



EMPIRICAL ARTICLE

# When one graph judgment leads to another: Signal detection analysis of base rate effects

Ethan C. Guthrie  and Anthony J. Bishara 

Department of Psychology, College of Charleston, Charleston, SC, USA

**Corresponding author:** Anthony J. Bishara; Email: [BisharaA@cofc.edu](mailto:BisharaA@cofc.edu)

**Received:** 31 July 2024; **Revised:** 11 November 2024; **Accepted:** 2 March 2025

**Keywords:** graph judgment; signal detection theory; bias; discriminability; lens model

## Abstract

Graphs can help people arrive at data-supported conclusions. However, graphs might also induce bias by shifting the amount of evidence needed to make a decision, such as deciding whether a treatment had some kind of effect. In 2 experiments, we manipulated the early base rates of treatment effects in graphs. Early base rates had a large effect on a signal detection measure of bias in future graphs even though all future graphs had a 50% chance of showing a treatment effect, regardless of earlier base rates. In contrast, the autocorrelation of data points within each graph had a larger effect on discriminability. Exploratory analyses showed that a simple cue could be used to correctly categorize most graphs, and we examine participants' use of this cue among others in lens models. When exposed to multiple graphs on the same topic, human judges can draw conclusions about the data, but once those conclusions are made, they can affect subsequent graph judgment.

## 1. Introduction

Suppose you are a doctor prescribing a new drug to a patient. This patient has high cholesterol, and you worry the new drug may worsen this condition. So, you measure the patient's blood cholesterol daily before and after starting them on the drug, adding each measurement to a graph. After several days, you look at the graph and make your final decision; will you continue the patient on the drug, or try a different treatment? This was an example of an interrupted time series; you observed a variable (blood cholesterol) before and after an interruption (the drug). Interrupted time series are not exclusive to medicine, however. Whether to adopt a new sleep schedule, try a new exercise routine, or begin piano lessons, interrupted time series are ubiquitous. Even when experimental evidence is available regarding the average effect of the intervention, one must still make the important decision: did this drug work for *me*? Graphical presentation of data is one way to facilitate this type of decision and is used in areas ranging from education to health to public policy (Franconeri et al., 2021; Garcia-Retamero and Cokely, 2017). However, while graphs facilitate these sorts of decisions, they do not eliminate the unfair interpretation of information. Even when data are presented graphically, human judges remain more likely to identify trends that fit with their previous beliefs (Freedman and Smith, 1996; Lee and Lee, 2022; Mena, 2023; Nyhan and Reifler, 2019). Could such beliefs be induced by graphs themselves, and if so, how do such beliefs influence judgments of later graphs?

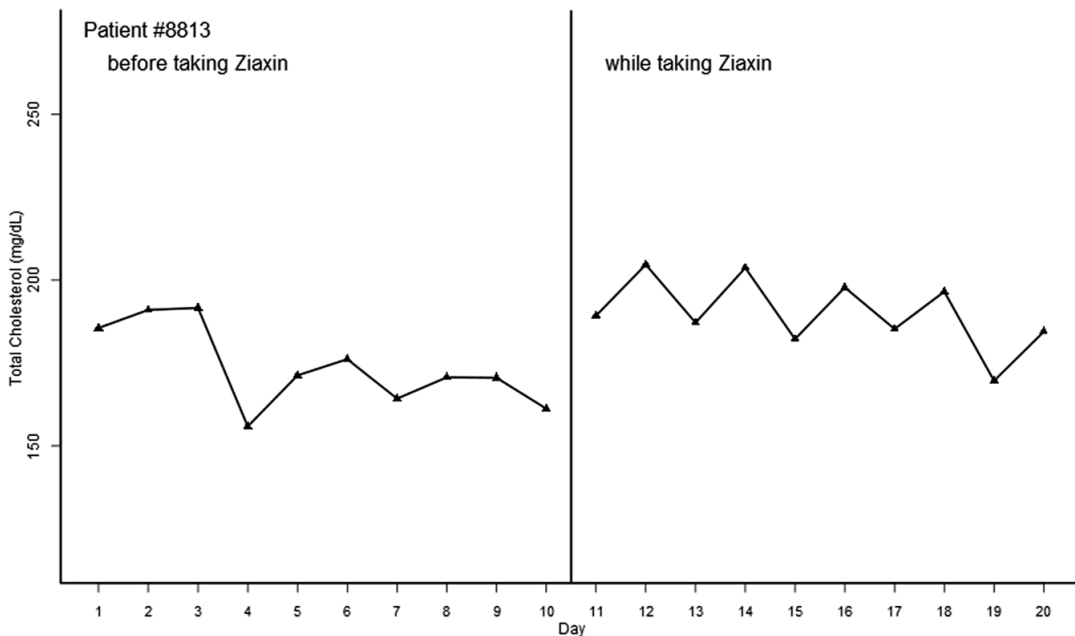


Figure 1. Example interrupted time series graph.

Table 1. The signal detection theory response matrix.

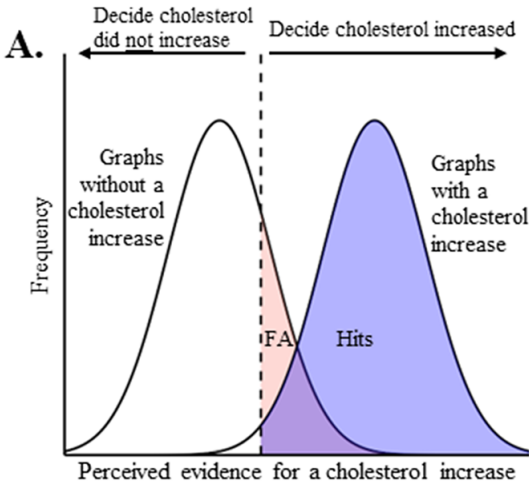
Base truth	Decision	
	Signal present	Signal absent
Signal present	Hits (7)	Misses (1)
Signal absent	False alarms (1)	Correct rejections (1)

Note: Parentheses indicate the number of responses by the hypothetical ‘first doctor’.

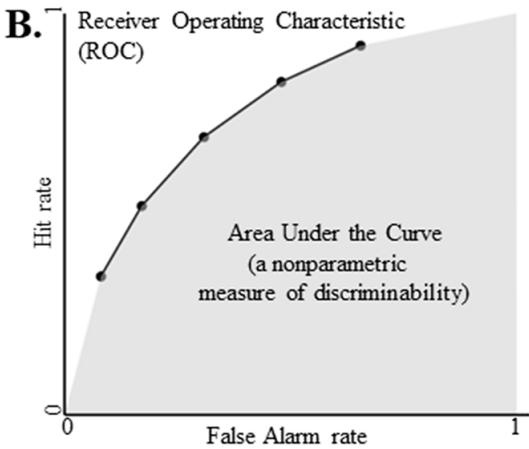
It is tempting to try to answer this question by assessing graph judgment via accuracy. However, the same accuracy can be achieved by vastly different means. To illustrate this, suppose a patient took a drug that may have increased their cholesterol, as shown in Figure 1. A doctor evaluates 10 graphs like this carefully and thereby achieves 80% accuracy. In contrast, a second doctor achieves the same accuracy by noting that cholesterol increases are common for this drug and—without evaluating any individual cases—summarily identifies all 10 cases as increases. Both doctors achieved the same accuracy, but the second doctor ignored the cases themselves. The second doctor’s background beliefs aided accuracy for the eight patients who truly had cholesterol increases but harmed accuracy for the 2 who did not. Accuracy fails to differentiate between true skill at detecting cholesterol increases and a predilection to favor a certain response because both contribute to accuracy.

In contrast, *Signal detection theory* provides a way to separate these 2 components: the ability to detect the cholesterol increase as discriminability, and a predilection in favor of that response as bias (e.g., Witt and Warden, 2021). To do so, as shown in Table 1, each response can be categorized by whether cholesterol truly increased and whether the doctor decided it increased. The first doctor had a hit rate of 0.88 because they correctly identified 7 of the 8 true increases, and a false-alarm rate of 0.50 because they falsely identified 1 of the 2 non-increases.

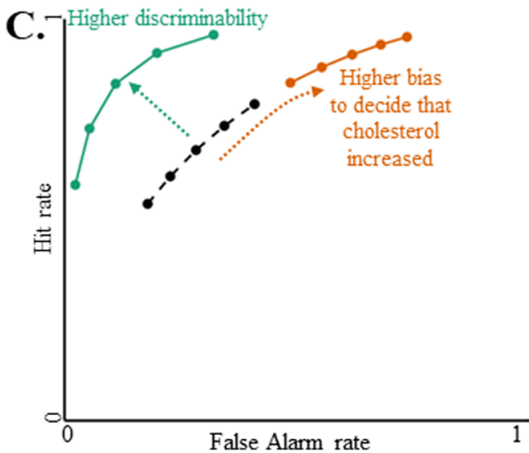
In a signal detection model, human judges decide that a signal is present when the perceived strength of evidence exceeds some threshold (see Figure 2A). The signal-detection framework considers



When judging each graph, participants decide whether the patient’s cholesterol increased after taking a new medication. If the perceived evidence for a cholesterol increase is higher than some mental threshold (dashed line), the decision is that cholesterol increased. When such an “increased cholesterol” decision is correct, it is a Hit. Otherwise, it is a False Alarm (FA). The actual experiments involved 6 response options, which could be represented by 5 thresholds.



In a Receiver Operating Characteristic (ROC) plot, the probabilities of Hits and False alarms are plotted as dots. Five dots are shown because participants responded on a 6-point Likert scale. The left-most dot represents the strictest threshold where only “Definitely an Increase” counted. The second dot from the left combines both “Probably...” and “Definitely an Increase” responses, and so on. When it is easy to discriminate between cholesterol increases and non-increases, Hits will be high, and False Alarms will be low, leading to curves near the upper left corner. The Area Under the Curve ( $A_g$ , the gray polygon) is a nonparametric measure of discriminability.



If a high proportion of early graphs show a cholesterol increase, it could lead participants to shift their thresholds to the left in panel A. This shift would result in an increase in both hits and false alarms, which would appear on an ROC plot as dots closer to the upper-right corner (panel C). Bias was quantified with the nonparametric measure  $\ln(B'_{12})$ .

