Normative evidence weighting and accumulation in correlated environments

Nathan Tardiff[1,2,3]

Jiwon Kang[3]

Joshua I. Gold[3]

[1] Department of Otorhinolaryngology, Perelman School of Medicine, University of Pennsylvania, United States

[2] Department of Psychology, New York University, United States

[3] Department of Neuroscience, Perelman School of Medicine, University of Pennsylvania, United States

* For correspondence: ntardiff@sas.upenn.edu

Abstract

The brain forms certain deliberative decisions following normative principles related to how sensory observations are weighed and accumulated over time. Previously we showed that these principles can account for how people adapt their decisions to the temporal dynamics of the observations (Glaze et al., 2015). Here we show that this adaptability extends to accounting for correlations in the observations, which can have a dramatic impact on the weight of evidence provided by those observations. We tested online human participants on a novel visual-discrimination task with pairwise-correlated observations. With minimal training, the participants adapted to uncued, trial-by-trial changes in the correlations and produced decisions based on an approximately normative weighting and accumulation of evidence. The results highlight the robustness of our brain's ability to process sensory observations with respect to not just their physical features but also the weight of evidence they provide for a given decision.

In their efforts to break the Enigma code during World War II, Alan Turning and his colleagues at Bletchley Park recognized the importance of the concept of a "weight of evidence" for making decisions: noisy or ambiguous evidence is most useful when the influence or weight it has on the ultimate decision depends on its uncertainty. For the case of two alternatives, they used a weight of evidence in the form of the logarithm of the likelihood ratio (i.e., the ratio of the likelihoods of each of the two alternative hypotheses, given the observations), or logLR. The logLR later became central to the sequential probability ratio test (SPRT), which was proven to provide certain optimal balances between the speed and accuracy of such decisions (Barnard, 1946; Wald, 1947; Wald and Wolfowitz, 1948). Recognizing the general nature of this formulation, Turing and colleagues noted that the logLR would be "an important aid to human reasoning and … eventually improve the judgment of doctors, lawyers, and other citizens" (Good, 1979).

The logLR, or scaled versions of it, has since become ubiquitous in models of decision-making, including sequential-sampling models related to the SPRT like the drift-diffusion model (DDM). These models capture many behavioral and neural features of human and animal decision-making for a broad range of tasks (Gold and Shadlen, 2007; Smith and Ratcliff, 2004). However, it is still unclear whether decision-makers compute the normative weight of evidence—the logLR—or instead rely on approximations or other heuristics (Brown et al., 2009; Hanks et al., 2011; Ratcliff et al., 2016; Ratcliff and McKoon, 2008). Furthermore, previous findings have come from studies that have been restricted mainly to tasks that simplify the computation of the logLR by providing only observations that are statistically independent of each other. Under these conditions, the logLR can be computed separately for each observation, often by scaling the strength of that observation by its signal and noise characteristics learned from past observations from the same source (Fig. 1a). These weights of evidence are added together (i.e.,

the logLRs are accumulated over time) to form the aggregated decision variable that governs the final choice.

More generally, however, non-independence in the statistics of the observations can have substantial effects on how those observations should be weighed to form effective decisions (Fig. 1b,c). If not accounted for appropriately, these effects can cause a decision-maker to over- or under-estimate the weight of available evidence and make suboptimal decisions. Such suboptimalities have real-world consequences. For example, misestimation of correlation patterns in mortgage defaults is thought to have played a role in triggering the global financial crisis of 2008 (Salmon, 2009). Neglecting correlations can contribute to false beliefs and ideological extremeness in social and political settings (Denter et al., 2021; Glaeser and Sunstein, 2009; Levy et al., 2022; Ortoleva and Snowberg, 2015). Likewise, correlations in the physical environment should, in principle, be leveraged to support perception (Geisler, 2008; Parise, 2016). Yet whether and how people account for correlations when making perceptual decisions is not well understood.

The goal of this preregistered study (https://osf.io/qj92c) was to test how humans form simple perceptual decisions based on observations with different degrees of correlation. We previously showed that both theoretically optimal (i.e., an ideal observer that maximizes decision accuracy) and human observers flexibly adjust how evidence is accumulated over time to account for the temporal dynamics of the sequentially presented observations (Glaze et al., 2015; Veliz-Cuba et al., 2016). Here we assess how both ideal and human observers weigh and accumulate evidence that is based on pairs of correlated observations (Fig. 2). Under these conditions, the normative decision process involves computing a weight of evidence by scaling each paired observation by a function of the underlying correlation, then accumulating that weight of evidence across pairs

until reaching a predefined bound that, like in the DDM, balances decision speed and accuracy. As we detail below, we found that people tend to follow these normative principles, accounting appropriately for the correlations (albeit based on slight misestimates of correlation magnitude) and demonstrating the robustness and flexibility with which our brains can appropriately weigh and accumulate evidence when making simple decisions.

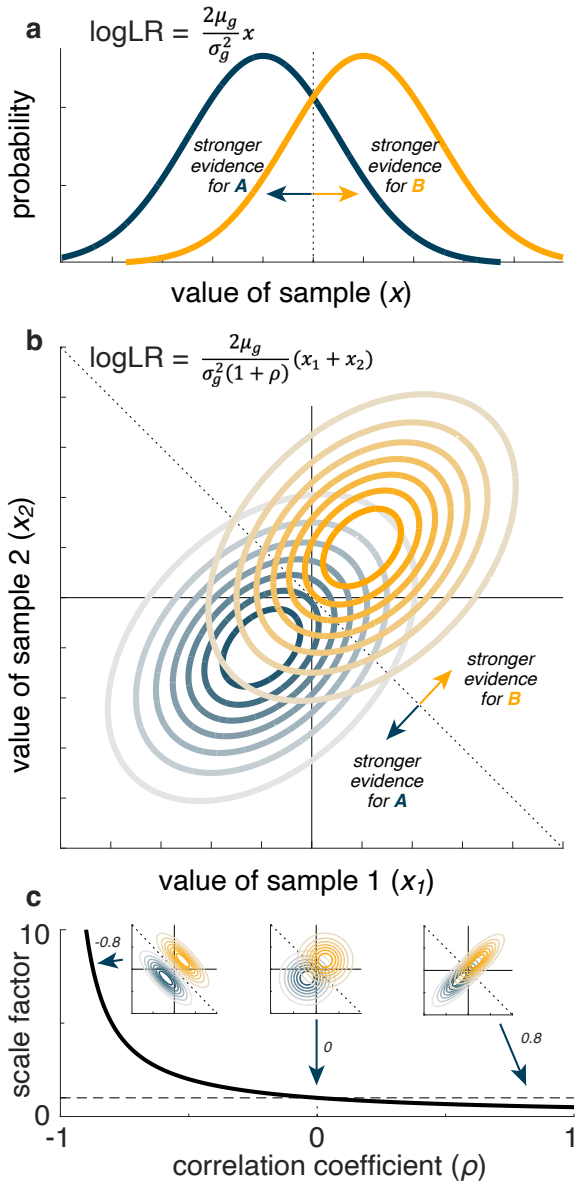

Figure 1. Illustration of how pairwise correlations can affect the weight of evidence (logLR) for the generative source of an observation. **a** Computing the logLR when the observation (x) is a single sample from one of two one-dimensional Gaussian distributions (labeled ***A*** and ***B***), with means $\pm\mu_g$ and equal variances (σ_g^2) (Gold and Shadlen, 2001). **b** Computing the logLR when the observation (x_1, x_2) is a pair of samples from one of two pairs of one-dimensional Gaussian distributions (labeled ***A*** and ***B***), with means $\pm\mu_g$, equal variances (σ_g^2), and correlation between the two Gaussians = ρ. **c** The normative scaling ($\frac{1}{1+\rho}$ term in **b**) of the observation plotted as a function of correlation sign and magnitude. The dashed horizontal line corresponds to scale factor = 1, which occurs at $\rho = 0$. The insets show three example pairs of distributions with different correlations, as indicated. The dotted lines in **a**, **b**, and the insets in **c** indicate the optimal decision boundary separating evidence for ***A*** versus ***B***.

Results

We tested 100 online participants performing a novel task that required them to form simple decisions about which of two latent sources generated the observed visual stimuli, in the presence of different correlation structures in the stimuli (Fig. 2). The task design was based on principles illustrated in Fig. 1: the normative weight of evidence for the identity of a source of paired observations from correlated, Gaussian random variables depends systematically on the magnitude and sign of the correlation (Fig. 1b). In this case, negative pairwise correlations provide, on average, stronger evidence with increasing correlation magnitude, because less overlap of the generative source distributions allows them to be more cleanly separated by the decision boundary (Fig. 1c, left inset). Positive pairwise correlations provide, on average, weaker evidence with increasing correlation magnitude, because more overlap of the generative source distributions causes them to be less cleanly separated by the decision boundary (Fig. 1c, right inset).

Participants reported the generative source (left or right) of noisy observations, depicted onscreen as the position of stars along a horizontal line (Fig. 2a). Observations were presented in

pairs. Each element of the pair had the same mean value across samples from the generative source, but the noise correlation within pairs was manipulated on a per-trial basis. We assigned each participant to a correlation-magnitude group ($|\rho|$ = 0.2, 0.4, 0.6, 0.8; 25 participants per group) in which the pairwise correlation on a given trial was drawn from three conditions: $-\rho$, 0, or $+\rho$. We equated task difficulty across participants by calibrating the means of the generative distributions (see Methods). We interleaved randomly the three correlation conditions with the two sources (left, right) and two levels of task difficulty (low, high), for 12 total conditions for each correlation-magnitude group.

Crucially, we adjusted the means of the generative distributions to ensure that the expected logLR (which we term the *evidence strength*) was constant across correlation conditions (Fig. 2b,c). For example, because negative correlations increase logLR (Fig. 1c), we used smaller differences in means for the negative-correlation conditions than for the zero-correlation conditions. As a result, we expect participants who make decisions by weighing the evidence according to the true logLR to produce identical distributions of choices and response times (RTs) across correlation conditions. In contrast, we expect participants who ignore the correlations to under-weigh the evidence provided by negative-correlation pairs (and thus take longer to accumulate evidence, leading to longer RTs and higher accuracy) and over-weigh the evidence provided by positive-correlation pairs (leading to shorter RTs and lower accuracy). We further expect strategies between these two extremes to have more mixed effects on choices and RTs, as we detail below.

