Normative evidence weighing 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 weighing 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 Turing 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 has since become ubiquitous in models of decision-making. Examples include sequential-sampling models related to the SPRT like the drift-diffusion model (DDM), which can 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). These models typically assume that the decision is formed by accumulating over time evidence from statistically independent observations until reaching a threshold value, or bound. The magnitude of this bound governs a trade-off between decision speed and accuracy (lower bounds emphasize speed, higher bounds emphasize accuracy; Heitz, 2014). The weight of evidence is computed as a scaled version of each observation (the scaling can be applied to the observations or to the bound, which are mathematically equivalent; Green and Swets, 1966) to form the logLR.

However, it is often unclear if and when decision-makers use 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). A major complication is that the logLR can be difficult to compute, because it depends on detailed knowledge of the statistical properties of the observations. One such property is the signal-to-noise ratio (SNR) of the observations. When the SNR is not directly accessible to the decision-maker (e.g., when signal strength is varied randomly from trial to trial, as is common for many laboratory tasks), it can be approximated using surrogates like elapsed decision time to help calibrate the weight of evidence (Drugowitsch et al., 2012; Hanks et al., 2011).

Another statistical property that has received less attention is non-independence of the observations, which can have substantial effects on how those observations should be weighed to form effective decisions (Fig. 1). If not accounted for appropriately, these effects can lead to over- or under-estimates of the weight of available evidence and 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 and others 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

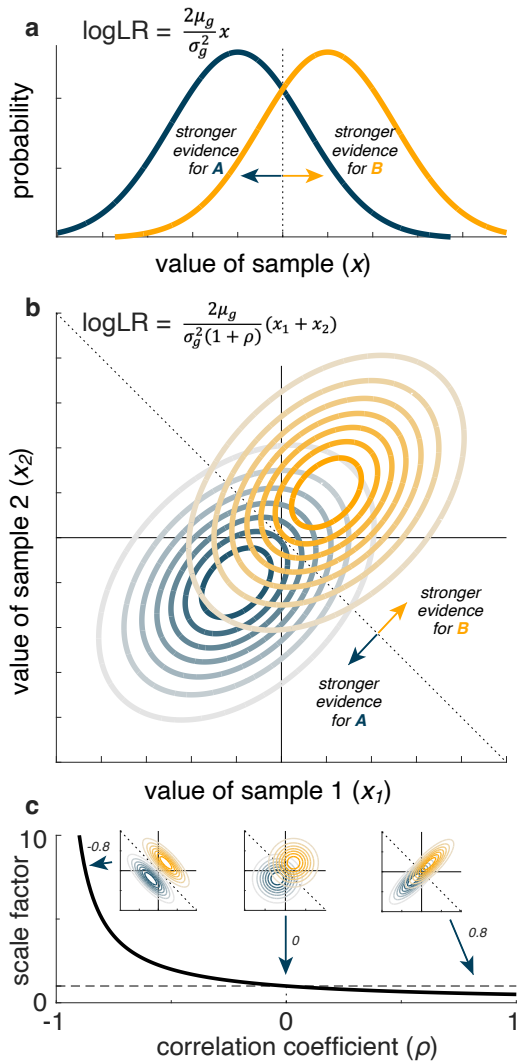

Figure 1. Illustration of how pairwise correlations can affect the weight of evidence (logLR) for the generative source of an observation. **a** Computing the logLR when the observation (x) is a single sample from one of two one-dimensional Gaussian distributions (labeled ***A*** and ***B***), with means $\pm\mu_g$ 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, correlation-dependent scaling of the weight of evidence ($\frac{1}{1+\rho}$ term in **b**) of the observation plotted as a function of the correlation. 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***.

accumulate evidence that is based on pairs of correlated observations (Fig. 2; we do not consider other forms of correlation, such as over time). Under these conditions, normative decisions use an accumulate-to-bound process that has been scaled appropriately to produce a correlation-dependent weight of evidence. 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.