**Figure 2.** A signal detection model of interrupted time-series graph judgment.

decisions to be the product of 2 separate phenomena: *discriminability* and *bias*. Generally, the purpose of decision aids such as graphs is to increase discriminability, or the extent to which signal and noise can be discriminated from one another (Witt and Warden, 2021). On the other hand, bias is represented by where the threshold is placed (vertical line in Figure 2A). The results of various thresholds are

summarized by a *receiver operating characteristic* (ROC) graph which plots the combination of hits and false alarms associated with each threshold (Peterson and Birdsall, 1953; see Figure 2B and 2C). When it is easier to discriminate between signal present and absent (between cholesterol increases and non-increases), hits will be high and false alarms will be low, leading to ROC points closer to the upper left. ROCs also highlight bias; a doctor who requires less evidence to believe a treatment works is biased in favor of treatment effects, leading to ROC points closer to the upper right in Figure 2C. The most common interpretation of such a bias is that it is due to shifting the threshold farther to the left in Figure 2A (in some situations, the same bias could also result from shifting both distributions in Figure 2A to the right by the same amount; Witt et al., 2015). Such a doctor will tend to generally decide that cholesterol increases more often, and so both hits and false alarms will be higher. The second doctor, who always decided cholesterol increased, is an extreme case of bias and would be represented by a dot exactly at the upper-right corner (1,1). Overall, unlike accuracy and related measures that conflate discriminability and bias, signal detection theory along with ROC plots disentangle these values (Gronlund et al., 2014).

Because bias is easier to manipulate than discriminability, comparisons based on accuracy are more likely to reflect changes in bias. Despite this, researchers could incorrectly conclude that variation in accuracy results from discriminability. For example, judges may be more likely to identify treatment effects as being present than absent in a variety of tasks, resulting in higher accuracy when effects are present (Bishara, Peller, et al., 2021). This does not, however, imply that judges are ‘better’ at identifying treatment effects; if judges were truly better at identifying when treatment effects are present, they would have to also be better at identifying when treatment effects are absent because effect presence/absence is a binary decision. Similarly, manipulations such as base rate which typically affect bias (Bohil and Maddox, 2001; Maddox and Bohil, 2005) can be misinterpreted as selectively making a judge better or worse at the task under certain conditions. For example, an increase in the base rate of treatment effect presence may lead to higher accuracy when an effect is present, but lower accuracy when an effect is absent. So base rate and effect presence are expected to interact on accuracy due to bias, though it would be clearer to have a direct measure of bias rather than this interaction. The signal detection measure of bias reduces the complexity in patterns of results by providing a measure that is directly affected. For example, there was initial enthusiasm about sequential lineups in witness identification. Some early studies showed that sequential lineups decreased false alarms with only a slight reduction in hits, and it was therefore concluded that sequential lineups were superior. However, later signal detection analyses suggested that sequential lineups biased witnesses against selecting any suspect (innocent or not) while reduced or at least did not improve discriminability (respectively, Wixted and Mickes, 2014; Kaesler et al., 2020; for related examples, see Anderson et al., 2005; Modirrousta-Galian and Higham, 2023).

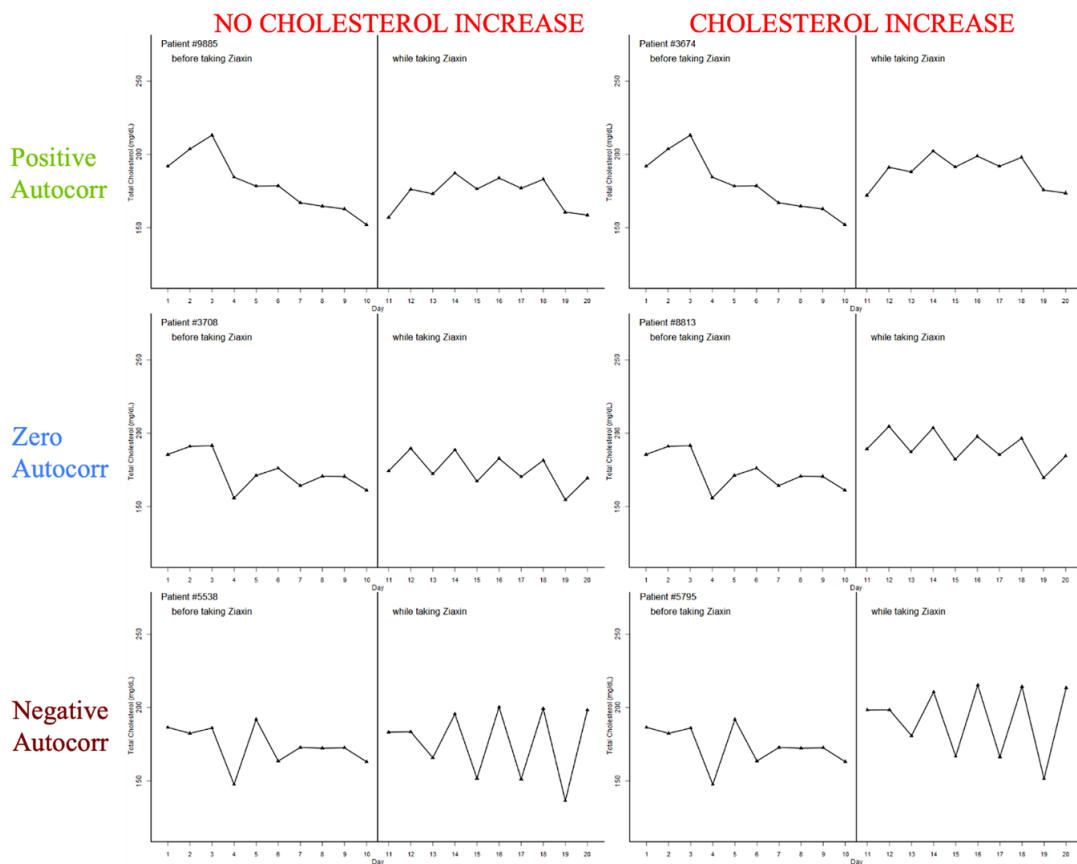
In a signal-detection framework, the shifting of criteria can be normative. When base rates of signal and noise conditions are unequal, a judge trying to maximize utility should establish a criterion that proportionally favors decisions more likely to be correct (Bohil and Maddox, 2001; Healy and Kubovy, 1981; Lynn and Barrett, 2014). That is, the doctor who concluded that all 10 patients showed a cholesterol increase was not entirely wrong to do so; 8 of the patients walked away with the correct diagnosis. In this way, the second doctor achieved similar patient outcomes while conserving effort. The shifting of criteria (bias) is therefore not inherently bad; it can be helpful to the extent that it reflects reality. Unfortunately, human judges sometimes fail to adequately account for base rates and therefore do not capitalize on their full potential (Healy and Kubovy, 1981; Kahneman and Tversky, 1973). Furthermore, when changes in context obsolesce previous beliefs, human judges may rely too heavily on early information and therefore not shift their criterion enough (Hogarth and Einhorn, 1992; Maddox and Bohil, 2005; C. R. Peterson and Ducharme, 1967). This effect is amplified if the judge does not receive immediate feedback (Maddox and Bohil, 2005). If the second doctor were to review case files from a population where cholesterol increases are less common, patient outcomes would likely suffer due to a failure to effectively shift criterion. Criteria are often affected by base rates, especially as compared to discriminability (Healy and Kubovy, 1981).

In addition to shifting criteria, initial information might change the weight that the decision-maker gives to different cues. For example, *predecisional information distortion* affects how the ‘given information’ itself is interpreted (DeKay, 2015). Relatedly, instead of maximizing utility, the goal of *motivated reasoning* is to reach a conclusion that affirms prior beliefs (Kunda, 1990). In this case, even disconfirming evidence (e.g., cues that contradict an earlier established belief) may be re-interpreted as supporting the desired conclusion. If this is the case, base rates might affect discriminability by changing which information is weighed more heavily.

Multiple sources of information (cues) are used to make decisions, but their relevance (validity) varies. For example, people sometimes attend to the slope of a graph after an interruption (e.g., introduction of a drug), while neglecting to compare it to the slope before the interruption (White, 2015; Zhang and Rottman, 2023). Suppose a patient had a steadily rising cholesterol level before trying a new drug. After starting the drug, the patient’s cholesterol continued to increase at the same rate. The doctor may see the rising cholesterol level and incorrectly conclude that the drug increases cholesterol. In some cases, less reliable cues may be used to ease the decision-making process. For example, the doctor may attend mostly to the data points immediately before and after starting the drug (days 10 and 11) to see if the change from baseline to treatment co-occurs with a change in cholesterol (Soo and Rottman, 2018). To achieve a more complete understanding of human decisions in these kinds of circumstances, a lens model displays the correlation of various cues with both human decisions and actual outcomes (Brunswick, 1956; Dhami and Mumpower, 2018). In this way, the lens model may offer insight into motivated reasoning and pre-decisional information distortion. The lens model provides some evidence on predecisional information distortion and motivated reasoning by comparing models in which base rates directly affect decisions versus ones in which base rates interact with other cues.

To better understand the effect of base rates, we compare that manipulation to another—the autocorrelation of the time series. Autocorrelation, or serial dependence, is the extent to which data points in a time series correlate with their previous values; if blood cholesterol is positively autocorrelated, a day of high blood cholesterol is more likely to be followed by another day of high blood cholesterol. Positive autocorrelation can make time series graphs appear smoother as error from one-time point partially spills over into the following time points. In past work, when judging interruptions amidst positive autocorrelation, participants often incorrectly inferred a treatment effect where none existed (Bishara, Peller, et al., 2021; Matyas and Greenwood, 1990). Positive autocorrelation both decreased discriminability and increased bias to decide that an interruption had some kind of effect (Bishara, Peller, et al., 2021). The bias effect likely occurred because participants evaluated a non-directional hypothesis: decide whether the interruption had *no effect* versus *any effect*. In that scenario, positive autocorrelation can create the illusion that a treatment effect occurred because it makes time series data more likely to drift in either direction. In contrast, the present experiments asked participants to evaluate a unidirectional hypothesis: deciding whether an interruption *increased* an outcome versus *not*. In this scenario, positive autocorrelation can create the illusion of a treatment effect only when it causes the time series data to drift upward. When autocorrelation causes the data to drift downward, it can create the illusion that a treatment effect did *not* occur. These 2 effects should largely cancel out with respect to bias, leading autocorrelation to primarily affect discriminability rather than bias. As a result, autocorrelation may contrast with early base rates, the latter of which is expected to primarily affect bias.