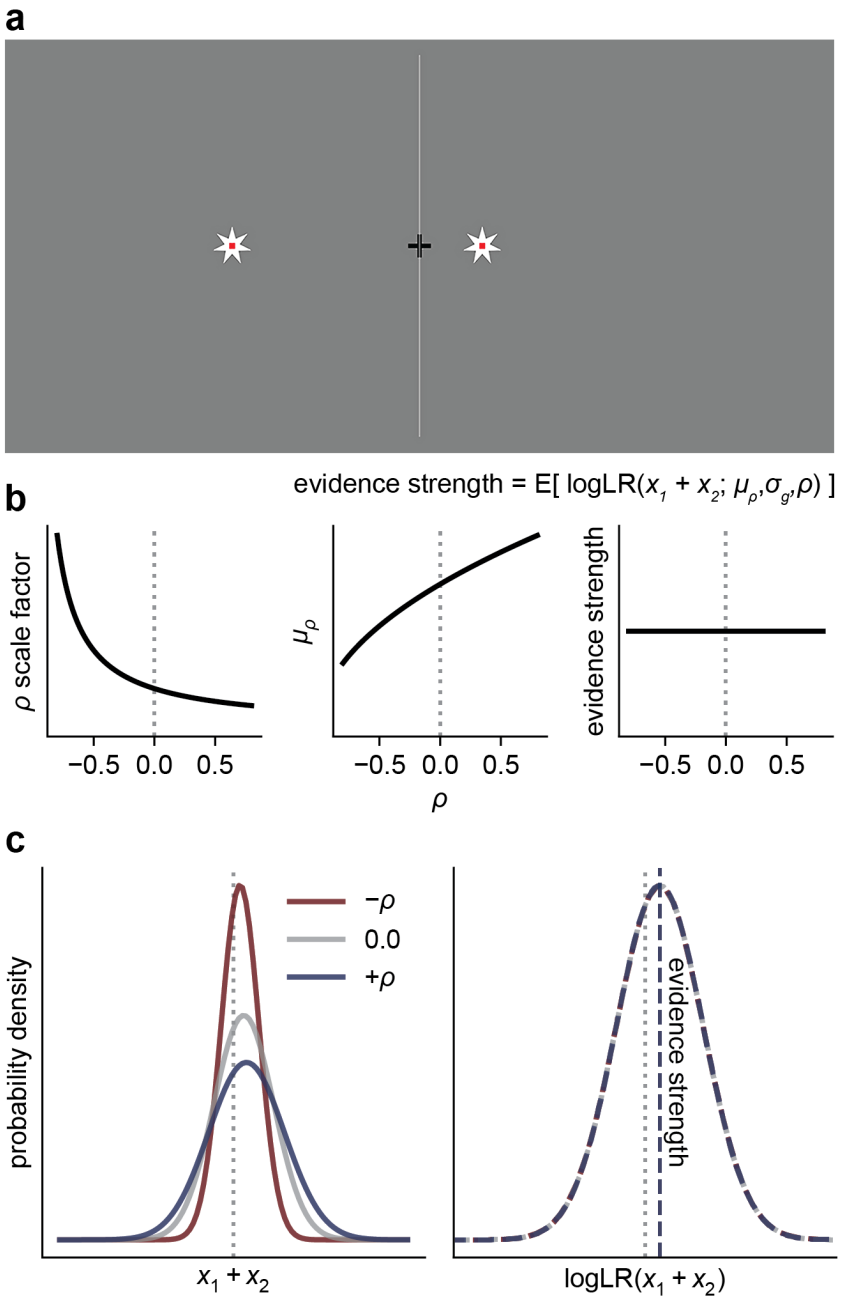

Figure 2. Task. **a** Human observers viewed pairs of stars (updated every 0.2 sec) and were asked to decide whether the stars were generated by a source on the left or right side of the screen. An example star pair is shown. The horizontal position of each star pair was drawn from a bivariate Gaussian distribution, with a mean and correlation that varied from trial-to-trial. **b** Because the normative correlation-dependent scale factor that converts observations to evidence (logLR) increases as the correlation decreases, we manipulated the mean of the generative distribution such that the expected logLR (evidence strength) was fixed across correlation conditions. **c** The generative distributions of the sum of individual star pairs, for three example correlation conditions. Decreasing the correlation has the effect of decreasing the standard deviation of the

sum distribution. By adjusting each correlation-specific generative mean (μ_ρ) in proportion to the correlation-dependent change in the standard deviation from the zero-correlation condition (i.e., μ_ρ=$\mu_0\sqrt{1+\rho}$), the true logLR distribution (i.e., of an ideal observer) is invariant to the correlation, and thus evidence strength remains fixed. Note that the sum-of-pairs distribution is equivalent to the bivariate distribution for the purposes of computing the logLR (see Methods).

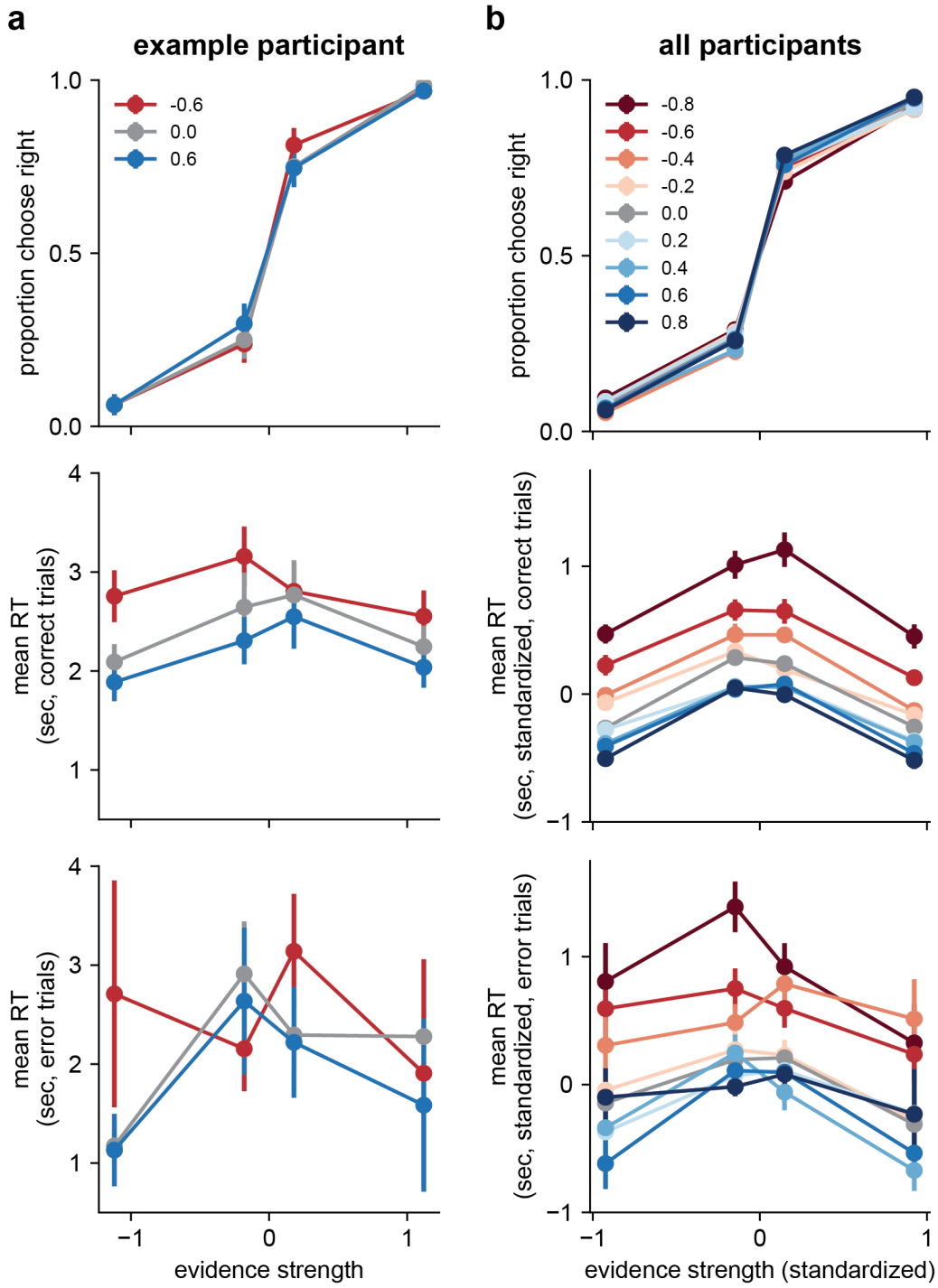

Figure 3. Effects of correlations on choice and RT. **a** Data from an example participant from the 0.6 correlation-magnitude group. Top: choices plotted as a function of evidence strength (abscissa) and correlation condition (see legend). Middle, Bottom: Mean RTs for correct and error trials, respectively. Error bars are within-participant standard errors of the mean (SEM). **b** Same as **a**, but data are averaged across all participants. Evidence strength was standardized to equal the mean evidence strength (expected logLR) for each condition, across participants. RT

was standardized by subtracting each participant's mean RT in the zero-correlation condition, separately for correct and error trials. Points and error bars are across-participant means and SEMs, respectively.

Human response times are influenced by correlated observations

The example participant in Fig. 3a exhibited behavioral patterns that were illustrative of the overall trends we observed. Specifically, their choice accuracy was affected by evidence strength (higher accuracy for stronger evidence) but not correlation (this participant was tested using correlation values of −0.6, 0.0, and 0.6). In contrast, their RTs were affected by both evidence strength and correlation, including faster responses on correct trials using stronger evidence and more positive correlations.

Likewise, across our sample of participants choices depended strongly on evidence strength but not correlation (Fig. 3b). Logistic models fit to individual participant's evidence-strength-dependent psychometric data demonstrated no benefit to fitting separate models per correlation condition versus a single model fit jointly to all three correlation conditions (mean ΔAIC = −4.14, protected exceedance probability [PEP] = 1.0 in favor of the joint model). This result held true at each correlation magnitude individually (all mean ΔAIC < −2.0, all PEP > 0.8).

In contrast, RTs were affected by both evidence strength and correlation, with a tendency of participants to respond faster for stronger evidence and more-positive correlations (Fig. 3b). A linear mixed-effects model fit to median RTs from correct trials confirmed these observations, indicating effects of evidence strength ($F(1,98.00) = 174.24$, $p < 0.001$), the sign of the correlation (negative, zero, positive) within participants ($F(2,131.56) = 219.96$, $p < 0.001$), and the interaction between the sign of the correlation and its magnitude between participants ($F(2, 131.56) = 81.04$, $p < 0.001$). That is, the effects of correlations on correct RTs were more

pronounced in participants tested using stronger correlations. Similar effects were also present on error trials (evidence strength: ($F(1,960.54) = 19.21$, $p < 0.001$), sign of correlation: ($F(2, 234.74) = 58.41$, $p < 0.001$), correlation sign *x* magnitude ($F(2, 233.48) = 13.50$, $p < 0.001$).

In short, the patterns of choice data that we observed were consistent with decisions that took into account the correlations in the observations, which by design were necessary to equate the weight of evidence across correlation conditions. In contrast, the patterns of RT data that we observed rule out the possibility that the participants used normative evidence weighting (i.e., based on the true logLR) to make their decisions, because if they did, their RTs would not depend on the correlation condition. These findings leave open a broad range of possible weighing strategies between the two extremes of an ideal observer and a naïve observer who ignores the correlations (i.e., assumes independence). The analyses detailed below aimed to more precisely identify where in that range our participants' strategies fell.

RTs are consistent with a decision bound on approximate logLR

We analyzed the RT data in more detail, based on principles from the DDM and SPRT, wherein simple decisions result from a process in which evidence is accumulated over time until reaching one of two fixed bounds, corresponding to the two choices. This process governs both the choice (which bound is reached first) and RT (when the bound is reached). We considered two forms of evidence with different scale factors (i.e., the scale value multiplied by each star position to compute its weight of evidence): 1) "naïve" were proportional to the generative mean, μ_g, of a pair of samples, which is a standard assumption in many implementations of the DDM (Palmer et al., 2005) but in this case ignores the correlations and thus does not produce a weight of

evidence equivalent to the true logLR; and 2) "true" were proportional to $\frac{\mu_g}{1+\rho}$, which takes into account the correlations and produces a weight of evidence equivalent to the true logLR.

Because we designed the task to present stimuli with equal expected logLR (evidence strength) across correlation conditions, decisions based on an accumulation of the true logLR to a fixed bound would have similar mean RTs across correlation conditions. In contrast, decisions based on an accumulation of the naïve logLR would be expected to have different effects for positive versus negative correlations. Ignoring positive correlations is equivalent to ignoring redundancies in the observations, which would lead to over-weighting the evidence and thus reaching the bound more quickly, corresponding to shorter RTs and lower accuracy. Ignoring negative correlations is equivalent to ignoring synergies in the observations, which would lead to under-weighting the evidence and thus reaching the bound less quickly, corresponding to longer RTs and higher accuracy (Fig. 1c).

As noted above, participants had RTs that were, on average, either relatively constant or slightly decreasing as a function of increasing correlations, particularly for larger correlations (Fig. 4a,b). These trends were not consistent with a decision process that used a fixed bound that ignored correlations. They also were not completely consistent with a decision process that used a fixed bound on the true logLR (because of the trend of decreasing RTs with increasing correlations). Instead, these results could be matched qualitatively to simulations that made decisions based on an approximation of logLR computed using underestimates of the correlation-dependent scale factor (Fig. 4c). We examined this idea more quantitatively using model fitting, detailed in the next section.