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 sign and magnitude of the correlation (Fig. 1b). For 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, while 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: ρ_{-}, 0, or ρ_{+}. 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 objective evidence strength) was constant across correlation conditions (Fig. 2b,c). For example, because negative correlations increase logLR for the same generative mean and standard deviation (Fig. 1c), we used smaller differences in means for the negative-correlation conditions than for the zero-correlation conditions. As a result, we expected participants who made 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 expected participants who ignored the correlations to under-weigh the evidence provided by negative-correlation pairs and over-weigh the evidence provided by positive-correlation pairs, which would affect RTs and/or choices. We further expected strategies between these two extremes to have more mixed effects on behavior, as we detail below.

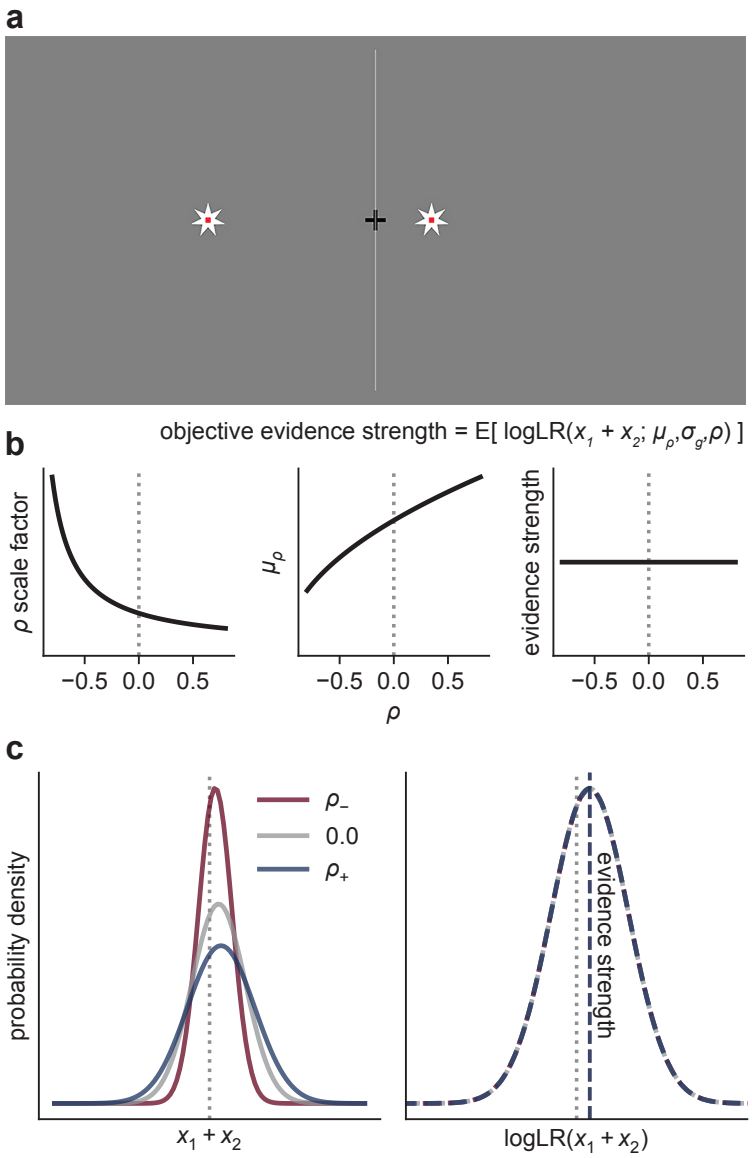

Figure 2. Task. **a** Human observers viewed pairs of stars (updated every 0.2 sec) and were asked to decide whether the stars were generated by a source on the left or right side of the screen. An example star pair is shown. The horizontal position of each star pair was drawn from a bivariate Gaussian 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 (objective 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., $\mu_\rho = \mu_0\sqrt{1+\rho}$), 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).