Finally, we also examined the effects of negative autocorrelation. As shown in Figure 3, negative autocorrelation causes an oscillatory pattern where high scores tend to be followed by low scores and vice versa; a day of high blood cholesterol is more likely to be followed by a day of low cholesterol, and vice versa. Such a pattern might occur, for example, if consumption of a high-cholesterol food leads to sensory-specific satiety, and therefore temporary avoidance of that food and similar ones. Negative autocorrelation has been less studied in graph judgment, but some results suggest that negative autocorrelation should decrease accuracy (Ximenes et al., 2009). The oscillating, highly variable pattern may visually obscure the mean height of data points, thus making it harder for a human to discriminate treatment effects. In contrast, if an algorithm can accurately measure the mean before the interruption



**Figure 3.** Example interrupted time-series graphs with various levels of autocorrelation.

and the mean afterward, the algorithm could better discriminate treatment effects. Furthermore, with negative autocorrelation, the residuals partially cancel out because a positive residual at one-time point usually leads to a negative residual at the next. An algorithm might leverage that cancellation, thereby achieving higher discriminability amidst negative autocorrelation than positive or even zero autocorrelation.

## 2. Experiment 1

The present experiment involved 3 early base rate conditions: Low, High, and Neutral base rates of treatment effects. These conditions were crossed with 3 autocorrelations: Positive, Negative, and Zero autocorrelation. Both factors were manipulated within subjects, and participants' discriminability and bias were compared. We hypothesized that early base rates would primarily influence bias while autocorrelation would primarily influence discriminability. Additionally, we explored how participants' judgment compared to a simple after-minus-before model (White, 2017) given only a single cue: the mean of the post-intervention period minus the mean of the pre-intervention period. Finally, we explored how participants' judgment could be described by a lens model with multiple cues.

### 2.1. Method

#### 2.1.1. Participants

One hundred and twenty undergraduates were recruited from introductory psychology courses at the College of Charleston. As pre-registered ([https://aspredicted.org/PCN\\_98J](https://aspredicted.org/PCN_98J)), a participant was excluded

if (1) any of the critical trials were uncompleted, (2) overall accuracy was less than or equal to chance (0.5), (3) median reaction time was less than 1 second, or (4) overall experiment completion time exceeded 1.5 hours. These rules resulted in the exclusion of 1 participant for accuracy, 7 participants for median reaction time, and 9 participants for experiment completion time. Due to unforeseen technological problems, 12 participants' data were lost. This resulted in a final sample size of  $n = 91$  (20 male and 71 female). Estimating power to detect autocorrelation's effect on discriminability, unpublished pilot data showed an effect size of  $\eta_p^2 = 0.32$ . The pilot means, standard deviations, and repeated measures correlations, along with the current sample size of 91, led to power to detect the same-sized effect of autocorrelation on discriminability in the current experiment to exceed 0.999 (via the Superpower R package, Lakens and Caldwell, 2021), though this could be an overestimate of power as the smaller sample size of the pilot data ( $n = 38$ ) may have led to a noisy estimate of effect size. No existing data were available to estimate the effect size of early base rates on bias, but power was 0.97 for detecting a medium-sized effect ( $\eta_p^2 = 0.06$ ) and 0.29 for a small-sized effect ( $\eta_p^2 = 0.01$ ; both medium and small estimates conservatively assumed no correlation among repeated measures; Faul et al., 2009). Age ranged from 18 to 38,  $M = 19.2$ ,  $SD = 2.48$ . Participants were compensated with credit for their introductory psychology course.

### 2.1.2. Design and materials

A 2 (Effect Present vs. Absent)  $\times$  3 (Autocorrelation Positive vs. Negative vs. Zero)  $\times$  3 (Early Base Rate: High vs. Low vs. Neutral) repeated measures factorial design was used. The Effect Present condition was defined by a positive (15) population mean difference between the pre- and post-intervention data. The population mean difference was 0 in the Effect Absent condition. The Early Base Rate High condition had 16 of the first 18 stimuli from the Effect Present condition, while Early Base Rate Low had 2 of the first 18, and Early Base Rate Neutral condition had 9 of the first 18. To manipulate autocorrelation, the population lag-1 autoregressive coefficient  $a$  was set to Negative ( $-0.75$ ), Positive ( $+0.75$ ), or Zero, representing how a given  $y$ -value is influenced by the  $y$ -value immediately preceding it. While some types of data show autoregressive tendencies with higher lag values (e.g., lag-2 where the current score is determined by the previous 2 values), lag-1 is generally acceptable for approximating behavioral interventions (Matyas and Greenwood, 1990).

Stimuli were generated using a 1-lag auto-regressive interrupted time-series model:

$$y_t = b + d_t + v_t, \quad (1)$$

where  $y_t$  was the score at time  $t$ , and  $t$  ranged from the 1st to the 20th day. A medication was initiated after the 10th day denoted by a vertical line (see Figures 1 and 3). The constant  $b$  represented the mean of the pre-intervention period ( $t \leq 10$ ) and was set to 175. The value  $d_t$  represented a population mean difference and was set to 0 for all  $t \leq 10$ . For  $t \geq 11$ ,  $d_t$  was either 0 in the Effect Absent condition or 15 in the Effect Present condition. To incorporate serial dependence in the model,  $v_t$  was influenced by previous time points via a 1-lag autoregressive model:

$$v_t = av_{t-1} + e_t, \quad (2)$$

where  $a$  was the population autocorrelation. An  $a$  of 0.00 represents no autocorrelation and denotes that  $v_t$  were generated independently of one another, an  $a$  of .75 represents a strong tendency for the current value to remain close to the previous, and an  $a$  of  $-0.75$  represents an equally strong tendency for the current value to differ from the previous. The value for the immediately previous time point is indicated by  $v_{t-1}$ . For random error,  $e_t$  was sampled from a random normal distribution with population mean = 0 and population standard deviation = 10. To make  $v_t$  approximately stationary even amidst autocorrelation,  $e_t$  was generated starting with  $t = -21$  to allow for a 'burn-in' period of 22 time points before the plot started at  $t = 1$  (22 was determined by the `arma.sim` function's default algorithm (R Core Team, 2020)).

Participants completed 3 blocks of 126 trials each. Each block consisted of 18 feedback trials followed by 108 critical (no feedback) trials. In feedback trials, the effect was present in 16 of 18 (Early Base Rate High), 9 of 18 (Neutral), or 2 out of 18 (Low) graphs. Of these 18 graphs, 6 came from each of the 3 autocorrelation conditions. Of the 108 critical trials, 54 trials had an effect present while the other 54 did not, and 18 of the 54 trials belonged to each of the autocorrelation conditions. For exploratory analyses of judgment over time, each set of 108 critical trials was divided into 6 mini-blocks consisting of 18 trials each, though sampling of conditions was not balanced across mini-blocks.

All trials within a block used one of the 3 fictitious drug names (Thalyde, Ofrocium, or Ziaxin), and the feedback trials within that block used one of 3 Early Base Rate conditions (High, Neutral, or Low). The assignment of drug names and Early Base Rate conditions to the 3 blocks was randomized for each participant with the constraint that each drug name and Early Base Rate condition would be used exactly once for every set of 3 participants. For each participant, the specific graph in each trial was parametrically sampled from a pool of candidate graphs.

$A_g$  was used as a nonparametric measure of discriminability.  $A_g$  is the area under the polygon produced by connecting ROC points (see Figure 2B). This measure was selected because it does not assume that signal and noise distributions are normal, nor does it assume that their variances are equal.  $A_g$  ranges from 0 to 1, where higher values indicate better discriminability, and 0.5 indicates chance performance. For bias, the nonparametric measure  $\ln\beta'_K$  was chosen for the same reasons (Kornbrot, 2006). An  $\ln\beta'_K$  of 0 indicates no bias, positive values indicate a bias to decide that a stimulus is from the signal distribution (cholesterol increased), and negative values indicate a bias to decide that a stimulus is from the noise distribution (cholesterol did not increase).

### 2.1.3. Procedure

Participants answered 3 open-ended questions about age, sex, and ethnicity. Participants were then instructed to imagine a scenario in which they were tasked with determining whether 3 different drugs had a side effect of increasing cholesterol. First, participants were shown an example graph depicting treatment day on the  $x$ -axis and blood cholesterol on the  $y$ -axis. This graph had some variance in the day-to-day measurements. The second and third example graphs, however, depicted both treatment effect sizes (Effect Present and Effect Absent) with zero variance. That is, in the Effect Present condition, participants were shown a horizontal line at height  $y = 175$  for days 1 to 10, followed by a horizontal line at height  $y = 190$  for days 11–20. For the Effect Absent condition, both horizontal lines were at  $y = 175$ . This was followed by the presentation of 2 more graphs depicting these 2 effect sizes with random error. The purpose of these graphs was to show what a mean cholesterol increase looked like in both error-free and more realistic data. Participants were instructed to look at each graph and respond to the following question: ‘Did the drug raise cholesterol?’ Participants responded by selecting a button from a 6-point scale. From left to right, the buttons read: ‘Definitely No Increase (Wager 3)’, ‘Probably No Increase (Wager 2)’, ‘Guess No Increase (Wager 1)’, ‘Guess An Increase (Wager 1)’, ‘Probably An Increase (Wager 2)’, and ‘Definitely An Increase (Wager 3)’.

Participants were informed that they would start with a score of 250, and with each question they would wager between 1 and 3 points depending on their response. Points would either be added to or subtracted from their score depending on if their response was correct or incorrect, respectively. For example, responding ‘Probably An Increase (Wager 2)’ in the Effect Present condition would result in a gain of 2 points, while the same response in the Effect Absent condition would result in a loss of 2 points. Participants were instructed to use all 6 response keys while maximizing their score.

After each feedback trial, the screen showed the participant’s current score, the change in their score from that trial, and the correct answer to that trial (‘Increase Present’ or ‘No Increase’). As a manipulation check, after finishing the 18 feedback trials for the drug, participants were asked what percentage of graphs for that drug appeared to be in the Effect Present condition.

Participants then completed 108 critical trials for that drug *without* feedback before moving on to the next drug. After all trials, participants were debriefed about the study’s purpose and their final score was shown.