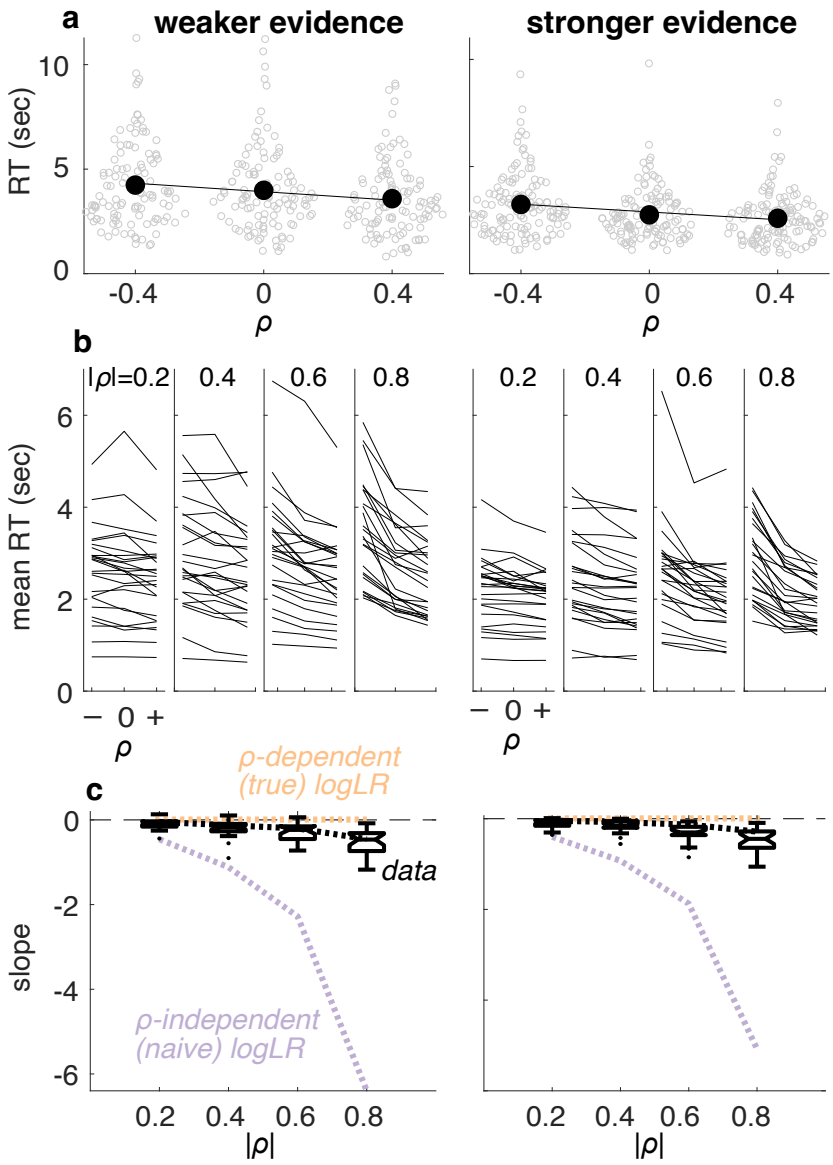

Figure 4. RTs were consistent with a bound on (approximate) logLR. **a** RTs measured from an example participant for the weaker (left) and stronger (right) evidence conditions. Unfilled points are data from individual trials. Filled points are means, lines are linear fits to those means. **b** Summary of mean RT versus correlation for all participants and conditions. Correlation-magnitude group is indicated at the top of each panel. Lines are data from individual participants. **c** Summary of slopes of linear fits to mean RT versus correlation for individual participants (as in **a**). Box-and-whisker plots show median, interquartile range, 90th percentiles, and outliers as a function of correlation-magnitude group. Colored lines are predicted relationships for decisions based on an accumulation of evidence to a fixed bound, where the weight of evidence was computed as correlation-independent (naïve) and correlation-dependent (true) logLR. The data are roughly consistent with decision processes that, on average, used a correlation-dependent logLR but based on a slight underestimate of the correlation-dependent scale factor (computed using $\frac{0.7}{1+\rho}$; black dashed lines).

Correlation-dependent adjustments affect the bound in a drift-diffusion model

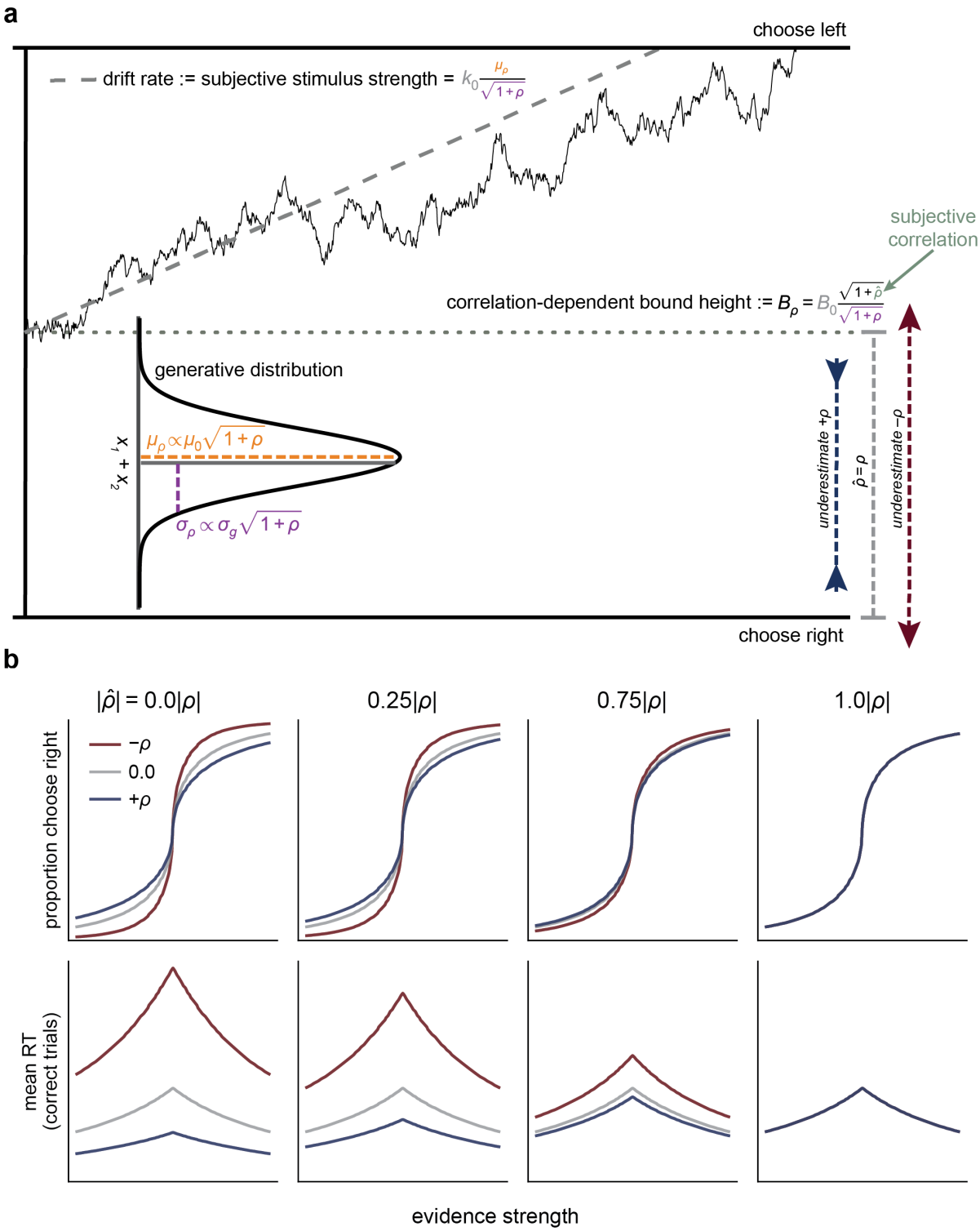

Fig. 5 A drift-diffusion model (DDM) captures normative evidence weighting via bound-height adjustments. **a** In the DDM, sensory observations are modeled as samples from a Gaussian distribution (in the continuum limit). Evidence is accumulated over time as the decision variable until it reaches one of the two bounds, which terminates the decision in favor of the choice corresponding to that bound (here for simplicity we show fixed bounds, but in the fitting detailed below we use collapsing bounds). For pairs of correlated observations, altering the correlation between the pairs is equivalent to changing the standard deviation of the generative distribution of the sum of each pair, which affects the drift rate plus the scaling of the bound height (see Methods). Normative evidence weighting in the DDM corresponds to correlation-dependent adjustments of the bound height (the decision rule) to account for the changes in the generative standard deviation. These changes in the bound height are functionally equivalent to scaling the observations to compute the true logLR. **b** Predictions from the DDM. Colors correspond to three simulated correlation conditions (see legend). Other parameters were chosen to approximate those found in fits to human data. Each column depicts predictions based on the same form of correlation-dependent bound scaling (see **a** and below) but with a different subjective correlation $\hat{\rho}$ (i.e., the correlation assumed by the observer), which was computed as a proportion of the objective correlation ρ (computed on Fisher-*z*-transformed correlations that were then back-transformed). Given equal expected logLR across correlation conditions, underestimating the correlation ($|\widehat{\rho}| < |\rho|$, first three columns) leads to performance differences between the conditions, where the magnitude of the differences is a function of the degree of underestimation. Only $\hat{\rho} = \rho$ (rightmost column) produces equal predicted performance across conditions.

To better understand how the participants formed correlation-dependent decisions, we developed a variant of the DDM that can account for pairwise-correlated observations. The DDM jointly accounts for choices and RTs according to a process that accumulates noisy evidence over time until reaching a decision bound (Fig. 5a). The model includes two primary components that govern the decision process and can be based on just two free parameters. The *drift rate* governs the average rate of information accumulation (Palmer et al., 2005; Ratcliff and McKoon, 2008). This term typically depends on the product of the strength or quality of the sensory observations (generally varied via the mean, or signal, of the evidence distribution, μ_g) and the decision-maker's sensitivity to those observations (the fit parameter *k;* i.e., $drift\ rate(\rho) = k\mu$. The *bound height* governs the decision criterion, or rule, which corresponds to the amount of evidence required to make a decision and controls the trade-off between decision speed and

accuracy (Heitz, 2014). This term is often just a single fit parameter representing symmetric bounds; i.e., *bound height* = B.

For our task, changing ρ affects the observation distribution (inset in Fig. 5a) in terms of both signal (the different values of μ_ρ we used to offset the effect of the correlation affected the "drift" part of "drift-diffusion") and noise (the different values of σ_ρ that reflect effects of correlations on the standard deviation of the sum-of-pairs distribution affect the "diffusion" part of "drift-diffusion"). To account for these effects in the DDM, which typically expresses both the drift rate and bound height in terms of the signal-to-noise ratio of the observation distribution (Palmer et al., 2005; Shadlen et al., 2006), we scaled the drift rate by the correlation-dependent component of the evidence noise:

$$drift\ rate(\rho) = k_0 \frac{\mu_\rho}{\sqrt{1+\rho}} = k_0 \mu_0,$$

where k_0 is the drift-rate parameter for the zero-correlation condition, which also includes the correlation-independent component of the observation noise, σ_g (i.e., $k_0 = \frac{k}{\sigma_g}$). Note that the correlation-dependent scale factor used here is the same factor we used to scale the generative means of the stars to ensure equal logLRs across correlation conditions ($\mu_\rho = \mu_0\sqrt{1+\rho}$). As a result, these factors cancel in the drift-rate equation above, which depends only on subjective (fit) sensitivity (k_0) and the objective strength of evidence expressed in terms of the zero-correlation condition (μ_0).

