use a wide range of experimental paradigms to replicate the empirical results obtained in this study and test the theoretical explanations.

Limitation of the Current Study

In the current study, we hypothesized that the evidence strength for both signal and noise stimuli was distributed as normal distribution, which is the assumption of normative SDT (Wickens, 2002; Wixted, 2020). However, it is also possible that the evidence strength in the tasks actually followed other distributions, such as the binomial distribution or exponential distribution (Glanzer et al., 2009). Furthermore, the ROC curve obtained in empirical data could not exactly reveal the form of the underlying signal and noise distributions (Glanzer et al., 2009). Previous study has shown that the estimation of LR in empirical experiments may be biased when the shape of the distributions assumed in the SDT model is different from that of the true distributions (Maloney & Thomas, 1991). Future studies should examine whether the expansion of objective LR criteria in easy tasks can also be observed in the SDT with other types of distributions.

Furthermore, the current study, as well as many previous studies about the setting of decision criteria in SDT (Glanzer et al., 2009, 2019; Semmler et al., 2018; Stretch & Wixted, 1998a), assumed that people rated confidence for their decisions directly based on the evidence strength for each stimulus. However, recent studies about metacognition suggest that the information utilized in decision making and confidence rating process may be different. After deciding whether each stimulus belongs to signal or noise category, people may lose some information when they subsequently rate confidence about their decision, or obtain more information during the confidence rating process through accumulation of additional evidence (Maniscalco & Lau, 2012; Pleskac & Busemeyer, 2010). Some computational models about metacognition have extended the classical SDT to account for the difference in evidence strength used for decision making and confidence rating (e.g., Fleming & Daw, 2017; Hu et al., 2021; Mamassian & de Gardelle, 2022; Shekhar & Rahnev, 2021), which should be considered by future studies concerning how people set decision and confidence criteria.

In Experiments 1 and 2, we assumed that participants rated confidence about their decision in 2AFC tasks using the difference in evidence strength between the two stimuli presented in the same trial (Wickens, 2002). However, recent studies have proposed the winner-take-all rule which argues that when rating their confidence in 2AFC tasks, people may solely rely on the evidence strength of the stimulus they have chosen in each trial, rather than on the difference in strength between the two stimuli (Hanczakowski et al., 2021; Miyoshi & Lau, 2020; Peters et al., 2017). Although the winner-take-all rule has been supported by some empirical and theoretical studies, it is based on the strength theory suggesting that people set confidence criteria directly on the axis of evidence strength, and thus is relatively difficult to explain how they adjust evidence strength criteria across levels of task difficulty to produce the three regularities observed in the current study. Future studies should try to combine the winner-take-all rule and the LR theory to account for the three regularities in 2AFC tasks.

Conclusion

Although the three regularities, including the mirror effect, variance effect and *z*ROC length effect, were observed in signal detection tasks that came from different cognitive domains (perception and memory) and were characterized by either equal or unequal SDT model, the decision criteria on the axis of objective LR reliably differed between difficulty levels and fanned out in easy (compared with difficult) condition. Thus, observing the three regularities is insufficient to conclude whether the decision criteria depend on objective LR. Furthermore, it is highly possible that people's subjectively estimated LR used for decision making should differ from objective LR, because relying on objective LR to make decisions would violate the constant LR account.

References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology*, *46*(2), 81–86. https://doi.org/10.1037/h0057955

Balakrishnan, J. D., & Ratcliff, R. (1996). Testing models of decision making using confidence ratings in classification. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(3), 615–633. https://doi.org/10.1037/0096-1523.22.3.615

Bruno, D., Higham, P. A., & Perfect, T. J. (2009). Global subjective memorability and the strength-based mirror effect in recognition memory. *Memory and Cognition*, *37*(6), 807–818. https://doi.org/10.3758/MC.37.6.807

Burgess, A. (1985). Visual signal detection. III. On Bayesian use of prior knowledge and cross correlation. *Journal of the Optical Society of America A: Optics and Image Science*, *2*(9), 1498–1507. https://doi.org/10.1364/JOSAA.2.001498

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33*(4), 497–505. https://doi.org/10.1080/14640748108400805

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language*, *55*(4), 447–460. https://doi.org/10.1016/j.jml.2006.06.003

DeCarlo, L. T. (2007). The mirror effect and mixture signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(1), 18–33. https://doi.org/10.1037/0278-7393.33.1.18