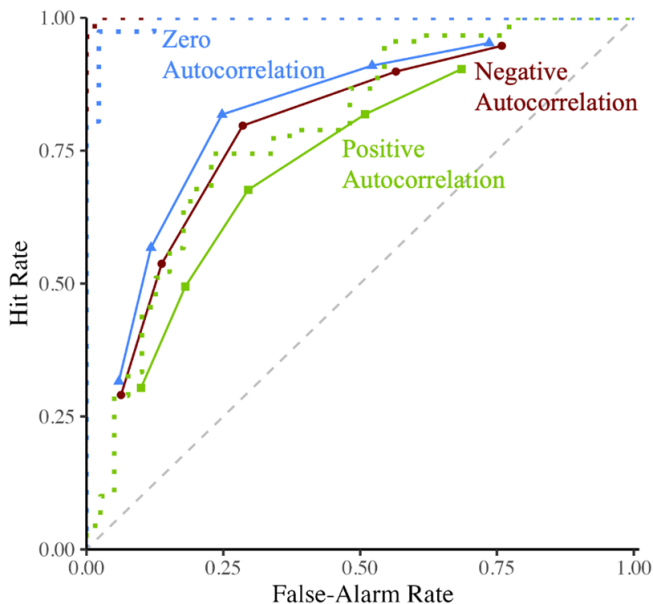
## 2.2. Results and discussion

### 2.2.1. Primary analyses

As pre-registered, a 3 (Autocorrelation: Negative, Positive, Zero)  $\times$  3 (Early Base Rate: Low, Neutral, High) repeated-measures ANOVA was performed for discriminability and then for bias. The Effect Present vs. Absent factor is already summarized in discriminability and bias measures, hence the 2-way ANOVA in a 3-factor design. The Greenhouse–Geisser correction was applied whenever Mauchly's test showed a significant violation of sphericity. As shown in Figure 4, Autocorrelation had a large effect on discriminability and a small effect on bias. In contrast, as shown in Figure 5, the opposite pattern occurred with Early Base rates, which had a large effect on bias and a small effect on discriminability (Table 2).

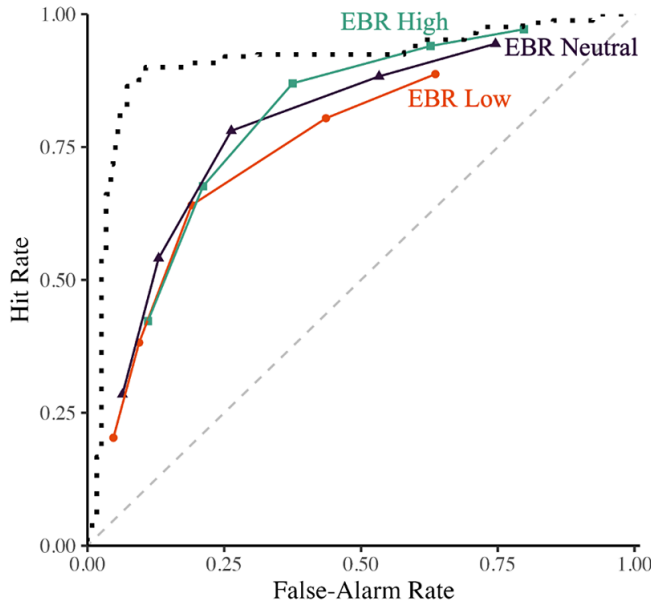
Autocorrelation had a large and significant effect on discriminability ( $F(1.71, 153.58) = 111.26$ ,  $p < .001$ ,  $\eta_G^2 = 0.15$ ). Tukey's HSD was conducted post-hoc for all significant primary effects. Because post-hoc tests were not preregistered, we report Bonferroni corrected  $p$ -values for them by multiplying each original  $p$  by the number of Tukey's HSD families used in the primary analyses (4). All 3 Autocorrelation conditions were significantly different from one another with respect to discriminability ( $ps < .001$ ); discriminability was highest in Zero Autocorrelation ( $M = 0.84$ ), followed by Negative Autocorrelation ( $M = 0.81$ ), and finally Positive Autocorrelation ( $M = 0.73$ ). Early Base Rates had a small but significant effect on discriminability ( $F(2, 180) = 7.91$ ,  $p = .001$ ,  $\eta_G^2 = 0.02$ ). Discriminability was significantly lower in Early Base Rate Low ( $M = 0.78$ ) compared to either of the other 2 conditions ( $ps < .04$ ;  $Ms = 0.80$  and  $0.81$ ). Discriminability was not affected by the interaction of Autocorrelation and Early Base Rates ( $F(3.63, 326.54) = 0.79$ ,  $p = .52$ ,  $\eta_G^2 = 0.002$ ). Importantly, Autocorrelation had a significantly larger effect on discriminability ( $\eta_G^2 = 0.15$ ) than did Early Base Rates (0.02), as indicated by a nonparametric percentile bootstrap of the difference (95% CI [0.09, 0.18]).

Autocorrelation had a small but significant effect on bias ( $F(2, 180) = 17.40$ ,  $p < .001$ ,  $\eta_G^2 = 0.01$ ). Bias was significantly lower in the Positive Autocorrelation condition ( $M = 0.20$ ) compared to either of the other 2 conditions ( $ps < .001$ ;  $Ms = 0.34$  and  $0.36$  for Negative Autocorrelation and Zero Autocorrelation, respectively). Early Base Rates had a large and significant effect on bias ( $F(1.71,$



**Figure 4.** Experiment 1: ROC of human and model judgment at varying levels of autocorrelation.

Note: The dotted lines represent after-minus-before model performance. ROC, Receiver Operating Characteristics.



**Figure 5.** Experiment 1: ROC of human and model judgment at varying early base rates.

Note: The dotted lines represent after-minus-before model performance. The model is not affected by the Early Base Rates because it ignores previous graphs when judging the current graph. ROC, Receiver Operating Characteristics; EBR, Early Base Rate.

**Table 2.** Experiment 1: mean discriminability and bias by Autocorrelation and Early Base Rate.

Early base rate	Autocorrelation							
	Negative			Zero	Positive			M
	Discriminability ( $A_g$ )				Bias ( $\ln\beta'_\kappa$ )			
High	0.82	0.85	0.75	<b>0.81</b>	0.95	0.91	0.67	<b>0.84</b>
Neutral	0.82	0.85	0.74	<b>0.80</b>	0.40	0.39	0.22	<b>0.34</b>
Low	0.80	0.82	0.71	<b>0.78</b>	-0.33	-0.23	-0.28	<b>-0.28</b>
M	<b>0.81</b>	<b>0.84</b>	<b>0.73</b>	<b>0.79</b>	<b>0.34</b>	<b>0.36</b>	<b>0.20</b>	<b>0.30</b>

Note: Boldface values indicate means.

154.15) = 81.64,  $p < .001$ ,  $\eta_G^2 = 0.28$ ). Bias varied significantly between all Early Base Rates conditions ( $ps < .001$ ), with the highest bias in Early Base Rate High ( $M = 0.84$ ), followed by Early Base Rate Neutral ( $M = 0.34$ ), and finally Early Base Rate Low ( $M = -0.28$ ). Autocorrelation and Early Base Rates also had a small but significant interaction for bias ( $F(3.63, 326.85) = 6.73$ ,  $p < .001$ ,  $\eta_G^2 = 0.01$ ). Early Base Rates had a significantly larger effect on bias ( $\eta_G^2 = 0.28$ ) than did Autocorrelation (0.01; 95% CI of the difference [0.21, 0.35]).

As pre-registered, a 3 (Autocorrelation: Negative, Positive, Zero)  $\times$  3 (Early Base Rate: Low, Neutral, High)  $\times$  2 (Effect: Present, Absent) repeated-measures ANOVA was conducted on accuracy. Accuracy was defined as (Hits + Correct Rejections) / (Hits + Correct Rejections + Misses + False Alarms). All ‘... An Effect’ responses, regardless of wager, were collapsed, and likewise, all ‘... No Effect’ responses were collapsed. As shown in Table 3, the Autocorrelation had a small main effect on accuracy ( $F(1.61, 144.84) = 84.32$ ,  $p < .001$ ,  $\eta_G^2 = 0.05$ ), as did Early Base Rates ( $F(2, 180) = 8.64$ ,  $p < .001$ ,  $\eta_G^2 = 0.01$ ) and Effect Presence ( $F(1, 90) = 5.18$ ,  $p = .03$ ,  $\eta_G^2 = 0.01$ ). The interaction of Early Base Rates and Effect Presence had a large effect ( $F(1.69, 151.96) = 75.71$ ,  $p < .001$ ,  $\eta_G^2 = 0.20$ ) while

**Table 3.** Experiment 1: mean accuracy by autocorrelation, Early Base Rate, and Effect Presence.

Variable	Accuracy	
	Effect present	Effect absent
Autocorrelation		
Negative	0.80	0.71
Zero	0.82	0.75
Positive	0.68	0.70
Early Base Rate		
Low	0.64	0.81
Neutral	0.78	0.74
High	0.87	0.62

the interaction between Autocorrelation and Effect Presence had a small effect ( $F(2, 180) = 45.49$ ,  $p < .001$ ,  $\eta_G^2 = 0.02$ ). Finally, the 3-way interaction of Autocorrelation, Early Base Rates, and Effect Presence had a small but significant effect ( $F(4, 360) = 6.78$ ,  $p < .001$ ,  $\eta_G^2 = 0.01$ ).