Implemented this way, scaling the strength of observations to compute a weight of evidence is equivalent to scaling the bound height to account for the correlation-dependent effects on both signal and noise, which again cancel (akin to the transformation in Fig. 2c, from the left to the

right panel; see Methods). However, our participants did not necessarily know the objective correlation (ρ) but instead relied on subjective internal estimates ($\hat{\rho}$). To account for possible misestimates, we assumed that their decision bounds were scaled by subjective estimates of correlation-dependent changes in the noise:

$$B_\rho = B_0 \frac{\sqrt{1+\hat{\rho}}}{\sqrt{1+\rho}},$$

where B_0 is the bound height for the zero-correlation condition, the numerator is the subjective component of the correlation-dependent scale factor, and the denominator reflects effect of the correlation on the diffusion process, as in the drift rate equation above. This formulation leads to the following predictions (Fig. 5b):

- If $\hat{\rho} = \rho$, then $B_\rho = B_0$: when correlations are estimated accurately, the drift rate and bound height are equal across correlations, giving equal average choices and RTs (e.g., Fig. 5b, right-most column).
- If $\hat{\rho} < \rho$ and $\rho > 0$, then $B_\rho < B_0$: when positive correlations are underestimated, the bound is lower than optimal because the evidence has been over-weighed, and average performance is faster but less accurate.
- If $\hat{\rho} < \rho$ and $\rho < 0$, then $B_\rho > B_0$: when negative correlations are underestimated, the bound is higher than optimal because the evidence has been under-weighed, and average performance is slower but more accurate.

In short, failing to (fully) account for correlations causes the decision-maker to set their bound according to a misestimate of the weight of evidence provided by each observation. The deviation of the bound height from the normative adjustment in turn alters the speed-accuracy tradeoff, with a greater qualitative effect on RTs compared to accuracy.

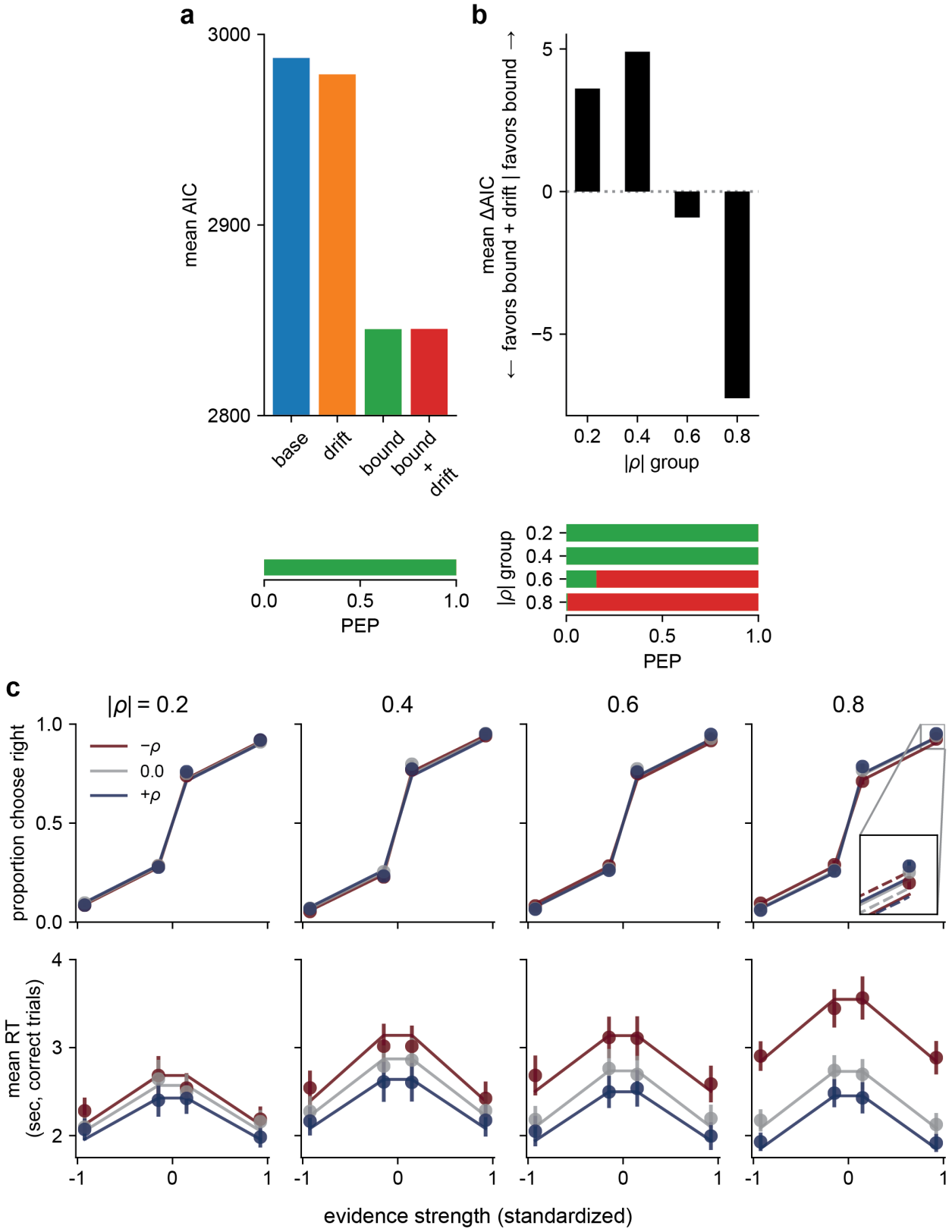
a
3000
2900
2800
mean AIC
base
drift
bound
bound + drift
b
5
0
−5
mean ΔAIC
← favors bound + drift | favors bound →
0.2
0.4
0.6
0.8
|ρ| group
0.0
0.5
1.0
PEP
|ρ| group
0.2
0.4
0.6
0.8
0.0
0.5
1.0
PEP
c
|ρ| = 0.2
0.4
0.6
0.8
1.0
0.5
0.0
proportion choose right
−ρ
0.0
+ρ
4
3
2
mean RT
(sec, correct trials)
-1
0
1
evidence strength (standardized)

Fig. 6. A DDM accounts for human behavior. **a** Model comparison: mean AIC (top) and protected exceedance probability (PEP; bottom), across all task conditions, for four different models, as labeled (see text for details). **b** Model comparison within each correlation-magnitude group, showing the difference in AIC between the bound and bound+drift models (top) and PEP (bottom). Bar colors in the PEP plots correspond to the model colors in the top panel of **a**. **c** Predictions from the DDM (lines) plotted against participant data (points) for choice (top) and RT (bottom) for each correlation-magnitude group (columns, labels at top). Predictions and data are averaged across participants. Colors correspond to the three correlation conditions (see legend). Error bars are SEM. Model predictions are derived from the best-fitting model at each magnitude (see **b**). The inset for the 0.8 group compares the fit of the bound+drift model (solid lines) to the bound model (dashed lines).

Human performance is consistent with correlation-dependent bound adjustments

Qualitatively, participants' patterns of choice and RTs across correlation conditions were consistent with bound-height adjustments that depended on slight underestimates of the objective correlation (compare Figs. 3b and 5b). To examine this idea more quantitatively and rule out alternative possibilities, such as adjustments to the drift rate, we fit four DDMs to each participant's data: 1) a *base* model, with no adjustment based on the correlation; 2) a *drift* model, with separate drift parameters for the negative, zero, and positive correlation conditions (k_{-}, k_{0}, k_{+}, respectively); 3) a *bound* model, which implements the normative bound scaling, where $\hat{\rho}_{-}$ and $\hat{\rho}_{+}$ are free parameters that estimate the subjective correlation in the negative and positive correlation conditions, respectively; and 4) a *bound+drift* model, which combines the adjustments from the drift and bound models. Prior to fitting these models, we confirmed that the DDM can generally account for choice and RTs in our novel task via fits to the zero-correlation condition. These fits confirmed that the DDM could account for the data, and that fits were improved with the addition of a linear collapsing bound, which we used in all subsequent fits (Figure 6—figure supplement 1).

These model fits indicated that correlation-dependent bound adjustments were more important than correlation-dependent drift adjustments for capturing differences in behavioral performance between correlation conditions (Fig. 6a, compare *bound* model to *base* and *drift* models). However, drift-rate adjustments were also useful for the high, but not low, correlation magnitudes (model comparison per group shown in Fig. 6b; best-fitting model per group shown in Fig. 6c; between-group random-effects model comparison ($|\rho| < 0.5$ versus $|\rho| > 0.5$), $p < 0.001$; Rigoux et al., 2014). These drift-rate adjustments at higher correlation magnitudes accounted for the fine-scale ordering of choice behavior across correlation conditions, which was opposite to that predicted by the *bound* model (Fig. 6c, 0.8 correlation inset). A likely explanation for this effect relates to our task design, which involved equating evidence strength (expected logLR) across correlation conditions via changes in the generative mean of the star positions that were necessarily more dramatic for higher versus lower correlation magnitudes (see Fig. 1c). If perceived star position is not a linear function of objective star position, different drift parameters per correlation would provide a better fit (Palmer et al., 2005). Put another way, the very small generative means in the −0.6 and −0.8 correlation conditions seem to provide weaker observed stimulus strength than predicted by a linear function between screen position and stimulus strength, leading to smaller drift parameters in these conditions (Figure 6—figure supplement 2).

Human performance is consistent with a weight of evidence based on the approximated correlation

As predicted, bound adjustments best accounted for behavioral differences across correlation conditions. Because these adjustments took the form of the normative bound scaling with a subjective correlation, we leveraged the fit correlations to ask how well participants estimated

the correlation, and whether deviations from equal performance across correlation conditions were caused by underestimation. To ensure that correlation effects on drift rate were accounted for in these analyses, we used the best-fitting model (either *bound* or *bound+drift*) per correlation-magnitude group.

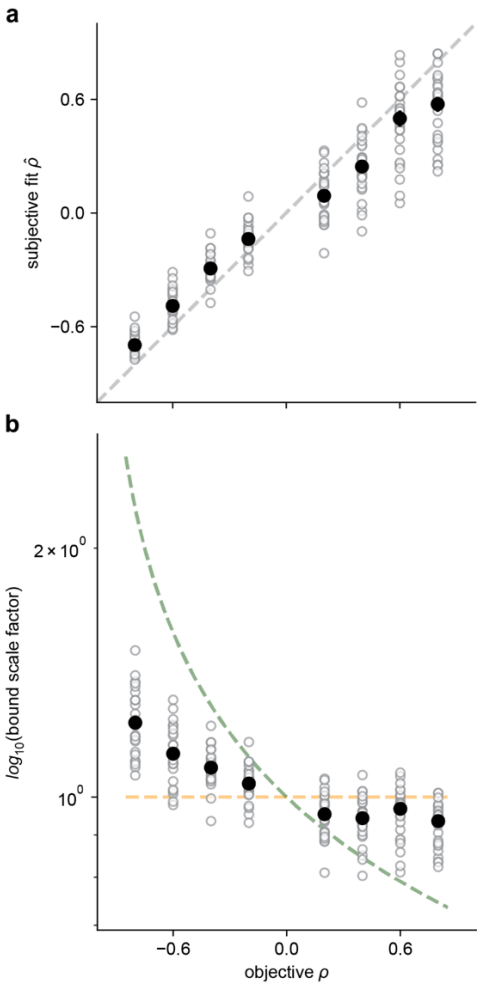

Fig. 7. Participants used near-optimal correlation estimates, with slight biases away from extreme values. **a** The subjective fit correlation ($\hat{\rho}$) from the DDM as a function of the objective correlation (ρ). Open circles are the fits from individual participants. Closed circles are the averages per correlation condition (averages were computed on Fisher *z*-transformed values and then back-transformed). Error bars (not visible in most cases) are SEM. The dashed line is the unity line. **b** The same data as in **a**, but plotted with the correlation-dependent bound scale factor ($\frac{\sqrt{1+\hat{\rho}}}{\sqrt{1+\rho}}$) on the ordinate. The orange dashed line corresponds to the normative scale factor. The green dashed line is the scale factor for a naïve observer that assumes zero correlation.

We found a strong relationship between the objective and subjective correlations (Fig. 7a; B = 0.71, $t(99) = 40.93$, $p < 0.001$, Fisher z-transformed scale), confirming that participants were sensitive to the correlations and used them to adjust their decision process in a near-optimal manner. However, the slope of this relationship was less than one. That is, participants underestimated the objective correlation, on average (test of $\rho - \hat{\rho}$ (Fisher z-transformed scale): B = −0.18, $t(99) = -12.18$, $p < 0.001$), consistent with our hypothesis that their deviations from normative behavior (i.e., unequal performance across correlation conditions) resulted from systematic underestimates of the generative correlation. These estimates also tended to be more variable for positive versus negative correlations (mean value of the standard deviation of Fisher z-transformed estimates across positive-correlation conditions = 0.22, 0.10 for negative correlation conditions), likely reflecting the weaker consequences of misestimating positive versus negative correlations on performance (Fig. 1c).