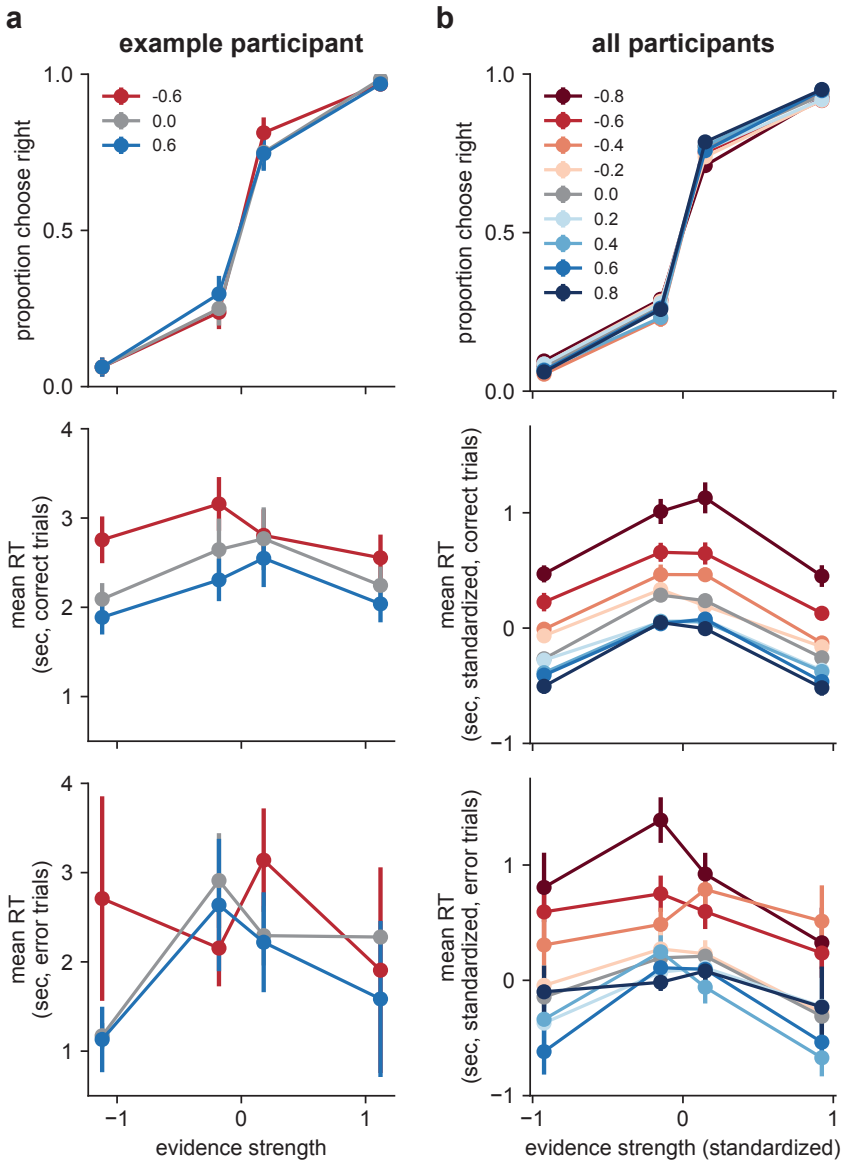

Figure 3. Effects of correlations on choice and RT. **a** Data from an example participant from the 0.6 correlation-magnitude group. Top: choices plotted as a function of evidence strength (abscissa) and correlation condition (see legend). Middle, Bottom: Mean RTs for correct and error trials, respectively. 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 objective 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 on 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 also 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 on RT 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 and RT 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. However, the RT data imply that this correlation dependence was not optimal (i.e., was not equivalent to using a weight of evidence based on the true logLR), because if it were, the RTs would not vary by 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, assuming that, like for the DDM and SPRT, the decision was formed by accumulating evidence over time until reaching one of two fixed bounds. This process governs both the choice (which bound is reached first) and RT (when the bound is reached). We considered evidence with weights determined by three different scale factors (i.e., the value by which to multiply each star position or, equivalently, divide each decision bound to govern the weight of evidence for a given set of task conditions): 1) "unscaled" evidence was taken directly as each star position; 2) "naïve" evidence was scaled by the generative mean, μ_g, of a pair of samples, which produces a weight of evidence equivalent to a mis-specified logLR that ignores the correlations; and 3) "true" evidence was scaled by $\frac{\mu_g}{1+\rho}$, 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 (objective 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 unscaled or naïve logLR would have different effects for positive versus negative correlations. Ignoring positive correlations is equivalent to ignoring redundancies in the observations, which would lead to over-weighing the evidence and thus reaching the bound more quickly, corresponding to shorter RTs. Ignoring negative correlations is equivalent to ignoring synergies in the observations, which would lead to under-weighing the evidence and thus reaching the bound less quickly, corresponding to longer RTs (Fig. 1c).