### 2.2.2. Exploratory analyses

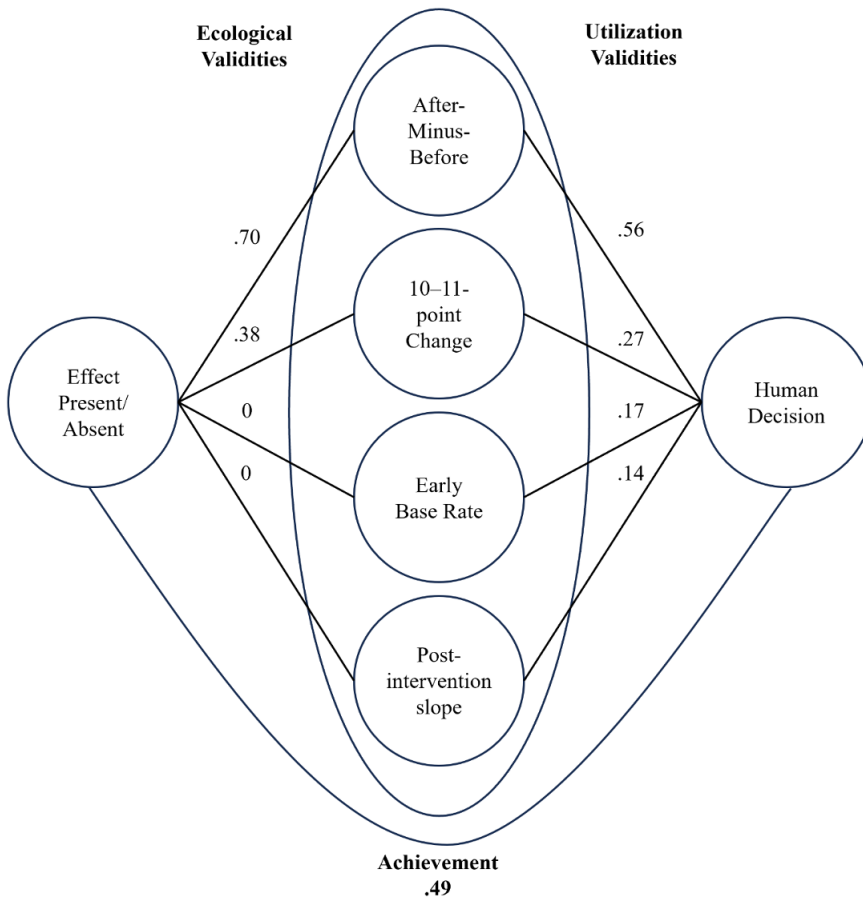
Participant performance was compared to the performance of a simple after-minus-before logistic regression model. For this model, each graph was summarized by a single numerical value: the mean of days 11–20 minus the mean of days 1–10 (White, 2017). This after-minus-before value was used as a predictor variable in logistic regression, and the binary outcome variable was whether each graph truly had a cholesterol increase (1) or not (0). To avoid evaluating the model on the same data it was trained on, which would overestimate the model's performance, this logistic regression model was fit to half of the graphs sampled randomly (the training stimuli). For the remaining graphs (the test stimuli), the model produced a  $p$ -value indicating the probability that cholesterol increased. Lower  $p$ -values indicated greater perceived evidence for a cholesterol increase. That is, lower  $p$ -values were represented farther to the right on the horizontal axis of Figure 2A.

This after-minus-before model's performance is shown by the dotted lines in Figures 4 and 5. The model outperformed human judgment in every condition with an average discriminability ( $A_g$ ) of 0.91. As it was with human judgment, model performance was reduced by positive autocorrelation ( $A_g = 0.79$ ). However, unlike human judgment, model performance was improved by negative autocorrelation ( $A_g = 0.999$ ). Negative autocorrelation caused error to partly cancel out across time points, making the sample mean of each phase more representative of the population means. This made the after-minus-before value an excellent cue. However, the oscillating visual pattern produced by negative autocorrelation (see Figure 3) may make it harder for humans to notice this cue. Note that, unlike human judgment, the model's performance was unaffected by early base rates because it ignores previous graphs when judging the current graph.

A lens model was used to provide more insight into what cues participants used to make decisions. First, a list of potential cues was created: after-minus-before value, observed autocorrelation, proportion of feedback trials in the Effect Present condition, the difference between time points 10 and 11, the pooled standard deviation, and the slope of the post-intervention phase. The first 3 cues were sample estimates of manipulated variables. The difference between the point immediately before and after the intervention (points 10 and 11) was included because it may be a simple way to estimate the correlation of a transition (Soo and Rottman, 2018). The pooled standard deviation of the pre-/post-intervention trials was included because people may make decisions like a  $t$ -test, considering not only the after-minus-before value (the numerator of a  $t$ -test) but also some measure of noise (the denominator; Obrecht

et al., 2007; White, 2017). The slope of the post-intervention phase was included because it can affect judgment, sometimes even in cases where it is unchanged relative to the pre-intervention phase (White, 2017; Zhang and Rottman, 2023). Finally, 2-way interactions between each cue and early base rates were considered because early base rates may affect how other information is processed following a motivated reasoning framework (Kunda, 1990). Examining every possible combination of these cues, 275 multiple logistic regression models were tested. In these models, the binary outcome variable was whether the participant gave one of the ‘. . .An Effect’ or ‘. . .No Effect’ response on critical (no-feedback) trials. Models were then assessed using their Bayesian information criterion (BIC), and the model with the lowest BIC was selected. This selected model had 4 cues: the after-minus-before value, the 10–11-point change, the proportion of feedback trials from the Effect Present condition, and the post-intervention slope. There were no interaction terms. All other cues and all interactions led to higher (worse) BICs, suggesting that they were less accurate descriptions of the cue utilization by participants.

The point-biserial correlation was then calculated for each predictor of the selected model to both participant responses and whether the stimulus had been generated from the population with an effect present. As shown in Figure 6, participants relied on the after-minus-before difference and the 10–11-point difference less than they should have; participants also relied on Early Base Rates and



**Figure 6.** *Experiment 1: lens model.*

*Note:* Numbers show the point-biserial correlation between each cue (middle set of circles) and the actual answer (left circle) or each cue and the human decision (right circle). After-minus-before is the mean of the post-intervention points (11–20) minus the mean of the pre-intervention points (1–10).

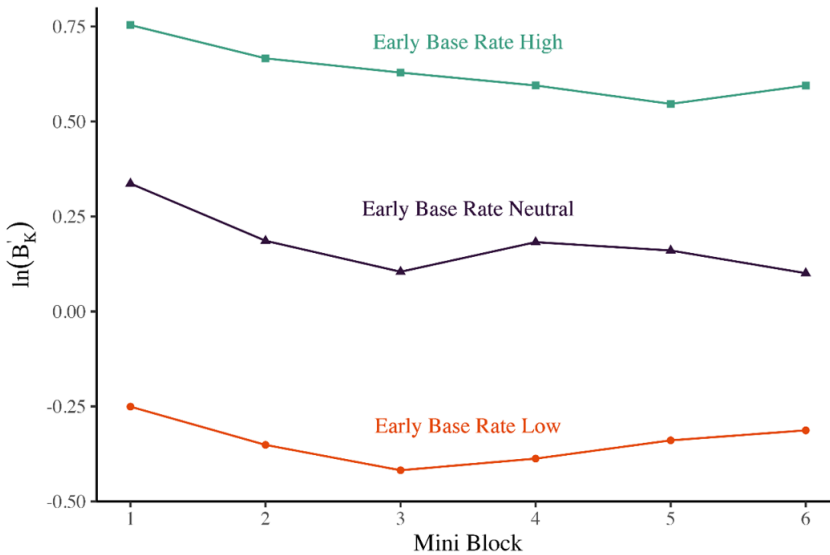


Figure 7. Experiment 1: bias across mini-blocks.

post-intervention slope, though neither of these had any correlation with the variable they were trying to predict ( $|r|s < 0.01$ ). As such, participant achievement (the correlation between effect presence and human decisions) was only moderate (0.49). In contrast, the after-minus-before model had a strong correlation (0.70) with the outcome variable.

Could the effect of base rates disappear over trials, perhaps as participants noticed that the base rates had reverted to 50% effect presence? Critical trials were divided into 6 mini-blocks, each comprised of 18 trials. A 3 (Early Base Rate High vs. Low vs. Neutral)  $\times$  6 (mini-blocks) within-subjects ANOVA was conducted on bias ( $\ln\beta'_K$ ). As shown in Figure 7, mini-block had a small main effect on bias ( $F(5, 450) = 5.86, p < .001, \eta^2_G = 0.01$ ). However, mini-block did not significantly interact with Early Base Rates ( $F(8.46, 761.68) = 0.99, p = .45, \eta^2_G = 0.002$ ), suggesting that the biasing effects of base rates failed to dissipate over time. Furthermore, when the ANOVA was restricted to mini-blocks 3–6, the main effect of mini-block became non-significant, suggesting that bias stabilized by mini-block 3 ( $F(3, 270) = 0.23, p = .88, \eta^2_G = .000$ ). Even when restricted to mini-blocks 3–6, the large effect of Early Base Rates remained ( $F(1.7, 152.71) = 65.5, p < .001, \eta^2_G = 0.22$ ).

For the manipulation check, participants estimated that the percentage of cholesterol increases in the feedback trials were 77.8%, 54.8%, and 27.3% in the Early Base Rates High, Neutral, and Low conditions respectively ( $F(1.83, 164.31) = 159.68, p < .001, \eta^2_G = .53$ ). Tukey’s HSD revealed that all conditions were significantly different from each other ( $ps < .001$ ).

### 3. Experiment 2

In the first experiment, Early Base Rates affected later graph judgments. However, as a manipulation check, participants were asked about these base rates immediately before the critical trials for the drug. Participants might have inferred that these base rates were important (after all, why else ask about them; Grice, 1975) and therefore made special effort to use early base rates with a drug for future judgments of that drug. Even without this inference, recalling the feedback trials would likely make them more memorable than trials that required no recall (Roediger and Karpicke, 2006). In Experiment 2, we moved this manipulation check to the end of each set of critical trials (each drug) to avoid these problems.

Experiment 2 also addressed a concern about interpreting the negative autocorrelation condition. As negative autocorrelation leads to jagged graphs (high data point followed by low followed by high, etc.), the variance of data points in such graphs was larger than in the other conditions. So, it is unclear whether negative autocorrelation itself or the larger variance that resulted from it caused lower discriminability. To address this issue, in Experiment 2, we included a condition where the autocorrelation was 0, but the population variance was approximately matched to that of the negative autocorrelation condition. Relatedly, this condition allowed us to explore the possibility that noise could enhance the effects of motivated reasoning, which could be shown by an interaction with other cues. This ‘Zero with Matched Noise’ condition replaced the Positive Autocorrelation condition from Experiment 1. The replacement of the Positive Autocorrelation condition allowed for the number of trials per condition to remain the same across experiments.