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Fleming, S. M. (2017). HMeta-d: Hierarchical Bayesian estimation of metacognitive efficiency from confidence ratings. *Neuroscience of Consciousness*, *2017*(1), Article nix007. https://doi.org/10.1093/nc/nix007

Fleming, S. M., & Daw, N. D. (2017). Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological Review*, *124*(1), 91–114. https://doi.org/10.1037/rev0000045

Fleming, S. M., Ryu, J., Golfinos, J. G., & Blackmon, K. E. (2014). Domain-specific impairment in metacognitive accuracy following anterior prefrontal lesions. *Brain: A Journal of Neurology*, *137*(10), 2811–2822. https://doi.org/10.1093/brain/awu221

Fox, C. R., & Rottenstreich, Y. (2003). Partition priming in judgment under uncertainty. *Psychological Science*, *14*(3), 195–200. https://doi.org/10.1111/1467-9280.02431

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, *91*(1), 1–67. https://doi.org/10.1037/0033-295X.91.1.1

Glanzer, M., & Adams, J. K. (1985). The mirror effect in recognition memory. *Memory and Cognition*, *13*(1), 8–20. https://doi.org/10.3758/BF03198438

Glanzer, M., Adams, J. K., & Iverson, G. (1991). Forgetting and the mirror effect in recognition memory: Concentering of underlying distributions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(1), 81–93. https://doi.org/10.1037/0278-7393.17.1.81

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 21–31. https://doi.org/10.1037/0278-7393.2.1.21

Glanzer, M., Hilford, A., Kim, K., & Maloney, L. T. (2019). Generality of likelihood ratio decisions. *Cognition*, *191*, Article 103931. https://doi.org/10.1016/j.cognition.2019.03.023

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin and Review*, *16*(3), 431–455. https://doi.org/10.3758/PBR.16.3.431

Gonzalez, R., & Wu, G. (1999). On the shape of the probability weighting function. *Cognitive Psychology*, *38*(1), 129–166. https://doi.org/10.1006/cogp.1998.0710

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley.

Hanczakowski, M., Butowska, E., Philip Beaman, C., Jones, D. M., & Zawadzka, K. (2021). The dissociations of confidence from accuracy in forced-choice recognition judgments. *Journal of Memory and Language*, *117*, Article 104189. https://doi.org/10.1016/j.jml.2020.104189

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, and Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Higham, P. A., Perfect, T. J., & Bruno, D. (2009). Investigating strength and frequency effects in recognition memory using type-2 signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 57–80. https://doi.org/10.1037/a0013865

Hilford, A., Glanzer, M., Kim, K., & Maloney, L. T. (2019). One mirror effect: The regularities of recognition memory. *Memory and Cognition*, *47*(2), 266–278. https://doi.org/10.3758/s13421-018-0864-y

Hilford, A., Maloney, L. T., Glanzer, M., & Kim, K. (2015). Three regularities of recognition memory: The role of bias. *Psychonomic Bulletin and Review*, *22*(6), 1646–1664. https://doi.org/10.3758/s13423-015-0829-0

Hintzman, D. L. (1994). On explaining the mirror effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 201–205. https://doi.org/10.1037/0278-7393.20.1.201

Hirshman, E. (1995). Decision processes in recognition memory: Criterion shifts and the list-strength paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(2), 302–313. https://doi.org/10.1037/0278-7393.21.2.302

Hu, X. (2023). *Likelihood ratio in signal detection theory*. https://osf.io/grcsj/

Hu, X., Yang, C., & Luo, L. (2023). Are the contributions of processing experience and prior beliefs to confidence ratings domain-general or domain-specific? *Journal of Experimental Psychology: General*, *152*(1), 28–44. https://doi.org/10.1037/xge0001257

Hu, X., Zheng, J., Su, N., Fan, T., Yang, C., Yin, Y., Fleming, S. M., & Luo, L. (2021). A Bayesian inference model for metamemory. *Psychological Review*, *128*(5), 824–855. https://doi.org/10.1037/rev0000270

Koop, G. J., Criss, A. H., & Pardini, A. M. (2019). A strength-based mirror effect persists even when criterion shifts are unlikely. *Memory and Cognition*, *47*(4), 842–854. https://doi.org/10.3758/s13421-019-00906-8