These biased, subjective correlation estimates resulted in slight but systematic deviations from optimal of the corresponding inferred decision bounds used by the participants. Specifically, inferred bounds deviated from optimal, on average (Fig. 7b; test of $\log_{10}$(bound scale factor): $B = -0.037$, $t(198) = -14.83$, $p < 0.001$). Because of the non-linear relationship between the correlation and the normative scale factor (Fig. 1c), the inferred bound scaling tended to be closer to optimal for positive correlations compared to negative correlations ($B = 0.021$, $t(198) = 4.26$, $p < 0.001$).

These correlation-dependent differences in the decision process did not seem to reflect ongoing adjustments that might involve, for example, feedback-driven learning specific to this task. In particular, the participants tended to exhibit some learning over the course of the task, involving

substantial decreases in RT (the mean ± sem difference in RT between the first and second half of the task, measured across participants, was 0.71 ± 0.06 sec, respectively, Mann-Whitney test for H_0: median difference = 0, $p < 0.001$) at the expense of only slight decreases in accuracy (0.02 ± 0.00% correct, $p = 0.004$). These trends reflected a tendency to use slightly higher drift rates (Fig. 8a) and lower decision bounds (Fig. 8b) in the latter half of the task, a pattern of results that is consistent with previous reports of practice effects for simple decisions (Balci et al., 2011; Dutilh et al., 2009). However, these adjustments were not accompanied by similar, systematic adjustments in the participants' subjective correlation estimates, which were similar in the first versus second half of the task (Fig. 8c). This conclusion was supported by a complementary analysis showing that linear changes in RT as a function of trial number within a session tended to be the same for positive- and negative-correlation trials, as expected for stable relationships between correlation and RT (Wilcoxon rank-sum test for H_0: median difference in slope = 0, $p < 0.05$ for just one of 8 evidence strength *x* correlation magnitude conditions, after accounting for multiple comparisons via Bonferroni correction). Thus, participants' decisions appeared to be based on relatively stable estimates of the stimulus correlations that could be determined and used effectively on a trial-by-trial basis.

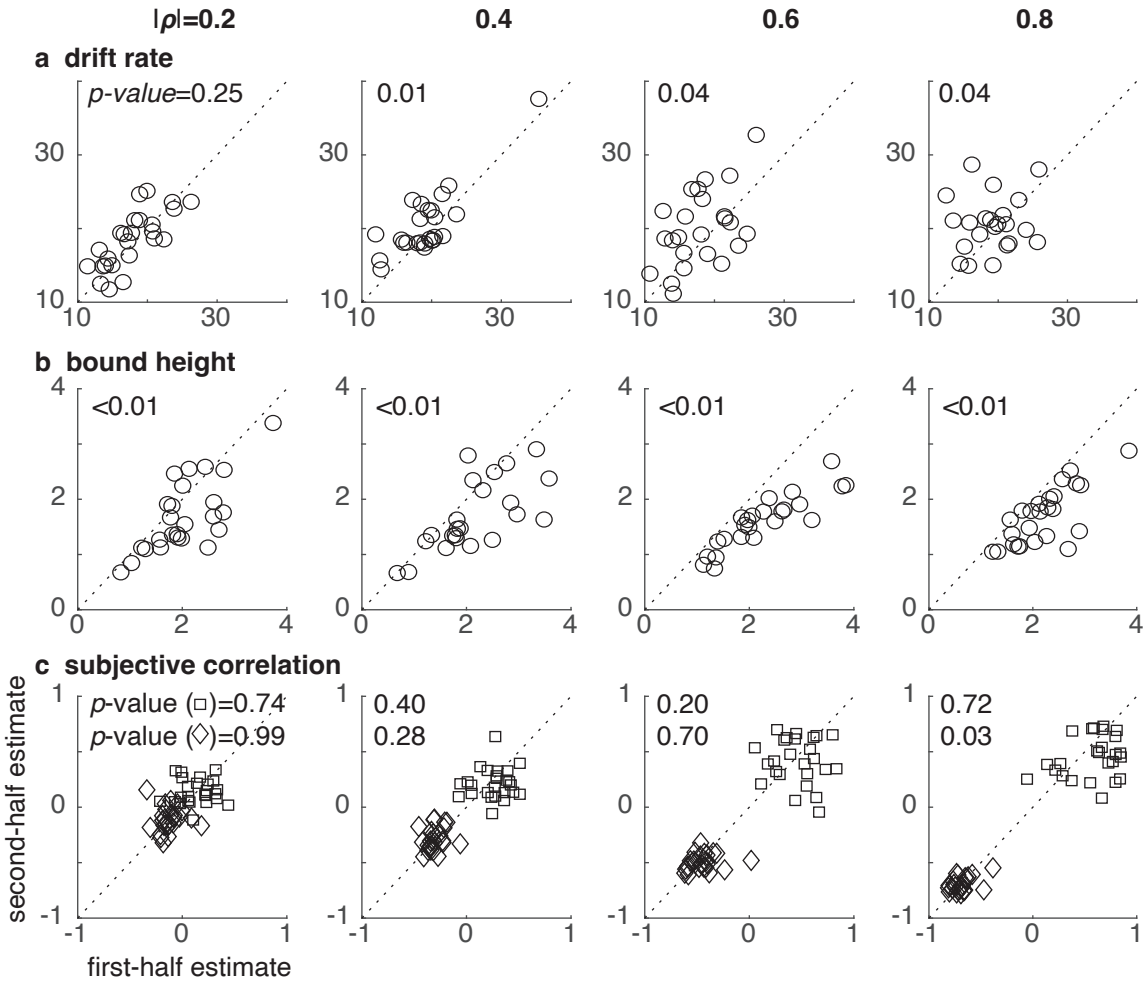

Fig. 8. Participants used stable estimates of the correlations, even as they adjusted other components of the decision process over the course of a session. Each panel shows a scatterplot of DDM parameters estimated using the first (abscissa) versus second (ordinate) half of trials from a given participant, from the best-fitting model per correlation-magnitude group. Points are data from individual participants. Columns are correlation condition, and rows are **a,** drift rate, k_0, and **b,** bound height, B. **c** shows estimates of positive ($\hat{\rho}_+$; squares) and negative ($\hat{\rho}_-$; diamonds) subjective correlations. P-values are for a Wilcoxon rank-sum test for H_0: median difference between the first- and second-half parameter estimates across participants = 0, uncorrected for multiple comparisons.

Discussion

This preregistered study addressed a fundamental question in perceptual decision-making: how do people convert sensory observations into a weight of evidence that can be used to form a decision about those observations? This question is important because evidence weighting affects how information is combined and accumulated over multiple sources and over time, ultimately governing the speed and accuracy of the decision process (Bogacz et al., 2006; Wald and Wolfowitz, 1948). To answer this question, we focused on correlations between observations, which are common in the real world, often ignored in laboratory studies, and can have a dramatic impact on the amount of evidence provided by a given set of observations. For simple perceptual decisions with correlated observations, the normative weight of evidence that accounts for these correlations can be expressed as a logLR. We showed that human participants make decisions that are approximately consistent with using this normative quantity, mitigating unintended shifts of the speed-accuracy tradeoff that would result from ignoring correlations. Below we discuss the implications of these findings for our understanding of the computations and mechanisms the brain uses to form simple decisions.

Previous support for the idea that human decision-makers can weigh evidence following normative principles has come from two primary lines of research. The first is studies of perceptual cue combination. Perceptual reports based on cues from multiple sensory modalities, or multiple cues from the same modality, often reflect weights of evidence that scale with the relative reliability of each cue, consistent with Bayesian theory (Ernst, 2005; Noppeney, 2021). The second is studies of evidence accumulation over time. The relationship between speed and accuracy for many decisions can be captured by models like the DDM that assume that the

underlying decision process involves accumulating quantities that are often assumed to be (scaled) versions of the logLR (Bogacz et al., 2006; Edwards, 1965; Gold and Shadlen, 2001; Laming, 1968; Stone, 1960).

Central to the interpretation of these studies, and ours, is understanding the scale factors that govern evidence weights. In their simplest forms, these scale factors are scalar values that are multiplied by stimulus strength to obtain the weight of evidence associated with each observed stimulus. These weights are then combined (e.g., by adding them together if they are in the form of logLR) to form a single decision variable, which is then compared to one or more criterion values (the bounds) to arrive at a final choice, as in the DDM (see Fig. 5a). Thus, as long as there is a linear relationship between stimulus strength and logLR, then using an appropriate, multiplicative scale factor to compute the weight of evidence (either scaling the observations or the bound, depending on the particular algorithmic implementation) can support normative decision-making.

These kinds of decision processes have been studied under a variety of conditions that have provided insights into how the brain scales stimulus strength to arrive at a weight of evidence. In the simplest evidence-accumulation paradigms, stimulus strength is held constant across decisions within a block. In this case, the appropriate scale factor can be applied in the same way to each decision within a block, and normative changes in scaling across blocks are equivalent to shifting the decision bound to account for changes in stimulus strength. Results from studies using these paradigms have been mixed: some found that participants vary their bound based on changes in stimulus strength across blocks (Malhotra et al., 2017; Starns and Ratcliff, 2012), whereas others found that participants adopt a fix bound across stimulus strengths (Balci et al., 2011). Interpretation of these studies is complicated by the fact that the participant is typically

assumed to have the goal of maximizing reward rate, which is a complicated function of multiple task parameters, including stimulus strength and timing (Bogacz et al., 2006; Zacksenhouse et al., 2010). Under such conditions, failure to take stimulus strength into account, or failure to do so optimally, could be a result of the particular strategy adopted by the decision-maker rather than a failure to accurately estimate a stimulus-strength-dependent scale factor. For example, several studies found that people deviate from optimal to a greater degree in low-stimulus-strength conditions because they value accuracy and not solely reward rate (Balci et al., 2011; Bohil and Maddox, 2003; Starns and Ratcliff, 2012). Additionally, deviations from optimal bounds have been found to depend on the uncertainty with which task timing is estimated, rather than uncertainty in estimates of stimulus strength (Zacksenhouse et al., 2010), suggesting that people can estimate stimulus strength even if they do not use it as prescribed by reward-rate maximization. By fixing expected logLR (evidence strength) across conditions, we avoided many of these potential confounds and isolated the effects of correlations on behavior.

More commonly, stimulus strength is varied from trial-to-trial in evidence-accumulation tasks. Under these conditions, the standard SPRT (and the DDM as its continuous-time equivalent; Bogacz et al., 2006) are no longer optimal (Deneve, 2012; Drugowitsch et al., 2012; Moran, 2015), because those models typically assume that the same scale factor is used on each trial, but different scale factors are needed to compute the normative weight of evidence (logLR) for different stimulus strengths. These considerations have led some to argue that it is highly unlikely that humans perform optimal computations, particularly under conditions of heterogenous stimulus strengths, because the precise stimulus statistics needed to compute the logLR are assumed to be unavailable or poorly estimated (Ratcliff et al., 2016; Ratcliff and McKoon, 2008). Relatedly, if decision-makers set their bounds according to the true logLR for

each stimulus (equivalent to the goal of maintaining, on average, the same level of accuracy across stimuli), the psychometric function should be flat as a function of stimulus strength, whereas RTs should decrease with increasing stimulus strength. That decisions are both more accurate and faster with increasing stimulus strength argues strongly against the idea that people set bounds based on a fixed expected accuracy or that the accumulated evidence is scaled exactly proportional to the logLR (Hanks et al., 2011).

However, several modeling and empirical studies have shown that it is possible to adjust how decisions are formed about stimuli whose statistics vary from trial to trial, in a manner that is consistent with trying to use optimal forms of the weight of evidence. These adjustments include scaling the decision variable and/or decision bounds within a trial according to online estimates of stimulus strength or some proxy thereof, particularly when the distribution of evidence-strength levels are known (Deneve, 2012; Drugowitsch et al., 2012; Hanks et al., 2011; Huang and Rao, 2013; Malhotra et al., 2018; Moran, 2015). One possible proxy for stimulus strength is the time elapsed within a trial (Drugowitsch et al., 2012; Hanks et al., 2011; Kiani and Shadlen, 2009; Malhotra et al., 2018): the more time has passed in a trial without reaching a decision bound, the more likely that the stimulus is weak. Under certain conditions, human decision-making behavior is consistent with such adjustments (Drugowitsch et al., 2012; Malhotra et al., 2017; Palestro et al., 2018).