### 3.1. Method

#### 3.1.1. Participants

One hundred and twenty undergraduates were recruited from introductory psychology courses at the College of Charleston. Criteria for participant exclusion were pre-registered ([https://aspredicted.org/YR\\_TQX](https://aspredicted.org/YR_TQX)) to be identical to Experiment 1. The criteria resulted in the exclusion of 9 participants for accuracy, 14 participants for median reaction time, and 21 participants for experiment completion time. Two participants were excluded for 3 criteria, and another 7 participants were excluded for 2 criteria. The remaining 24 participants excluded from analysis were excluded for only one criterion. Lastly, the data of 6 participants were corrupted due to an internal error with Qualtrics. This resulted in a final sample size of  $n = 81$  (18 male and 65 female). With this sample size, power to detect the effect of base rates on bias exceeded .99, as estimated via Experiment 1 means, standard deviations, and repeated measures correlations (Lakens and Caldwell, 2021). We had no previous data regarding the Zero Autocorrelation with Matched Noise condition’s effect on discriminability, but power to detect a medium effect ( $\eta_p^2 = 0.06$ ) was 0.95, and power to detect a small effect ( $\eta_p^2 = 0.01$ ) was 0.26 (Faul et al., 2009). Age ranged from 18 to 32 ( $M = 19.6$ ,  $SD = 2.43$ ).

#### 3.1.2. Design and materials

A 2 (Effect Presence: Present vs. Absent)  $\times$  3 (Autocorrelation: Negative vs. Zero vs. Zero with Matched Noise)  $\times$  3 (Early Base Rate: High vs. Low vs. Neutral) repeated measures factorial design was used. The Negative Autocorrelation and Zero Autocorrelation conditions were identical to those of Experiment 1, and the Positive Autocorrelation condition was replaced with a ‘Zero with Matched Noise’ condition. In the Zero with Matched Noise condition,  $a$  was set to 0, but the standard deviation of  $e_i$  was set at 15.45937. This made the stimuli’s population variance approximately that of the Negative Autocorrelation condition (see Supplementary Material for details on this approximation). Like Experiment 1, stimuli falling into each of these 3 conditions were shuffled within blocks, but with the added specification that the 18-trial mini-blocks were fully balanced with respect to effect presence and autocorrelation.

#### 3.1.3. Procedure

Participants’ experience in this experiment was identical to Experiment 1 with the sole exception that participants were only asked about the portion of trials appearing to raise cholesterol after the final critical trial for each drug, rather than after the feedback trials.

### 3.2. Results and discussion

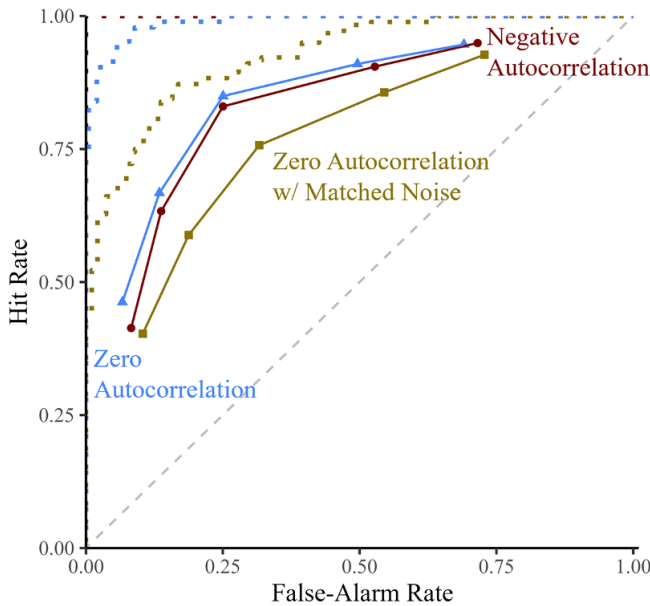
#### 3.2.1. Primary analyses

As in Experiment 1, Autocorrelation primarily affected discriminability, whereas Early Base rates primarily affected bias (Table 4). As shown in Figure 8, the main effect of Autocorrelation on

**Table 4.** Experiment 2: mean discriminability and bias by autocorrelation and Early Base Rate.

Early base rate	Autocorrelation							
	Zero w/matched			$M(A_g)$	Zero w/matched			$M(\ln\beta'_K)$
	Negative	Zero	Noise		Negative	Zero	Noise	
Discriminability ( $A_g$ )			Bias ( $\ln\beta'_K$ )					
High	0.84	0.87	0.78	<b>0.83</b>	0.91	0.92	0.87	<b>0.90</b>
Neutral	0.85	0.85	0.77	<b>0.82</b>	0.60	0.63	0.53	<b>0.59</b>
Low	0.81	0.84	0.76	<b>0.81</b>	0.18	0.21	0.17	<b>0.19</b>
<i>M</i>	<b>0.84</b>	<b>0.85</b>	<b>0.77</b>	<b>0.82</b>	<b>0.56</b>	<b>0.59</b>	<b>0.52</b>	<b>0.56</b>

Note: Boldface values indicate means.

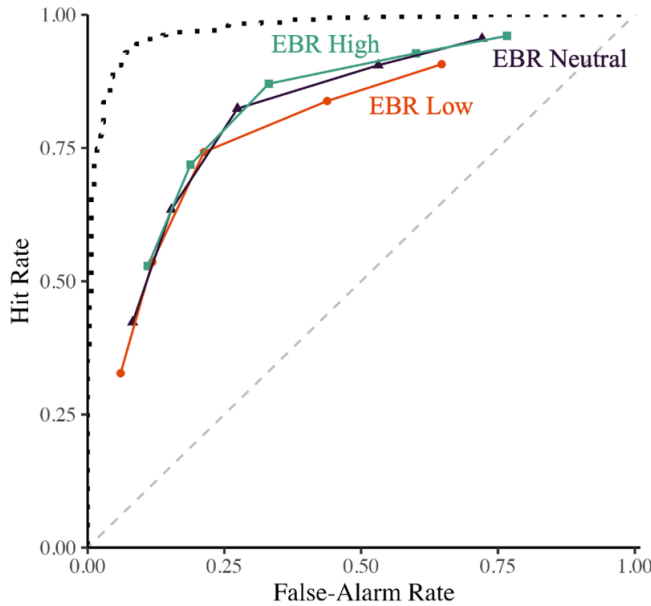


**Figure 8.** Experiment 2: ROC of human and model judgment at varying levels of autocorrelation.

Note: The dotted lines represent after-minus-before model performance. ROC, Receiver Operating Characteristics.

discriminability was significant and of medium effect size ( $F(2, 160) = 91.83, p < .001, \eta_G^2 = 0.11$ ). Tukey’s HSD was again conducted post-hoc for all significant primary effects. Because post-hoc tests were not preregistered, we report Bonferroni corrected  $p$ -values for them by multiplying each original  $p$  by the number of Tukey HSD families used in the primary analyses (3). All 3 conditions of Autocorrelation were significantly different from one another with respect to discriminability ( $ps < .03$ ); Zero Autocorrelation showed the highest discriminability ( $M = 0.85$ ), followed by Negative Autocorrelation ( $M = 0.83$ ) and finally Zero with Matched Noise ( $M = 0.77$ ). As shown in Figure 9, the main effect of Early Base Rates was small but significant ( $F(2, 160) = 4.56, p = 0.01, \eta_G^2 = 0.01$ ). Discriminability was significantly higher in Early Base Rate High ( $M = 0.83$ ) compared to Early Base Rate Low ( $M = 0.81; p = .047$ ). The 2 factors did not interact ( $F(4, 320) = 1.87, p = .12, \eta_G^2 = 0.004$ ). Autocorrelation had a significantly larger effect on discriminability ( $\eta_G^2 = 0.11$ ) than did Early Base Rates (0.01; 95% CI of the difference [0.06, 0.15]).

Like Experiment 1, Early Base Rates had a large and significant effect on bias ( $F(1.51, 120.60) = 44.08, p < .001, \eta_G^2 = 0.17$ ). Bias was again significantly different between all conditions



**Figure 9.** Experiment 2: ROC of human and model judgment at varying EBR.

Note: The dotted lines represent after-minus-before model performance. ROC, Receiver Operating Characteristics; EBR, Early Base Rate.

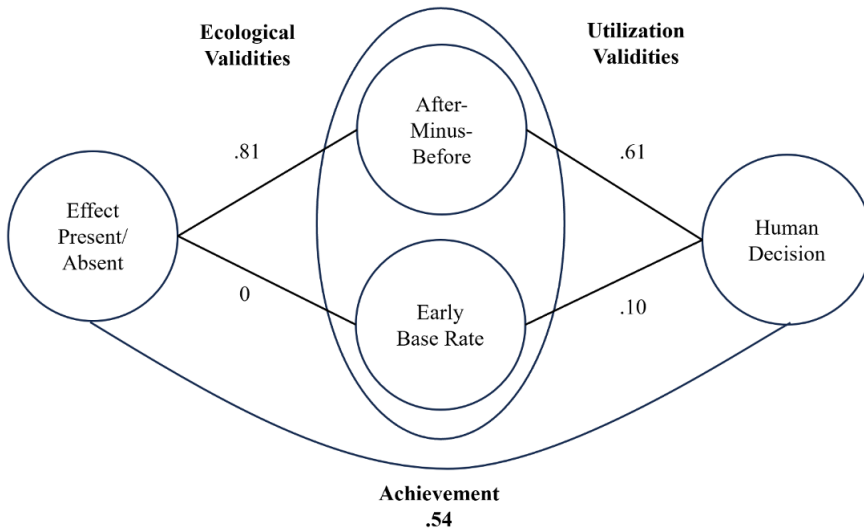
**Table 5.** Experiment 2: mean accuracy by autocorrelation, Early Base Rates, and Effect Presence.

Variable	Accuracy ( <i>M</i> )	
	Effect present	Effect absent
Autocorrelation		
Negative	0.83	0.75
Zero with matched noise	0.76	0.68
Zero	0.85	0.75
Early Base Rate		
Low	0.74	0.79
Neutral	0.82	0.73
High	0.87	0.67

( $p < .001$ ): highest in Early Base Rate High ( $M = 0.90$ ), then Early Base Rate Neutral ( $M = 0.59$ ), and lowest in Early Base Rate Low ( $M = 0.19$ ). Autocorrelation was almost significant with respect to bias ( $F(1.82, 145.39) = 2.81, p = .07, \eta_G^2 = 0.002$ ), and the 2 factors did not interact ( $F(4, 320) = 0.30, p = .88, \eta_G^2 = 0.000$ ). Early Base Rates had a significantly larger effect on bias ( $\eta_G^2 = 0.17$ ) than did Autocorrelation (0.002; 95% CI of the difference [0.10, 0.24]).