Kubovy, M. (1977). A possible basis for conservatism in signal detection and probabilistic categorization tasks. *Perception and Psychophysics*, *22*(3), 277–281. https://doi.org/10.3758/BF03199690

Lau, H. C. (2007). A higher order Bayesian decision theory of consciousness. In R. Banerjee, & B. Chakrabarti (Eds.), *Progress in brain research* (Vol. 168, pp. 35–48). Elsevier. https://doi.org/10.1016/S0079-6123(07)68004-2

Lynn, S. K., Wormwood, J. B., Barrett, L. F., & Quigley, K. S. (2015). Decision making from economic and signal detection perspectives: Development of an integrated framework. *Frontiers in Psychology*, *6*, Article 952. https://doi.org/10.3389/fpsyg.2015.00952

Ma, W. J. (2019). Bayesian decision models: A primer. *Neuron*, *104*(1), 164–175. https://doi.org/10.1016/j.neuron.2019.09.037

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology Press. https://doi.org/10.4324/9781410611147

Maloney, L. T., & Thomas, E. A. C. (1991). Distributional assumptions and observed conservatism in the theory of signal detectability. *Journal of Mathematical Psychology*, *35*(4), 443–470. https://doi.org/10.1016/0022-2496(91)90043-S

Maloney, L. T., & Zhang, H. (2010). Decision-theoretic models of visual perception and action. *Vision Research*, *50*(23), 2362–2374. https://doi.org/10.1016/j.visres.2010.09.031

Mamassian, P., & de Gardelle, V. (2022). Modeling perceptual confidence and the confidence forced-choice paradigm. *Psychological Review*, *129*(5), 976–998. https://doi.org/10.1037/rev0000312

Maniscalco, B., & Lau, H. (2012). A signal detection theoretic approach for estimating metacognitive sensitivity from confidence ratings. *Consciousness and Cognition*, *21*(1), 422–430. https://doi.org/10.1016/j.concog.2011.09.021

Martins, A. C. R. (2006). Probability biases as Bayesian inference. *Judgment and Decision Making*, *1*(2), 108–117. https://doi.org/10.1017/S1930297500002321

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760. https://doi.org/10.1037/0033-295X.105.4.734-760

McCurdy, L. Y., Maniscalco, B., Metcalfe, J., Liu, K. Y., de Lange, F. P., & Lau, H. (2013). Anatomical coupling between distinct metacognitive systems for memory and visual perception. *The Journal of Neuroscience*, *33*(5), 1897–1906. https://doi.org/10.1523/JNEUROSCI.1890-12.2013

McGraw, A. P., Mellers, B. A., & Ritov, I. (2004). The affective costs of overconfidence. *Journal of Behavioral Decision Making*, *17*(4), 281–295. https://doi.org/10.1002/bdm.472

Mickes, L., Hwe, V., Wais, P. E., & Wixted, J. T. (2011). Strong memories are hard to scale. *Journal of Experimental Psychology: General*, *140*(2), 239–257. https://doi.org/10.1037/a0023007

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin and Review*, *14*(5), 858–865. https://doi.org/10.3758/BF03194112

Miyoshi, K., & Lau, H. (2020). A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychological Review*, *127*(5), 655–671. https://doi.org/10.1037/rev0000184

Osth, A. F., & Dennis, S. (2015). Sources of interference in item and associative recognition memory. *Psychological Review*, *122*(2), 260–311. https://doi.org/10.1037/a0038692

Osth, A. F., Dennis, S., & Heathcote, A. (2017). Likelihood ratio sequential sampling models of recognition memory. *Cognitive Psychology*, *92*, 101–126. https://doi.org/10.1016/j.cogpsych.2016.11.007

Peters, M. A. K., Thesen, T., Ko, Y. D., Maniscalco, B., Carlson, C., Davidson, M., Doyle, W., Kuzniecky, R., Devinsky, O., Halgren, E., & Lau, H. (2017). Perceptual confidence neglects decision-incongruent evidence in the brain. *Nature Human Behaviour*, *1*(7), Article 139. https://doi.org/10.1038/s41562-017-0139

Pleskac, T. J., & Busemeyer, J. R. (2010). Two-stage dynamic signal detection: A theory of choice, decision time, and confidence. *Psychological Review*, *117*(3), 864–901. https://doi.org/10.1037/a0019737