The 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 (either with or without additional scaling). They also were not completely consistent with a decision process that used a fixed bound on the true logLR, because of the dependency of RT on the 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.

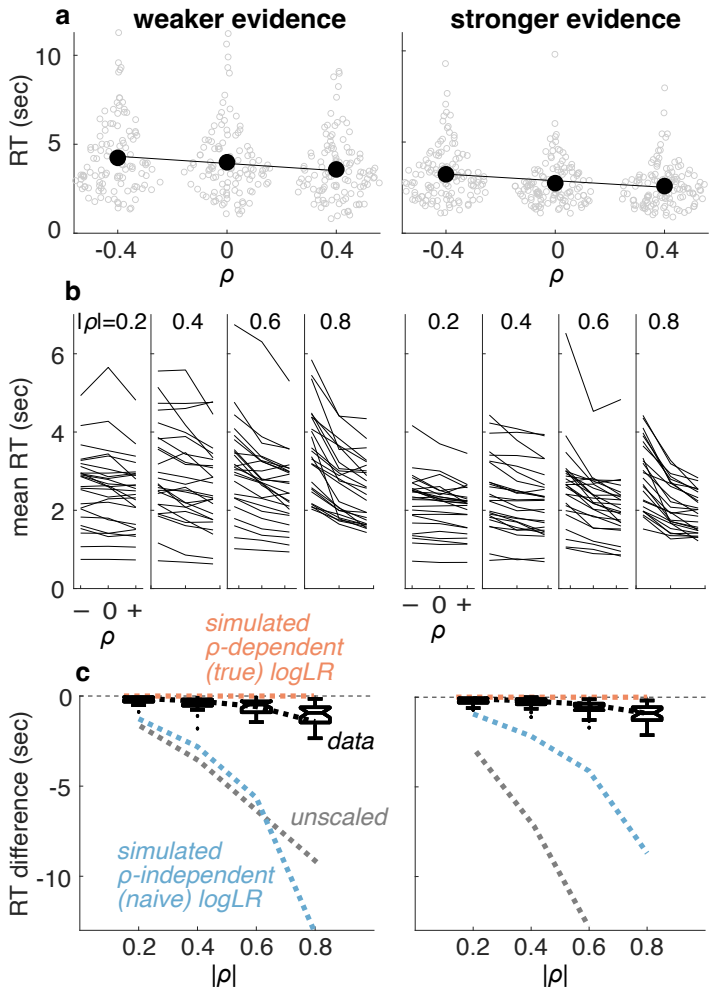

Figure 4. RTs were consistent with a bound on (approximate) logLR. **a** RTs measured from an example participant for the weaker (left) and stronger (right) evidence conditions. Unfilled points are data from individual trials. Filled points are means, lines are linear fits to those means. **b** Summary 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 differences in mean RT between the positive- versus negative-correlation condition 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 unscaled, correlation-independent (naïve), or 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 $\frac{\mu_g}{1+0.9\rho}$; black dashed lines).