In terms of accuracy, as shown in Table 5, the Autocorrelation had a small main effect ( $F(2, 160) = 101.59, p < .001, \eta_G^2 = 0.05$ ) while the main effect of Effect Presence was moderate ( $F(1, 80) = 17.7, p < .001, \eta_G^2 = 0.07$ ). The interaction of Effect Presence and Early Base Rates was moderate ( $F(1.70, 136.34) = 38.07, p < .001, \eta_G^2 = 0.09$ ). All other effects and interactions were nonsignificant.





**Figure 10.** Experiment 2: lens model.

Note: After-minus-before is the mean of the post-intervention days (11–20) minus the mean of the pre-intervention days (1–10).

**3.2.2. Exploratory analyses**

Participant performance was again compared to the performance of a simple after-minus-before model. A similar pattern emerged, with model performance exceeding human performance in all conditions, as shown in Figure 8.

The model with the lowest BIC in Experiment 2 included just 2 cues: the after-minus-before value and the Early Base Rate (see Figure 10). As before, participants relied most heavily on the after-minus-before value, though not as much as they should have. Participants again relied on the Early Base Rate despite it having no correlation with effect presence. The key difference is that participants made use of neither the post-intervention slope nor the 10–11-point difference. Reliance on the post-intervention slope may be more likely amidst positive autocorrelation, as the smoothing effect of positive autocorrelation makes (illusory) directional trends more apparent. Positive autocorrelation may also increase the salience of the 10–11-point difference by smoothing out other transitions between days.

Like Experiment 1, mini-block had a small main effect on bias ( $F(4.36, 349.18) = 4.01, p = 0.003, \eta_G^2 = 0.01$ ), but there was no significant interaction between mini-block and Early Base Rates ( $F(7.79, 623.15) = 1.22, p = .29, \eta_G^2 = 0.003$ ). Again, when the ANOVA was restricted to mini-blocks 3–6, the main effect of mini-block became non-significant ( $F(2.61, 208.78) = 0.35, p = .76, \eta_G^2 = 0.000$ ), but the effect of Early Base Rates remained ( $F(1.52, 121.98) = 35.69, p < .001, \eta_G^2 = .13$ ).

For the manipulation check, participants estimated that the percentage of cholesterol increases across all trials (feedback and critical) were 63.1%, 55.5%, and 52.8% for the drug in the Early Base Rates High, Neutral, and Low conditions respectively ( $F(2, 160) = 12.54, p < .001, \eta_G^2 = .05$ ). Tukey’s HSD revealed that the drug in the Early Base Rate High condition received significantly higher estimates than the drugs in the other 2 conditions ( $ps < .01$ ).

**4. General discussion**

After viewing 18 graphs, judges became biased by the base rate of that initial information. Indeed, even as context changed and their newly established beliefs obsolesced, human judges failed to adjust their beliefs to make them more consistent with the task environment. In Experiment 1, the effect of Early Base Rates on bias may have been amplified by conversational norms and/or recall practice induced by

the manipulation check. That is, asking participants to recall the base rate of the early graphs may have made participants attend to and remember this information more strongly. However, in Experiment 2, when this manipulation check was moved to the end of each block, it could no longer affect graph judgments of that drug. Nevertheless, the feedback trial base rates still had a large effect on the bias of later critical trial judgments. Likewise, decision-makers in other contexts may also fail to adjust their beliefs adequately (Healy and Kubovy, 1981; Kahneman and Tversky, 1973), especially when they receive limited or delayed feedback on the veracity of their judgments (Maddox and Bohil, 2005). Believing their decisions to be correct, the decision maker may cease to adjust their criterion once they feel they understand the task; indeed, after about 54 trials (Mini-Block 3), bias stabilized and did not adjust further, continuing to bias future decisions.

We have interpreted the effects of base rates on bias as being due to a change in the response threshold (Figure 2A dotted line). However, an alternative interpretation is that it is due to a shift of both the signal and noise distributions' means by the same amount in the same direction (Witt et al., 2015; also see Modirrousta-Galian and Higham, 2023). For example, bias in the high base rate condition could also be due to shifting both distributions to the right while keeping the response threshold constant. One could imagine that the high base rate condition caused participants to opportunistically seek out any graph features, even spurious ones, which could be construed to support a cholesterol increase. This would increase the perceived evidence for cholesterol increases for both effect present and effect absent trials. Such an interpretation predicts that base rates would interact with the utilization of other graph cues. We found no evidence of such an interaction in the lens models, though of course there could be relevant graph features that were not considered in these models. The results showed a primacy effect where the early probability of effects in graphs had a large impact on later judgments. However, the relative importance of primacy versus recency likely depends on the task characteristics (Hogarth and Einhorn, 1992). A primacy effect could have occurred here because feedback was removed after 18 trials with a drug. The lack of feedback could lower the weight given to critical trials when updating judgments of the base rate. Additionally, primacy may also be due to the length of the task (126 graphs per drug), which could encourage participants to develop firmer and firmer beliefs, and thereby make smaller and smaller adjustments in later trials (Hogarth and Einhorn, 1992). This primacy effect was useful here mainly for a methodological reason: it allowed us to observe a base rate effect on bias while equating the base rates of critical trials (always 50%).

Had these experiments used a text-based rather than graphical presentation format, the outcome might have been different. Graphical presentation of data may increase sensitivity to trends such as the post-intervention slope (White, 2017; Zhang and Rottman, 2023). Because the post-intervention slope was utilized by participants in Experiment 1 but was not predictive, reduced utilization of this cue could have actually increased discriminability. Base rates have been shown to bias human judges in a variety of contexts, so text-based presentation is unlikely to diminish this effect (Han and Dobbins, 2008; Maddox and Bohil, 2005). In fact, if mean difference utilization decreased with text-based presentation (e.g., White, 2015), it is possible that base rates could make up a larger portion of their decision-making process. Another finding of note is that unlike previous research (Han and Dobbins, 2008; Healy and Kubovy, 1981), base rates had a small but significant effect on discriminability such that the High Early Base Rate condition had a small but consistent advantage over the Low Early Base Rate condition. It is possible that low early base rates lead to slightly less effort in searching for a difference between before versus after the medication was started. Nevertheless, this discriminability effect was more subtle and appeared to be restricted to the 3 more relaxed criteria (the 3 right-most dots in Figures 5 and 9).

One concern is that these base rate effects might transfer narrowly to only stimuli with similar surface features (e.g., line graphs with 20 days on the x-axis, an interruption after day 10, etc.). However, there were large changes in bias from block to block where drug names changed while those surface features remained relatively constant. This pattern suggests that base rates affected participants' higher-order beliefs about each drug (e.g., belief that Ziaxin usually raises cholesterol). In other words, it is plausible that these base rate effects could transfer to other stimuli (e.g., text or tables), so long they involved the same drug.

The discriminability of human judgment was reduced by both autocorrelation and noise. The finding that autocorrelation reduces discriminability is not new, but the lens model analysis does put it in a new view. Compared to a simple after-minus-before model, informal human judgment was worse in every case. Conversely, the simple after-minus-before cue was quite effective. Note that these experiments used a simple population-generating model which included a possible mean difference, but no trends (non-zero slopes). Interrupted time series with trends would make this simple cue less diagnostic.

Negative autocorrelation improved model performance while worsening human performance. The reduction in discriminability of both human and model decisions when stimuli had increased variance compared to negative autocorrelation suggests that autocorrelation uniquely affects judgment in a way that variance alone does not. Increased model performance suggests that negative autocorrelation increased the viability of the after-minus-before value as a decision cue. This is likely because negative autocorrelation creates positive and negative residuals in an alternating pattern which partly cancels out. It is unclear why negative autocorrelation reduced human discriminability, but one possible explanation is that the jagged pattern of negatively autocorrelated data makes means (and in turn after-minus-before values) harder to judge by eye.

Positive autocorrelation reduced both human performance and model performance. Interestingly, observed autocorrelation was not included as a parameter in the selected lens model for either experiment nor were any autocorrelation interactions. This suggests that positive autocorrelation might not affect how human judges use other cues, including the after-minus-before value. Instead, positive autocorrelation likely makes the after-minus-before value a less useful cue (as supported by the dotted lines in Figure 4).

The failure of human judgment to outperform even a simple model is unfortunate given the prevalence of single-case time series judgment in a variety of contexts (Matyas and Greenwood, 1990; White, 2015, 2017). The limited discriminability of informal human judgment is likely caused by both sub-optimal cue utilization and inconsistency. Inconsistency in human judgment reduces discriminability because judgment often varies based on irrelevant contextual factors (Goldberg, 1968). As a result, an individual's decisions vary not only from case to case, but even in the same case when viewed in different contexts (Ashton, 2000). Signal detection theory has been instrumental in analyzing human and model performance. Had analyses been limited to accuracy, it would have been harder to spot what factors made the task more difficult and what factors biased participants (Anderson et al., 2005; Bishara, Li, et al., 2021; Modirrousta-Galian and Higham, 2023; Wixted and Mickes, 2014). For example, in Experiment 1 early base rates affected accuracy, suggesting that base rates affected task difficulty; signal detection theory revealed base rates primarily affected bias rather than discriminability. These results show a circumstance where initial beliefs alter a response criterion, even without those beliefs interacting with other potential cues (for a related example, see Vicente et al., 2023).

Graph literacy was not measured in the experiments reported here. Participants were non-experts, specifically undergraduates, and so their graph literacy likely varies (Galesic and Garcia-Retamero, 2011). In previous work, time series graph judgment has been examined both in a general U.S. adult population and in experts (graduate students in behavioral analysis programs, most of whom had experience with single-subject interrupted time series graphs; Bishara et al., 2021). The 2 groups' average performance was similar. Furthermore, the experts showed about the same impairment due to autocorrelation as the general population did. It may be hard for even experts to adjust for autocorrelation because the perceptual features it creates are similar to those of noise (or lack thereof). For example, a smooth line produced by high autocorrelation may be mistaken as data with little noise, and the low perceived noise of the data may in turn cause judges to infer that effects are more likely to be present (White, 2017). While expertise is unlikely to moderate the effects of autocorrelation, its potential to moderate base rate usage in graph judgment is less clear.