Pouget, A., Drugowitsch, J., & Kepecs, A. (2016). Confidence and certainty: Distinct probabilistic quantities for different goals. *Nature Neuroscience*, *19*(3), 366–374. https://doi.org/10.1038/nn.4240

Ren, X., Wang, M., & Zhang, H. (2018). Context effects in the judgment of visual relative-frequency: Trial-by-trial adaptation and non-linear sequential effect. *Frontiers in Psychology*, *9*, Article 1691. https://doi.org/10.3389/fpsyg.2018.01691

See, K. E., Fox, C. R., & Rottenstreich, Y. S. (2006). Between ignorance and truth: Partition dependence and learning in judgment under uncertainty. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(6), 1385–1402. https://doi.org/10.1037/0278-7393.32.6.1385

Semmler, C., Dunn, J., Mickes, L., & Wixted, J. T. (2018). The role of estimator variables in eyewitness identification. *Journal of Experimental Psychology: Applied*, *24*(3), 400–415. https://doi.org/10.1037/xap0000157

Shekhar, M., & Rahnev, D. (2021). The nature of metacognitive inefficiency in perceptual decision making. *Psychological Review*, *128*(1), 45–70. https://doi.org/10.1037/rev0000249

Singer, M., & Wixted, J. T. (2006). Effect of delay on recognition decisions: Evidence for a criterion shift. *Memory and Cognition*, *34*(1), 125–137. https://doi.org/10.3758/BF03193392

Smith, M., & Ferrell, W. R. (1983). The effect of base rate on calibration of subjective probability for true–false questions: Model and experiment. In P. Humphreys, O. Svenson, & A. Vári (Eds.), *Analysing and aiding decision processes* (Vol. 14, pp. 469–488). North-Holland. https://doi.org/10.1016/S0166-4115(08)62251-7

Snodgrass, J. G., & Corwin, J. (1988). Pragmatics of measuring recognition memory: Applications to dementia and amnesia. *Journal of Experimental Psychology: General*, *117*(1), 34–50. https://doi.org/10.1037/0096-3445.117.1.34

Spanton, R. W., & Berry, C. J. (2020). The unequal variance signal-detection model of recognition memory: Investigating the encoding variability hypothesis. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, *73*(8), 1242–1260. https://doi.org/10.1177/1747021820906117

Starns, J. J., & Ratcliff, R. (2014). Validating the unequal-variance assumption in recognition memory using response time distributions instead of ROC functions: A diffusion model analysis. *Journal of Memory and Language*, *70*, 36–52. https://doi.org/10.1016/j.jml.2013.09.005

Stretch, V., & Wixted, J. T. (1998a). Decision rules for recognition memory confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1397–1410. https://doi.org/10.1037/0278-7393.24.6.1397

Stretch, V., & Wixted, J. T. (1998b). On the difference between strength-based and frequency-based mirror effects in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1379–1396. https://doi.org/10.1037/0278-7393.24.6.1379

Turner, B. M., Van Zandt, T., & Brown, S. (2011). A dynamic stimulus-driven model of signal detection. *Psychological Review*, *118*(4), 583–613. https://doi.org/10.1037/a0025191

Ungemach, C., Chater, N., & Stewart, N. (2009). Are probabilities overweighted or underweighted when rare outcomes are experienced (rarely)? *Psychological Science*, *20*(4), 473–479. https://doi.org/10.1111/j.1467-9280.2009.02319.x

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(3), 613–625. https://doi.org/10.1037/0096-1523.16.3.613

Verde, M. F., & Rotello, C. M. (2007). Memory strength and the decision process in recognition memory. *Memory and Cognition*, *35*(2), 254–262. https://doi.org/10.3758/BF03193446

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press.

Wixted, J. T. (2020). The forgotten history of signal detection theory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(2), 201–233. https://doi.org/10.1037/xlm0000732

Wixted, J. T., & Gaitan, S. C. (2002). Cognitive theories as reinforcement history surrogates: The case of likelihood ratio models of human recognition memory. *Animal Learning and Behavior*, *30*(4), 289–305. https://doi.org/10.3758/BF03195955

Zhang, H., & Maloney, L. T. (2012). Ubiquitous log odds: A common representation of probability and frequency distortion in perception, action, and cognition. *Frontiers in Neuroscience*, *6*, Article 1. https://doi.org/10.3389/fnins.2012.00001

Received October 13, 2022
Revision received April 4, 2023
Accepted April 17, 2023 ■