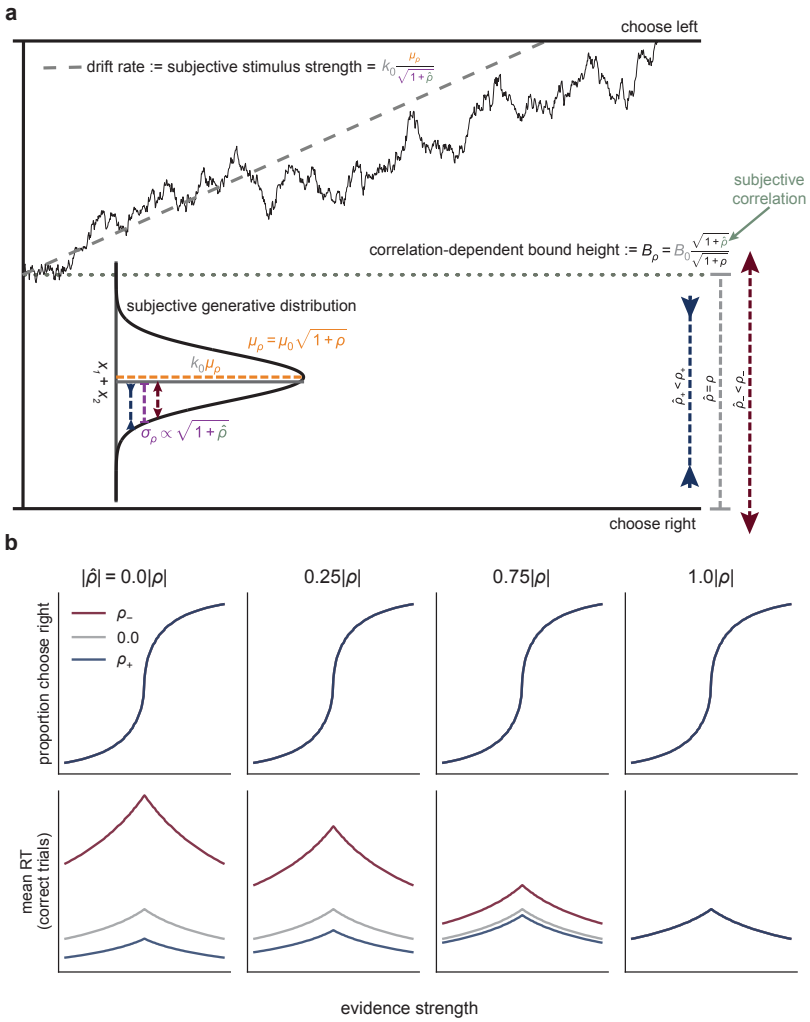

Figure 5. A drift-diffusion model (DDM) captures normative evidence weighing via bound-height adjustments. **a** In the DDM, sensory observations are modeled as samples from a Gaussian distribution (in the continuum limit). Evidence is accumulated over time as the decision variable until it reaches one 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). We designed the task such that this scaling effect on drift rate was countered exactly by correlation-dependent changes in the mean of the generative distribution. Normative evidence weighing corresponds to correlation-dependent adjustments of the bound height that are functionally equivalent to scaling the observations to compute the true logLR. **b** Predictions from a DDM that implements normative bound-height adjustments but allows for subjective misestimates of the correlation. Colors correspond to three simulated correlation conditions (see legend and headings). Other parameters were chosen to approximate the fits to human data. Each column depicts predictions based on the same form of correlation-dependent bound scaling 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 ($|\hat{\rho}| < |\rho|$, first three columns) leads to RT differences across the conditions, where the magnitude of the differences depends on the degree of underestimation (note that misestimating ρ does not cause correlation-dependent changes in choice patterns predicted by the DDM, because choice depends on the product of the drift and bound, in which the subjective terms cancel). Only $\hat{\rho} = \rho$ (rightmost column) produces exactly equal predicted RTs across conditions.

Correlation-dependent adjustments in a drift-diffusion model

To better understand how the participants formed correlation-dependent decisions, we developed variants 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. 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 observation 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). We used a single fit parameter representing symmetric bounds (*bound height* = *B*). Following previous approaches, we also included an additional parameter (t_B) to govern the rate of a linear "collapse" of the bounds over time, which can serve to calibrate the weight of evidence when the objective evidence strength is varied from trial to trial (Drugowitsch et al., 2012; Hanks et al., 2011).