Our results imply that outside relatively simple cases involving statistically independent observations, elapsed time cannot serve as a sole proxy for stimulus strength. In particular, pairs of correlated observations can complicate the relationship between stimulus quality and elapsed time: in our task, negative correlations lead to slower decisions than positive correlations if the observations are treated as uncorrelated, when in fact stimulus strength is stronger for negative

correlations than positive correlations. Therefore, in more general settings elapsed time should be combined with other relevant statistics, such as the correlation. In support of this idea, our data are consistent with decisions that used collapsing bounds, which can be a proxy for stimulus strength under an elapsed-time heuristic, but those collapsing bounds alone could not account for the clear behavioral adjustments to trial-to-trial differences in stimulus correlations.

Our results are in stark contrast with the literature on correlations in behavioral economics, which suggests that people fail to use correlations appropriately to inform their decision-making. When combining information from multiple sources (e.g., for financial forecasts: Budescu and Yu, 2007; Enke and Zimmermann, 2017; Hossain and Okui, 2021; Maines, 1996, 1990; or constructing portfolios of correlated assets: Eyster and Weizsacker, 2016; Laudenbach et al., 2022), most participants exhibit "correlation neglect" (i.e., partially or fully failing to account for correlations), which often leads to reduced decision accuracy. Positive correlations have also been proposed to lead to overconfidence, which has been attributed to failing to account for redundancy (Eyster and Rabin, 2010; Glaeser and Sunstein, 2009; Ortoleva and Snowberg, 2015) or to the false assumption that consistency among information sources suggests higher reliability (Kahneman and Tversky, 1973).

These discrepant results are likely a result of the vast differences in task designs between those studies and ours. Those tasks tended to present numerical stimuli representing either small samples of correlated sources or explicitly defined correlation coefficients, often in complicated scenarios. Under such conditions, participants may fail to recognize the correlation and its importance, or they may not be statistically sophisticated enough to adjust for it even if they do (Enke and Zimmermann, 2017; Maines, 1996). In contrast, highly simplified task structures increase the ability to account for correlations (Enke and Zimmermann, 2017). Nevertheless,

even in simplified cases, decisions in descriptive scenarios likely rely on very different cognitive mechanisms than decisions that, like ours, are based directly on relatively low-level sensory stimuli. For example, decisions under risk can vary substantially when based on description versus direct experience (Hertwig and Erev, 2009), and giving passive exposure to samples from distributions underlying two correlated assets has been shown to alleviate correlation neglect in subsequent allocation decisions (Laudenbach et al., 2022).

These differences likely extend to how and where in the brain correlations are represented and used (or not) to inform different kinds of decisions. For certain perceptual decisions, early sensory areas may play critical roles. For example, when combining multiple visual cues to estimate slant, some observers' estimates are consistent with assuming a correlation between cues, which is sensible because the cues derive from the same retinal image and likely overlapping populations of neurons (Oruç et al., 2003; Rosas et al., 2007). The combination of within-modality cues is thought to be encapsulated within the visual system, such that observers have no conscious access to the individual cues (Girshick and Banks, 2009; Hillis et al., 2002). These results suggest that the visual system may have specialized mechanisms for computing correlations among visual stimuli (which may or may not involve the well-studied, but different, phenomena of correlations in the patterns of firing rates of individual neurons; Cohen and Kohn, 2011) that are different than those used to support higher-order cognition.

The impact of correlations on the weight of evidence ultimately depends on the type of correlation and its relationship to other statistical features of the task environment and to intrinsic correlations in the brain (Averbeck and Lee, 2006; Bhardwaj et al., 2015; Hossain and Okui, 2021; Hu et al., 2014; Moreno-Bote et al., 2014). We showed that this impact can be substantial, and that human decision-makers' sensitivity to correlations does not seem to require extensive,

task-specific learning and can be adjusted flexibly from one decision to the next. Further work that pairs careful manipulation of task statistics with neural measurements could provide insight into how the brain tracks stimulus correlations and computes the weight of the evidence to support effective decision-making behaviors under different conditions.

Data and code availability

The datasets generated and analyzed for this article are available at https://osf.io/qygkc/. The analysis code for this article is available at https://github.com/TheGoldLab/Analysis_Tardiff_Kang_Correlated.

Acknowledgements

NT was supported by a T32 training grant from the National Institutes of Health [MH014654]. JG was supported by a CRCNS grant from the National Science Foundation [220727]. JK was funded by the Penn Undergraduate Research Mentorship program (PURM). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank Long Ding for helpful comments on the manuscript.

Author contributions

Nathan Tardiff: Conceptualization, Methodology, Software, Validation, Formal analysis, Data curation, Writing—original draft, Writing—review & editing, Visualization, Supervision, Project administration. Jiwon Kang: Conceptualization, Methodology, Software, Investigation, Data curation, Writing—original draft. Joshua I. Gold: Conceptualization, Methodology, Software, Formal analysis, Resources, Writing—original draft, Writing—review & editing, Visualization, Supervision, Project administration, Funding acquisition.

Methods

Participants

One hundred human participants took part in this online study (42 male, 43 female, 3 other, 12 N/A; median age: 24 yrs, range 18–53, 1 N/A), each of whom provided informed consent via button press. Human protocols were approved and determined to be Exempt by the University of Pennsylvania Internal Review Board (IRB protocol 844474). Participants were recruited using the Prolific platform (https://www.prolific.com/). They were paid a base amount of $9.00 for a projected completion time of 1 hour. They also could receive a bonus of up to $8, depending on task performance (see below).

Behavioral task

The task was developed in PsychoPy (v. 2021.1.4; Peirce et al., 2019), converted to JavaScript (PyschoJS), and run on the online experiment hosting service Pavlovia (https://pavlovia.org/), via functionality integrated into PsychoPy. On each trial, the participant saw a sequence of observations. Each observation consisted of two stars displayed simultaneously. The stars' horizontal positions were generated from a bivariate Gaussian distribution with equal means and variances for each star position and a correlation between star positions that changed from trial-trial-to-trial (the generative distribution), while their vertical position was fixed in the center of the display. The stars were generated by either a "left" source or a "right" source, varied randomly from trial-to-trial. The two sources were equidistant from the vertical midline of the screen, corresponding to equal means of the generative distribution with opposite signs. To prevent stars from being drawn past the edge of the display, their positions were truncated to a maximum value of 0.7, in units of relative window height. For a standard 16:9 monitor at full

screen, this procedure implies that positions could not take on values past 78.8% of the distance from the center of the screen to the edge. Within a trial, new observations were generated from the underlying source distribution every 0.2 sec. Participants were instructed to indicate whether the stars were being generated by the left or the right source once they believed they had accumulated enough noisy information to make an accurate decision.

Each participant was assigned randomly to one of the four correlation-magnitude groups ($|\rho|$ = 0.2, 0.4, 0.6, or 0.8; 25 participants per group) and completed 768 trials, which were divided into 4 blocks of 192 trials, with brief breaks in between. Within each block, there were 12 different stimulus conditions: 2 sources x 2 evidence strengths x 3 correlations. The source (left, right), evidence strength (high, low), and correlation ($-\rho$, 0.0, $+\rho$) were varied pseudo-randomly from trial-to-trial, per participant. Within each block, the trials were divided into 16 sets, with one trial of each condition per set. Each condition was presented in random order within a set, such that all 12 conditions were presented once before the next repetition of a given condition, resulting in 64 total repetitions of each condition across the experiment. Participants received 1 point for each correct choice and −2 points for each incorrect choice (floor of 0 points). The total number of points received by the end of the task was divided by the total possible points (768), and that proportion of $8 was awarded as the bonus.

Prior to completing the main task, each participant completed first a set of training trials, then a staircase procedure to standardize task difficulty across participants. We used a 3-down, 1-up staircase procedure to identify each participant-specific evidence-strength threshold (i.e., by varying the mean of the star-generating distribution while holding its standard deviation at a constant value of 0.1, in units of relative window height) that resulted in a target accuracy of 79.4% in the zero-correlation condition (García-Pérez, 1998). Staircase trials were presented at a

fixed-duration of 1.4 sec, which in pilot data was roughly the mean RT in the free-response paradigm used in the main task, to equate the amount of information provided to each participant and avoid potential individual differences in the speed-accuracy trade-off. The high and low evidence strength conditions used in the main task were then defined as 0.4 and 2.5 times each participant's evidence-strength threshold, respectively.

Because the staircase procedure should standardize accuracy across participants, a performance that is much lower than the target accuracy can be interpreted as a failure of the staircase procedure, a failure of the participant to maintain engagement in the task, or both. Therefore, we kept recruiting participants until we had 25 in each correlation-magnitude group with task performance at 70% or higher. No more than three candidate participants in each group were excluded due to this criterion.

Ideal-observer analysis

Bivariate-Gaussian observations (x_1, x_2) with equal means μ_g, standard deviations σ_g, and correlation ρ are distributed as:

$$p(x_1, x_2|S) = \frac{1}{2\pi\sigma_g^2\sqrt{1-\rho^2}} exp(-\frac{1}{2(1-\rho^2)}[(\frac{x_1-\mu_g}{\sigma_g})^2 + (\frac{x_2-\mu_g}{\sigma_g})^2 - 2\rho\frac{(x_1-\mu_g)(x_2-\mu_g)}{\sigma_g^2}]),$$

where S is the generative source. For the problem of choosing between two such generative sources, S_0 and S_1, the normative weight of evidence can be computed using the log-likelihood ratio, logLR = $\log(\frac{p(x_1,x_2|S_1)}{p(x_1,x_2|S_0)})$. For sources with means μ_0 and μ_1 and equal σ_g and ρ, the logLR reduces to:

$$logLR_{S_1,S_0}(x_1, x_2) = \frac{(\mu_1-\mu_0)}{\sigma_g{}^2(1+\rho)}[(x_1 + x_2) - (\mu_1 + \mu_0)].$$

Our task had equal and opposite generative means, $\mu_0 = -\mu_1 = \mu_g$. Under these conditions, the logLR further simplifies to:

$$logLR_{S_1,S_0}(x_1, x_2) = \frac{2\mu_g}{\sigma_g^2(1+\rho)}(x_1 + x_2).$$

This formulation makes clear that the logLR is a weight of evidence composed of the sum of the observations (for our task corresponding to the horizontal locations of the two stars) times a scale factor that depends on the generative properties of the sources. Because this logLR, which is expressed in terms of bivariate observations (x_1, x_2), depends only on the sum of the observations, it is equivalent to a logLR expressed in terms of univariate observations composed of the sum of each pair (i.e., $logLR_{S_1,S_0}(x_1 + x_2)$; see Fig. 2c). The logLR for a sequence of these (identically distributed, paired) observations is the sum of the logLRs for the individual (paired) observations.

We defined the evidence strength for a given condition as the expected value of the logLR for a single (paired) observation:

$$\text{(expected) evidence strength} := \mathrm{E}\left[logLR_{S_1,S_0}(x_1, x_2)\right] = \frac{4\mu_g^2}{\sigma_g^2(1+\rho)}.$$

Therefore, to equate the evidence strength between two conditions with equal σ_g, but one with correlation = 0 and one with correlation $\rho \neq 0$, we adjusted the generative mean of condition ρ to offset the correlation-dependent scale factor $\frac{1}{1+\rho}$:

$$\mathrm{E}\left[logLR_\rho\right] = \mathrm{E}[logLR_0]$$

$$\frac{4\mu_\rho^2}{\sigma_g^2(1+\rho)} = \frac{4\mu_0^2}{\sigma_g^2}$$

$$\mu_\rho = \mu_0\sqrt{1+\rho}.$$

Drift-diffusion modeling (DDM)

In the DDM, noisy evidence is accumulated into a decision variable until reaching one of the two bounds, representing commitment to one of two choices (e.g., left or right). In general, the average rate of accumulation is governed by the drift rate:

$$drift\ rate = k\frac{\mu_g}{\sigma_g},$$

where μ_g and σ_g are the mean and standard deviation of the generative distribution of the observations (which, as detailed above, for our task can be expressed as the distribution of sums of pairwise observation, $x_1 + x_2$). The drift parameter k captures subjective scaling of the stimulus strength (i.e., the signal-to-noise ratio, $\frac{\mu_g}{\sigma_g}$), which accounts for individual differences in perceptual sensitivity and other factors.