Graph judgment remains a powerful tool for the exploration and dissemination of data. Graphs can enhance understanding of trends as well as reduce misperceptions (White, 2017). Nonetheless, previous research shows that graph judgment can be affected by background beliefs (Freedman and Smith, 1996; Lee and Lee, 2022; Mena, 2023; Nyhan and Reifler, 2019). As shown by the experiments here, these

background beliefs can be created by graphs themselves. Furthermore, these beliefs can affect behavior largely through bias, that is, a change in a decision threshold. Finally, even though such bias can be adaptive while those beliefs are valid, the bias can persist even when those beliefs cease to be so. Overall, when exposed to multiple graphs on the same topic, human judges can draw conclusions about the data, but once those conclusions are made, they can affect subsequent graph judgment.

**Data availability statement.** Raw data are available at <https://osf.io/pf239/>.

**Acknowledgments.** We thank Stephen Short and Benjamin Rottman for helpful advice, and Craig Tanton for help with data collection.

**Funding statement.** This research received no specific grant funding from any funding agency, commercial or not-for-profit sectors.

**Competing interest.** The authors declare none.

## References

- Anderson, R. B., Doherty, M. E., Berg, N. D., & Friedrich, J. C. (2005). Sample size and the detection of correlation—A signal detection account: Comment on Kareev (2000) and Juslin and Olsson (2005). *Psychological Review*, *112*(1), 268–279.
- Ashton, R. H. (2000). A review and analysis of research on the test–retest reliability of professional judgment. *Journal of Behavioral Decision Making*, *13*(3), 277–294.
- Bishara, A. J., Li, J., & Conley, C. (2021). Informal versus formal judgment of statistical models: The case of normality assumptions. *Psychonomic Bulletin & Review*, *28*(4), 1164–1182. <https://doi.org/10.3758/s13423-021-01879-z>
- Bishara, A. J., Peller, J., & Galuska, C. M. (2021). Misjudgment of interrupted time-series graphs due to serial dependence: Replication of Matyas and Greenwood (1990). *Judgment and Decision Making*, *16*(3), 687–708. <https://doi.org/10.1017/S1930297500007786>
- Bohil, C. J., & Maddox, W. T. (2001). Category discriminability, base-rate, and payoff effects in perceptual categorization. *Perception & Psychophysics*, *63*(2), 361–376. <https://doi.org/10.3758/BF03194476>
- Brunswick, E. (1956). *Perception and the representative design of psychological experiments*. Berkeley, CA: University of California Press. <https://doi.org/10.1525/9780520350519>
- DeKay, M. L. (2015). Predecisional information distortion and the self-fulfilling prophecy of early preferences in choice. *Current Directions in Psychological Science*, *24*(5), 405–411. <https://doi.org/10.1177/0963721415587876>
- Dhami, M. K., & Mumpower, J. L. (2018). Kenneth R. Hammond's contributions to the study of judgment and decision making. *Judgment and Decision Making*, *13*(1), 1–22. <https://doi.org/10.1017/S1930297500008780>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Franconeri, S. L., Padilla, L. M., Shah, P., Zacks, J. M., & Hullman, J. (2021). The science of visual data communication: What works. *Psychological Science in the Public Interest*, *22*(3), 110–161. <https://doi.org/10.1177/15291006211051956>
- Freedman, E. G., & Smith, L. D. (1996). The role of data and theory in covariation assessment: Implications for the theory-ladenness of observation. *Journal of Mind and Behavior*, *17*(4), 321–343.
- Galesic, M., & Garcia-Retamero, R. (2011). Graph literacy: A cross-cultural comparison. *Medical Decision Making*, *31*(3), 444–457. <https://doi.org/10.1177/0272989X10373805>
- Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *59*(4), 582–627. <https://doi.org/10.1177/0018720817690634>
- Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist*, *23*(7), 483–496. <https://doi.org/10.1037/h0026206>
- Gronlund, S. D., Wixted, J. T., & Mickes, L. (2014). Evaluating eyewitness identification procedures using Receiver Operating Characteristic analysis. *Current Directions in Psychological Science*, *23*(1), 3–10. <https://doi.org/10.1177/0963721413498891>
- Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics, Vol. 3: Speech acts* (pp. 41–58). New York, NY: Academic Press.
- Han, S., & Dobbins, I. G. (2008). Examining recognition criterion rigidity during testing using a biased-feedback technique: Evidence for adaptive criterion learning. *Memory & Cognition*, *36*(4), 703–715. <https://doi.org/10.3758/MC.36.4.703>
- Healy, A. F., & Kubovy, M. (1981). Probability matching and the formation of conservative decision rules in a numerical analog of signal detection. *Journal of Experimental Psychology: Human Learning and Memory*, *7*(5), 344–354. <https://doi.org/10.1037/0278-7393.7.5.344>
- Hogarth, R. M., & Einhorn, H. J. (1992). Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, *24*(1), 1–55. [https://doi.org/10.1016/0010-0285\(92\)90002-J](https://doi.org/10.1016/0010-0285(92)90002-J)

- Kaesler, M., Dunn, J. C., Ransom, K., & Semmler, C. (2020). Do sequential lineups impair underlying discriminability? *Cognitive Research: Principles and Implications*, 5(1), 35. <https://doi.org/10.1186/s41235-020-00234-5>
- Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review*, 80(4), 237–251. <https://doi.org/10.1037/h0034747>
- Kornbrot, D. E. (2006). Signal detection theory, the approach of choice: Model-based and distribution-free measures and evaluation. *Perception & Psychophysics*, 68(3), 393–414. <https://doi.org/10.3758/BF03193685>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, 4(1), 251524592095150. <https://doi.org/10.1177/2515245920951503>
- Lee, N., & Lee, S. (2022). Visualizing science: The impact of infographics on free recall, elaboration, and attitude change for genetically modified foods news. *Public Understanding of Science*, 31(2), 168–178. <https://doi.org/10.1177/096366252111034651>
- Lynn, S. K., & Barrett, L. F. (2014). “Utilizing” signal detection theory. *Psychological Science*, 25(9), 1663–1673. <https://doi.org/10.1177/0956797614541991>
- Maddox, W. T., & Bohil, C. J. (2005). Optimal classifier feedback improves cost-benefit but not base-rate decision criterion learning in perceptual categorization. *Memory & Cognition*, 33(2), 303–319. <https://doi.org/10.3758/BF03195319>
- Matyas, T. A., & Greenwood, K. M. (1990). Visual analysis of single-case time series: Effects of variability, serial dependence, and magnitude of intervention effects. *Journal of Applied Behavior Analysis*, 23(3), 341–351. <https://doi.org/10.1901/jaba.1990.23-341>
- Mena, P. (2023). Reducing misperceptions through news stories with data visualization: The role of readers’ prior knowledge and prior beliefs. *Journalism*, 24(4), 729–748. <https://doi.org/10.1177/14648849211028762>
- Modirrousta-Galian, A., & Higham, P. A. (2023). Gamified inoculation interventions do not improve discrimination between true and fake news: Reanalyzing existing research with Receiver Operating Characteristic analysis. *Journal of Experimental Psychology: General*, 152(9), 2411–2437. <https://psycnet.apa.org/doi/10.1037/xge0001395>
- Nyhan, B., & Reifler, J. (2019). The roles of information deficits and identity threat in the prevalence of misperceptions. *Journal of Elections, Public Opinion and Parties*, 29(2), 222–244. <https://doi.org/10.1080/17457289.2018.1465061>
- Obrecht, N. A., Chapman, G. B., & Gelman, R. (2007). Intuitive *t*-tests: Lay use of statistical information. *Psychonomic Bulletin & Review*, 14(6), 1147–1152. <https://doi.org/10.3758/BF03193104>
- Peterson, C. R., & Ducharme, W. M. (1967). A primacy effect in subjective probability revision. *Journal of Experimental Psychology*, 73(1), 61–65. <https://doi.org/10.1037/h0024139>
- Peterson, W. W., & Birdsall, T. G. (1953). The theory of signal detectability: Part I. The general theory. Electronic Defense Group, Technical Report, 13.
- R Core Team (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from <https://www.R-project.org/>.
- Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Soo, K. W., & Rottman, B. M. (2018). Causal strength induction from time series data. *Journal of Experimental Psychology: General*, 147(4), 485–513. <https://doi.org/10.1037/xge0000423>
- Vicente, L., Blanco, F., & Matute, H. (2023). I want to believe: Prior beliefs influence judgments about the effectiveness of both alternative and scientific medicine. *Judgment and Decision Making*, 18(1), 1–15. <https://doi.org/10.1017/jdm.2022.3>
- White, P. A. (2015). Causal judgements about temporal sequences of events in single individuals. *Quarterly Journal of Experimental Psychology*, 68(11), 2149–2174. <https://doi.org/10.1080/17470218.2015.1009475>
- White, P. A. (2017). Causal judgments about empirical information in an interrupted time series design. *Quarterly Journal of Experimental Psychology*, 70(1), 18–35. <https://doi.org/10.1080/17470218.2015.1115886>
- Witt, J. K., Taylor, J. E. T., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, 44(3), 289–300. <https://doi.org/10.1068/p7908>
- Witt, J. K., & Warden, A. C. (2021). Better sensitivity to linear and nonlinear trends with position than with color. *Journal of Vision*, 21(5), 1–13. <https://doi.org/10.1167/jov.21.5.12>
- Wixted, J. T., & Mickes, L. (2014). A signal-detection-based diagnostic-feature-detection model of eyewitness identification. *Psychological Review*, 121(2), 262–276. <https://doi.org/10.1037/a0035940>
- Ximenes, V. M., Manolov, R., Solanas, A., & Quera, V. (2009). Factors affecting visual inference in single-case designs. *The Spanish Journal of Psychology*, 12(2), 823–832. <https://doi.org/10.1017/S1138741600002195>
- Zhang, Y., & Rottman, B. M. (2023). Causal learning with interrupted time series data. *Judgment and Decision Making*, 18, e30. <https://doi.org/10.1017/jdm.2023.29>

**Cite this article:** Guthrie, E. C. and Bishara, A. J. (2025). When one graph judgment leads to another: Signal detection analysis of base rate effects. *Judgment and Decision Making*, e20. <https://doi.org/10.1017/jdm.2025.4>