For our task, changes in ρ affected the standard deviation of the distribution of sum-of-pairs observations, σ_ρ (i.e., $\sigma_\rho \propto \sqrt{1+\rho}$). The DDM typically assumes that this observation variability is captured by a noise term (the "diffusion" part of "drift-diffusion") that is used to normalize both the drift rate and bound height (Palmer et al., 2005). Conventionally, this scaling is implicit: by assuming that the diffusion is constant across conditions, this scaling is simply subsumed into the drift-rate and bound-height parameters, and the diffusion parameter is set to one. In contrast, our models accounted for the correlation-dependence of this diffusion term

explicitly so that we could fit a single model to the correlation-dependent data from each participant. Critically, this correlation-dependent diffusion was a component of the internal decision process and not the external task features. Thus, we formulated it using a pair of free parameters that represented subjective estimates of the negative ($\hat{\rho}_{SD-}$) and positive ($\hat{\rho}_{SD+}$) correlations affecting the standard deviation of the internal noise distribution, as follows.

For the drift rate, dividing by the component of the noise distribution (σ_ρ) that depended on the appropriate $\hat{\rho}_{SD}$ yields:

$$drift\ rate(\rho) = k\frac{\mu_\rho}{\sigma_\rho} = k_0\mu_0\frac{\sqrt{1+\rho}}{\sqrt{1+\hat{\rho}_{SD}}},$$

where k_0 is the drift-rate parameter for the zero-correlation condition. Here we also express μ_ρ (the "drift" part of "drift-diffusion") as the ρ-dependent generative mean that for our task was set to $\mu_0\sqrt{1+\rho}$ (see Fig. 2). This formulation allows for both linear and non-linear relationships between objective stimulus properties (governed by task parameters μ_0 and ρ) and subjective stimulus strength (governed by fit parameters k_0 and $\hat{\rho}_{SD}$,). As such, it leaves open the possibility of a variety of participant-specific suboptimalities in the decision process, which are often found for decisions requiring the simultaneous accumulation of evidence from multiple sources (Kang et al., 2021; Luyckx et al., 2020; Usher et al., 2019; Wyart et al., 2015; cf. Rangelov et al., 2024).

For the bound height, dividing by the component of the noise distribution (σ_ρ) that depended on the subjective correlation yielded a similar dependence on $\sqrt{1+\hat{\rho}_{SD}}$. However, the correlation dependence of the bound was further complicated by an additional scaling to convert the sum-of-pairs observations to logLR, which also depended on ρ (using as a scale factor the inverse of

$2\mu_g/\sigma_g(1+\rho)$; see Fig. 1b). Because this conversion also applied to internal quantities, again we used a subjective estimate of the correlation to scale the bound, $\hat{\rho}_B$ (which we assumed could differ from $\hat{\rho}_{SD}$, as detailed below). Specifically, we scaled the bound as (see Methods for details):

$$B_\rho = B_0 \frac{\frac{(1+\hat{\rho}_B)}{\sqrt{1+\rho}}}{\sqrt{1+\hat{\rho}_{SD}}},$$

where B_0 is the bound height for the zero-correlation condition (which subsumes the components of the logLR scale factor that do not depend on ρ).

Consider two cases that illustrate how correlation misestimates could affect the decision process. First, an observer's internal encoding of the observation distribution could underestimate the correlation magnitude (i.e., $|\hat{\rho}_{SD}| < |\rho|$), but they then use this underestimate normatively to compute the weight of evidence. For this "suboptimal encoding" case, $\hat{\rho}_B = \hat{\rho}_{SD} = \hat{\rho}$ and

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

This formulation leads to the following predictions (Fig. 5b):

- If $\hat{\rho} = \rho$, then $drift\ rate(\rho) = drift\ rate(0)$ and $B_\rho = B_0$: when correlations are perceived and 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 $drift\ rate(\rho) > drift\ rate(0)$ and $B_\rho < B_0$: when positive correlations are underestimated, the subjective evidence strength is greater than the objective evidence strength, corresponding to a higher drift rate and lower bound than when the correlations are estimated correctly. These effects cancel for the psychometric

function (which depend only on the product of drift rate and bound height), leaving accuracy unchanged, but tend to produce faster RTs.