There is an arbitrary degree of freedom in these and related models, which form equivalence classes when the decision variable and decision bound are both scaled in the same way (Green and Swets, 1966; Palmer et al., 2005). Fixing this extra degree of freedom in the DDM is typically accomplished by setting $\sigma_g = 1$, which causes the drift rate and bound height to be scaled implicitly by the standard deviation of the observation distribution (Palmer et al., 2005). This formulation is straightforward when stimulus strength is varied only via changes in signal, μ_g, and not noise (the "diffusion" in "drift-diffusion"), σ_g. However, our task included correlation-dependent changes in both signal and noise. In particular, the correlation scales the standard deviation of the sum distribution by $\sqrt{1+\rho}$. Therefore, we explicitly set the noise across correlation conditions, which is accomplished by scaling the drift rate and bound height

by the relative change in the generative standard deviation induced by the correlation. Thus, for our task the drift rate for correlation ρ is:

$$drift\ rate(\rho) = k_0 \frac{\mu_\rho}{\sqrt{1+\rho}}.$$

where μ_ρ is the correlation-specific generative mean, accounting for changes in signal; k_0 is the subjective drift parameter, which is implicitly scaled by σ_g under the unit variance assumption ($k_0 = \frac{k}{\sigma_g}$); and $\sqrt{1+\rho}$ is the relative change in the generative standard deviation induced by the correlation. By specifying the noise with this latter term, the unit variance assumption of the DDM is maintained and k_0 can accurately reflect subjective scaling of stimulus strength across correlation conditions. Because we manipulated μ_ρ to offset the effect of the correlation on the noise, the drift rate across correlation conditions reduces to:

$$drift\ rate(\rho) = k_0 \mu_0,$$

reflecting the equality of stimulus strength across correlation conditions (see Fig 2c).

The bound height is set by parameter B, which is equal to the amount of evidence that must be accumulated to reach a decision, from a starting point equidistant between the bounds. To maintain equal performance across correlation conditions, the bound height must be scaled to account for the normative weight of evidence. Because the drift rate is related monotonically to the logLR, it is guaranteed that there exists a bound height that satisfies this requirement (Gold and Shadlen, 2001; Green and Swets, 1966). Accordingly, the correlation-specific bound height (B_ρ) for correlation ρ was adjusted relative to the bound height for the zero-correlation condition (B_0) as:

$$B_\rho = B_0 \sqrt{1+\rho},$$

where $\sqrt{1+\rho}$ is the scale factor that ensures that the bound represents the same amount of accumulated evidence across correlation conditions, in units proportional to the true logLR. This scale factor can be derived analytically (Appendix A in the Supplementary Material).

However, decision-makers like our participants do not necessarily know the objective correlation (ρ) but instead must rely on a subjective internal estimate ($\hat{\rho}$). Furthermore, to set the noise under the unit variance assumption of the DDM, like the drift rate, the bound must also be scaled by the relative change in the generative standard deviation across correlation conditions. Therefore, the final correlation-dependent bound height adjustment for correlation ρ was:

$$B_\rho = B_0 \frac{\sqrt{1+\hat{\rho}}}{\sqrt{1+\rho}}.$$

Concretely, $B_0\sqrt{1+\hat{\rho}}$ sets the observer's subjective decision rule, and $\frac{1}{\sqrt{1+\rho}}$ sets the noise. The ratio in this formulation makes clear that the normative correlation-dependent scale factor equalizes the weight of evidence needed to make a decision across correlation conditions (i.e., $B_\rho = B_0$). Note that the normative scale factor could also be implemented in the DDM as a scaling of both the signal and noise, as in the scaling of the observations in the formula for the logLR. We choose to scale the bound instead both for parsimony and to maintain the typical interpretation of the DDM parameters as assigning perceptual factors and decisional factors to the drift rate and bound height, respectively.

All models also included a non-decision time, ndt, that captures the contributions to RT that are not determined by decision formation (e.g., sensory or motor processing). Therefore, RT for a single simulation of the DDM is given by $t_s + ndt$, where t_s is the time at which the bound is reached. Finally, all models included a lapse rate, λ, which mixes the RT distribution determined

by the drift-diffusion process with a uniform distribution in proportion to λ (i.e., $\lambda = 0.01$ computes the predicted RT distribution as a weighted average of 99% the DDM distribution and 1% a uniform distribution).

To empirically validate the ability of the DDM to account for our data (note that the DDM is the continuous-time equivalent of the discrete-time SPRT, which like the generative process in our task is a random-walk process, and one can be used to approximate the other; Edwards, 1965; Smith, 1990; Bogacz et al., 2006), we fit a basic four-parameter DDM (k_0, B_0, *ndt*, λ) to each participant's data from the zero-correlation condition. These fits could qualitatively account for the data but were improved by the addition of a collapsing bound (Figure 6—figure supplement 1). Therefore, all models in the main analyses included a linear collapsing bound, where parameter t_B determined the rate of linear collapse, such that total bound height at time t is $2(B - t_B t)$. Choice commitment occurs when one of the bounds is reached, which happens when $|x(t)| \geq B - t_B t$, where $x(t)$ is the value of the decision variable at time t.

To isolate the mechanisms underlying behavioral adjustments to the correlations, we fit four different DDM variants to each participant's data, jointly to all three correlation conditions. Unless otherwise specified, parameters were shared across conditions. The *base* model was the same as the five-parameter model described above (k_0, B_0, t_B, *ndt*, λ). The *bound* model accounted for correlation-based adjustments using the normative form of the bound scale factor derived above with separate fit subjective correlation parameters for the $-\rho$ and $+\rho$ conditions ($\hat{\rho}_-$ and $\hat{\rho}_+$, respectively) for a total of seven free parameters. The *drift* model instead accounted for correlation-based adjustments by fitting separate drift rates to each correlation condition (k_-, k_0, k_+); it also had seven parameters. Finally, the *bound+drift* model included both separate drift parameters and correlation-based bound adjustments, for a total of nine free parameters.

The DDMs were fit to participant's full empirical RT distributions, using PyDDM (Shinn et al., 2020). Maximum-likelihood optimization was performed using differential evolution (Storn and Price, 1997), a global-optimization algorithm suitable for estimating the parameters of high-dimensional DDMs (Shinn et al., 2020). We also used PyDDM to generate predictions for the expected performance of an observer that uses the normative form of the bound-height adjustment defined above with $\hat{\rho}$ chosen to explore different levels of correlation underestimation, where $|\rho|$ = 0.6, and other model parameters were chosen to approximate the average parameters from participants in the 0.6 correlation-magnitude group.

Data analysis

We conducted statistical analyses in Matlab (Mathworks) and R (R Core Team, 2020). We excluded from analysis trials with RTs <0.3 sec or >15 sec, which are indicative of off-task behavior. This procedure removed 0.8% of the data across participants.

To analyze choice behavior, we fit logistic models to each participant's choices using maximum-likelihood estimation. The basic logistic function was:

$$P(R) = \lambda + \frac{1 - 2\lambda}{1 + e^{-(\beta_0 + \beta_e \, x \, E[logLR])}},$$

where $P(R)$ is the probability that the subject chose the right source, β_e determines the slope of the psychometric function as a function of evidence strength (expected logLR), β_0 is a fixed offset, and λ is a lapse rate that sets the lower and upper asymptotes of the logistic curve. We fit two models per participant to assess whether choices were dependent on the correlations: 1) a *joint* model, in which the three free parameters were shared across the three correlation conditions; and 2) a *separate* model, in which a logistic function was fit separately to each correlation condition (nine free parameters).

To assess whether RTs were affected by the correlations, we fit linear mixed-effects models to median RTs per condition, separately for correct and error trials. The predictors included evidence strength (low, high), correlation condition (−ρ, 0.0, +ρ), and correlation magnitude (0.2, 0.4, 0.6, 0.8), as well as the interaction between evidence strength and correlation magnitude and the interaction between correlation condition and correlation magnitude. Evidence strength and correlation condition were effect coded, and correlation magnitude was *z*-scored and entered as a continuous covariate. The models were fit using lme4 (Bates et al., 2015b). When possible, we fit the maximal model (i.e., random intercepts for subjects and random slopes for all within-subjects variables). In cases where the maximal model failed to converge or yielded singular fits, we iteratively reduced the random-effects structure until convergence (Bates et al., 2015a). Significance was assessed via ANOVA using Kenward-Roger *F*-tests with Satterthwaite degrees of freedom, using the car package (Fox and Weisberg, 2019).

To assess the relationship between the objective correlations and the subjective fit correlations, we fit linear mixed-effects models to the Fisher-*z*-transformed correlations. To quantify the average deviation of the subjective correlation from the objective correlation and the average deviation of the bound scale factors computed with the subjective versus objective correlations, we reversed the signs of the deviations for the negative-correlation conditions so underestimates and overestimates for negative and positive correlations would have the same sign.

Model Comparison

We assessed goodness-of-fit for the logistic and drift-diffusion models using Akaike information criteria (AIC). We also used AIC values in a Bayesian random-effects analysis, which attempts to identify the model among competing alternatives that is most frequent in the population. This analysis produced a protected exceedance probability (PEP) for each model, which is the

probability that the model is the most frequent in the population, above and beyond chance (Rigoux et al., 2014b). We computed PEPs using the VBA toolbox (Daunizeau et al., 2014).

References

Averbeck BB, Lee D. 2006. Effects of Noise Correlations on Information Encoding and Decoding. *J Neurophysiol* **95**:3633–3644. doi:10.1152/jn.00919.2005

Balci F, Simen P, Niyogi R, Saxe A, Hughes JA, Holmes P, Cohen JD. 2011. Acquisition of decision making criteria: reward rate ultimately beats accuracy. *Atten Percept Psychophys* **73**:640–657. doi:10.3758/s13414-010-0049-7

Barnard GA. 1946. Sequential Tests in Industrial Statistics. *Supplement to the Journal of the Royal Statistical Society* **8**:1. doi:10.2307/2983610

Bates D, Kliegl R, Vasishth S, Baayen H. 2015a. Parsimonious Mixed Models. doi:arXiv:1506.04967

Bates D, Mächler M, Bolker B, Walker S. 2015b. Fitting linear mixed-effects models using lme4. *J Stat Softw* **67**:1–48. doi:10.18637/jss.v067.i01

Bhardwaj M, Carroll S, Ma WJ, Josić K. 2015. Visual Decisions in the Presence of Measurement and Stimulus Correlations. *Neural Comput* **27**:2318–2353. doi:10.1162/NECO_a_00778

Bogacz R, Brown E, Moehlis J, Holmes P, Cohen JD. 2006. The physics of optimal decision making: A formal analysis of models of performance in two-alternative forced-choice tasks. *Psychol Rev* **113**:700–765. doi:10.1037/0033-295X.113.4.700

Bohil CJ, Maddox WT. 2003. On the generality of optimal versus objective classifier feedback effects on decision criterion learning in perceptual categorization. *Mem Cognit* **31**:181–198. doi:10.3758/BF03194378

Brown S, Steyvers M, Wagenmakers E-J. 2009. Observing evidence accumulation during multi-alternative decisions. *J Math Psychol* **53**:453–462. doi:10.1016/j.jmp.2009.09.002

Budescu D V., Yu H. 2007. Aggregation of opinions based on correlated cues and advisors. *J Behav Decis Mak* **20**:153–177. doi:10.1002/bdm.547

Cohen MR, Kohn A. 2011. Measuring and interpreting neuronal correlations. *Nat Neurosci* **14**:811–819. doi:10.1038/nn.2842