- If $\hat{\rho} > \rho$ and $\rho < 0$, then $drift\ rate(\rho) < drift\ rate(0)$ and $B_\rho > B_0$: when negative correlations are underestimated, the subjective evidence strength is less than the objective evidence strength, corresponding to a lower drift rate and higher bound than when the correlations are estimated correctly. These effects cancel for the psychometric function, leaving accuracy unchanged, but tend to produce slower RTs.

Second, an observer could encode the true ρ but underestimate this correlation when computing the weight of evidence ($\hat{\rho}_B$). For this "suboptimal weighing" case, $\hat{\rho}_{SD} = \rho$ and

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

Unlike suboptimal encoding, suboptimal weighing affects the speed-accuracy trade-off and thus predicts changes in both accuracy and RT (compare Fig. 5 to Fig. 5—figure supplement 1).

Scaling the bound in these formulations follows conventions of the DDM, as detailed above, to facilitate interpretation of the parameters. These formulations also raise an apparent contradiction: the "predefined" bound is scaled by subjective estimates of the correlation, but the correlation was randomized from trial to trial and thus could not be known in advance. However, scaling the bound in these ways is mathematically equivalent to using a fixed bound on each trial and scaling the observations to approximate logLR (see Methods). This equivalence implies that in the brain, effectively scaling a "predefined" bound could occur when assigning a weight of evidence to the observations as they are presented.