Daunizeau J, Adam V, Rigoux L. 2014. VBA: A probabilistic treatment of nonlinear models for neurobiological and behavioural data. *PLoS Comput Biol* **10**:e1003441. doi:10.1371/journal.pcbi.1003441

Deneve S. 2012. Making Decisions with Unknown Sensory Reliability. *Front Neurosci* **6**:21473. doi:10.3389/fnins.2012.00075

Denter P, Dumav M, Ginzburg B. 2021. Social Connectivity, Media Bias, and Correlation Neglect. *The Economic Journal* **131**:2033–2057. doi:10.1093/ej/ueaa128

Drugowitsch J, Moreno-Bote R, Churchland AK, Shadlen MN, Pouget A. 2012. The Cost of Accumulating Evidence in Perceptual Decision Making. *The Journal of Neuroscience* **32**:3612–3628. doi:10.1523/JNEUROSCI.4010-11.2012

Dutilh G, Vandekerckhove J, Tuerlinckx F, Wagenmakers E-J. 2009. A diffusion model decomposition of the practice effect. *Psychon Bull Rev* **16**:1026–1036. doi:10.3758/16.6.1026

Edwards W. 1965. Optimal strategies for seeking information: Models for statistics, choice reaction times, and human information processing. *J Math Psychol* **2**:312–329. doi:10.1016/0022-2496(65)90007-6

Enke B, Zimmermann F. 2017. Correlation Neglect in Belief Formation. *Rev Econ Stud* **86**:313–332. doi:10.1093/restud/rdx081

Ernst MO. 2005. A Bayesian View on Multimodal Cue Integration In: Knoblich G, Thornton IM, Grosjean M,, Shiffrar M, editors. Human Body Perception From The Inside Out. New York: Oxford University PressNew York, NY. pp. 105–132. doi:10.1093/oso/9780195178371.003.0006

Eyster E, Rabin M. 2010. Naïve Herding in Rich-Information Settings. *Am Econ J Microecon* **2**:221–243. doi:10.1257/mic.2.4.221

Eyster E, Weizsacker G. 2016. Correlation Neglect in Portfolio Choice: Lab Evidence. *SSRN Electronic Journal*. doi:10.2139/ssrn.2914526

Fox J, Weisberg S. 2019. An R Companion to Applied Regression, 3rd ed. ed. Thousand Oaks, CA: Sage.

García-Pérez MA. 1998. Forced-choice staircases with fixed step sizes: asymptotic and small-sample properties. *Vision Res* **38**:1861–1881. doi:10.1016/S0042-6989(97)00340-4

Geisler WS. 2008. Visual Perception and the Statistical Properties of Natural Scenes. *Annu Rev Psychol* **59**:167–192. doi:10.1146/annurev.psych.58.110405.085632

Girshick AR, Banks MS. 2009. Probabilistic combination of slant information: Weighted averaging and robustness as optimal percepts. *J Vis* **9**:8. doi:10.1167/9.9.8

Glaeser EL, Sunstein CR. 2009. Extremism and Social Learning. *Journal of Legal Analysis* **1**:263–324. doi:10.4159/jla.v1i1.10

Glaze CM, Kable JW, Gold JI. 2015. Normative evidence accumulation in unpredictable environments. *Elife* **4**. doi:10.7554/eLife.08825

Gold JI, Shadlen MN. 2007. The Neural Basis of Decision Making. *Annu Rev Neurosci* **30**:535–574. doi:10.1146/annurev.neuro.29.051605.113038

Gold JI, Shadlen MN. 2001. Neural computations that underlie decisions about sensory stimuli. *Trends Cogn Sci* **5**:10–16. doi:10.1016/S1364-6613(00)01567-9

Good IJ. 1979. Studies in the History of Probability and Statistics. XXXVII A. M. Turing's Statistical Work in World War II. *Biometrika* **66**:393. doi:10.2307/2335677

Green DM, Swets JA. 1966. Signal detection theory and psychophysics. New York: John Wiley.

Hanks TD, Mazurek ME, Kiani R, Hopp E, Shadlen MN. 2011. Elapsed decision time affects the weighting of prior probability in a perceptual decision task. *Journal of Neuroscience* **31**:6339–6352. doi:10.1523/JNEUROSCI.5613-10.2011

Heitz RP. 2014. The speed-accuracy tradeoff: history, physiology, methodology, and behavior. *Front Neurosci* **8**:86875. doi:10.3389/fnins.2014.00150

Hertwig R, Erev I. 2009. The description–experience gap in risky choice. *Trends Cogn Sci* **13**:517–523. doi:10.1016/j.tics.2009.09.004

Hillis JM, Ernst MO, Banks MS, Landy MS. 2002. Combining Sensory Information: Mandatory Fusion Within, but Not Between, Senses. *Science (1979)* **298**:1627–1630. doi:10.1126/science.1075396

Hossain T, Okui R. 2021. Belief Formation Under Signal Correlation. *SSRN Electronic Journal*. doi:10.2139/ssrn.3218152

Hu Y, Zylberberg J, Shea-Brown E. 2014. The Sign Rule and Beyond: Boundary Effects, Flexibility, and Noise Correlations in Neural Population Codes. *PLoS Comput Biol* **10**:e1003469. doi:10.1371/JOURNAL.PCBI.1003469

Huang Y, Rao RPN. 2013. Reward Optimization in the Primate Brain: A Probabilistic Model of Decision Making under Uncertainty. *PLoS One* **8**:e53344. doi:10.1371/journal.pone.0053344

Kahneman D, Tversky A. 1973. On the psychology of prediction. *Psychol Rev* **80**:237–251. doi:10.1037/h0034747

Kiani R, Shadlen MN. 2009. Representation of Confidence Associated with a Decision by Neurons in the Parietal Cortex. *Science (1979)* **324**:759–764. doi:10.1126/science.1169405

Laming DRJ. 1968. Information theory of choice-reaction times. New York: Academic Press.

Laudenbach C, Ungeheuer M, Weber M. 2022. How to Alleviate Correlation Neglect in Investment Decisions. *Manage Sci* **69**:3400–3414. doi:10.1287/mnsc.2022.4535

Levy G, Barreda IM de, Razin R. 2022. Persuasion with Correlation Neglect: A Full Manipulation Result. *Am Econ Rev Insights* **4**:123–138. doi:10.1257/aeri.20210007

Maines LA. 1996. An experimental examination of subjective forecast combination. *Int J Forecast* **12**:223–233. doi:10.1016/0169-2070(95)00623-0

Maines LA. 1990. The Effect of Forecast Redundancy on Judgments of a Consensus Forecast's Expected Accuracy. *Journal of Accounting Research* **28**:29. doi:10.2307/2491245

Malhotra G, Leslie DS, Ludwig CJH, Bogacz R. 2018. Time-varying decision boundaries: insights from optimality analysis. *Psychon Bull Rev* **25**:971–996. doi:10.3758/s13423-017-1340-6

Malhotra G, Leslie DS, Ludwig CJH, Bogacz R. 2017. Overcoming indecision by changing the decision boundary. *J Exp Psychol Gen* **146**:776–805. doi:10.1037/xge0000286

Moran R. 2015. Optimal decision making in heterogeneous and biased environments. *Psychon Bull Rev* **22**:38–53. doi:10.3758/s13423-014-0669-3

Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A. 2014. Information-limiting correlations. *Nat Neurosci* **17**:1410–1417. doi:10.1038/nn.3807

Noppeney U. 2021. Perceptual Inference, Learning, and Attention in a Multisensory World. *Annu Rev Neurosci* **44**:449–473. doi:10.1146/annurev-neuro-100120-085519

Ortoleva P, Snowberg E. 2015. Overconfidence in Political Behavior. *American Economic Review* **105**:504–535. doi:10.1257/aer.20130921

Oruç İ, Maloney LT, Landy MS. 2003. Weighted linear cue combination with possibly correlated error. *Vision Res* **43**:2451–2468. doi:10.1016/S0042-6989(03)00435-8

Palestro JJ, Weichart E, Sederberg PB, Turner BM. 2018. Some task demands induce collapsing bounds: Evidence from a behavioral analysis. *Psychon Bull Rev* **25**:1225–1248. doi:10.3758/s13423-018-1479-9

Palmer J, Huk AC, Shadlen MN. 2005. The effect of stimulus strength on the speed and accuracy of a perceptual decision. *J Vis* **5**:1–1. doi:10.1167/5.5.1

Parise C V. 2016. Crossmodal Correspondences: Standing Issues and Experimental Guidelines. *Multisens Res* **29**:7–28. doi:10.1163/22134808-00002502

Peirce J, Gray JR, Simpson S, MacAskill M, Höchenberger R, Sogo H, Kastman E, Lindeløv JK. 2019. PsychoPy2: Experiments in behavior made easy. *Behav Res Methods* **51**:195–203. doi:10.3758/S13428-018-01193-Y/FIGURES/3

R Core Team. 2020. R: A language and environment for statistical computing.

Ratcliff R, McKoon G. 2008. The diffusion decision model: theory and data for two-choice decision tasks. *Neural Comput* **20**:873–922. doi:10.1162/neco.2008.12-06-420

Ratcliff R, Smith PL, Brown SD, McKoon G. 2016. Diffusion Decision Model: Current Issues and History. *Trends Cogn Sci* **20**:260–281. doi:10.1016/j.tics.2016.01.007

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. 2014a. Bayesian model selection for group studies - Revisited. *Neuroimage* **84**:971–985. doi:10.1016/j.neuroimage.2013.08.065

Rigoux L, Stephan KE, Friston KJ, Daunizeau J. 2014b. Bayesian model selection for group studies - Revisited. *Neuroimage* **84**:971–985. doi:10.1016/j.neuroimage.2013.08.065

Rosas P, Wichmann FA, Wagemans J. 2007. Texture and object motion in slant discrimination: Failure of reliability-based weighting of cues may be evidence for strong fusion. *J Vis* **7**:3. doi:10.1167/7.6.3

Salmon F. 2009. Recipe for Disaster: The Formula That Killed Wall Street. *Wired*.

Shadlen MN, Hanks TD, Churchland AK, Kiani R, Yang T. 2006. The Speed and Accuracy of a Simple Perceptual Decision: A Mathematical Primer In: Doya K, Ishii S, Pouget A, Rao RPN, editors. Bayesian Brain: Probabilistic Approaches to Neural Coding. The MIT Press. pp. 208–237. doi:10.7551/mitpress/9780262042383.003.0010

Shinn M, Lam NH, Murray JD. 2020. A flexible framework for simulating and fitting generalized drift-diffusion models. *Elife* **9**:1–27. doi:10.7554/eLife.56938

Smith PL. 1990. A note on the distribution of response times for a random walk with Gaussian increments. *J Math Psychol* **34**:445–459. doi:10.1016/0022-2496(90)90023-3

Smith PL, Ratcliff R. 2004. Psychology and neurobiology of simple decisions. *Trends Neurosci* **27**:161–168. doi:10.1016/j.tins.2004.01.006

Starns JJ, Ratcliff R. 2012. Age-related differences in diffusion model boundary optimality with both trial-limited and time-limited tasks. *Psychon Bull Rev* **19**:139–145. doi:10.3758/s13423-011-0189-3

Stone M. 1960. Models for choice-reaction time. *Psychometrika* **25**:251–260. doi:10.1007/BF02289729

Storn R, Price K. 1997. Differential Evolution - A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces. *Journal of Global Optimization* **11**:341–359. doi:10.1023/A:1008202821328

Veliz-Cuba A, Kilpatrick ZP, Josić K. 2016. Stochastic Models of Evidence Accumulation in Changing Environments. *SIAM Review* **58**:264–289. doi:10.1137/15M1028443

Wald A. 1947. Sequential analysis. Oxford, England: John Wiley.

Wald A, Wolfowitz J. 1948. Optimum Character of the Sequential Probability Ratio Test. *The Annals of Mathematical Statistics* **19**:326–339. doi:10.1214/aoms/1177730197

Zacksenhouse M, Bogacz R, Holmes P. 2010. Robust versus optimal strategies for two-alternative forced choice tasks. *J Math Psychol* **54**:230–246. doi:10.1016/j.jmp.2009.12.004