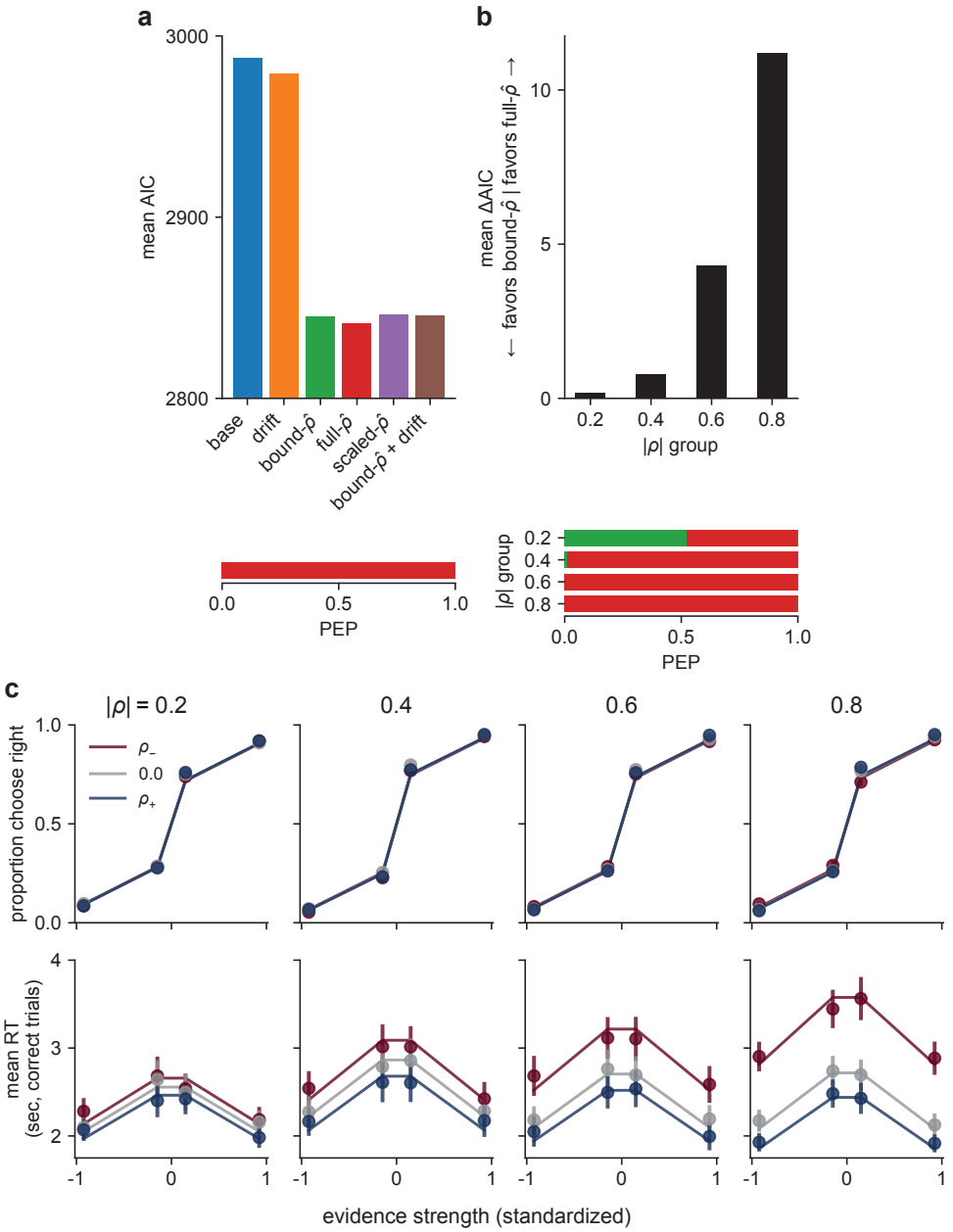

Figure 6. A DDM accounts for human behavior. **a** Model comparison: mean AIC (top) and protected exceedance probability (PEP; bottom), across all participants, for six different models, as labeled (see text for details). **b** Model comparison within each correlation-magnitude group, showing the difference in AIC between the *full-$\hat{\rho}$* and *bound-$\hat{\rho}$* models (top) and PEP over all models (bottom). Bar colors in the PEP plots correspond to the model colors in the top panel of **a**. **c** Predictions from the *full-$\hat{\rho}$* 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.

Human performance is consistent with correlation-dependent bound adjustments

Qualitatively, the participants' patterns of correlation-dependent RTs but correlation-independent choices were consistent with slight underestimates of the correlation (see Fig. 4) resulting from suboptimal encoding (compare Figs. 3b and 5b), not suboptimal weighing (compare Figs. 3b and 5—figure supplement 1b). To examine these effects more quantitatively, we fit six DDMs to each participant's data (see Methods for details). Prior to fitting these models, we confirmed that the DDM with a linear collapsing bound could generally account for choice and RT data from the zero-correlation condition (Fig. 6—figure supplement 1). All models included five basic free parameters: one for the drift rate (k_0), two for the bound (B_0, t_B), one accounting for sensory and motor ("non-decision") processing times (*ndt*), and one for lapses (λ). The models differed in whether and how they accounted for changes in the correlation.

The six models we used were: 1) a *base* model, which included no adjustments to evidence weighing based on the correlation; 2) a *drift* model, which included unconstrained, correlation-dependent adjustments in the drift rate, but not the bound (by fitting separate drift parameters for the negative, zero, and positive correlation conditions: k_-, k_0, k_+, respectively); 3) a *bound-*$\hat{\rho}$ model, which included correlation-dependent adjustments to the bound based on subjective estimates of the correlation (i.e., suboptimality in evidence weighing, with extra free parameters $\hat{\rho}_{B-}$ and $\hat{\rho}_{B+}$); 4) a *full-*$\hat{\rho}$ model, which included normative evidence weighing that depended on subjective estimates of the correlation (i.e., suboptimality in encoding, where $\hat{\rho}_B = \hat{\rho}_{SD} = \hat{\rho}$, with extra free parameters $\hat{\rho}_-$ and $\hat{\rho}_+$); 5) a *scaled-*$\hat{\rho}$ model, which permitted suboptimality in both encoding and weighing (with extra free parameters $\hat{\rho}_{SD-}$, $\hat{\rho}_{SD+}$, $\hat{\rho}_{B-}$, and $\hat{\rho}_{B+}$); and 6) a *bound-*$\hat{\rho}$ + *drift* model, which was the same as the *bound-*$\hat{\rho}$ model but also included unconstrained, correlation-dependent adjustments in the drift rate (with extra free parameters k_-,

k_+, $\hat{\rho}_{B-}$, and $\hat{\rho}_{B+}$). For all models that did not explicitly fit $\hat{\rho}_{SD}$, we set $\hat{\rho}_{SD} = \rho$, which corresponds to the standard DDM assumption that subjective stimulus strength is related linearly to objective stimulus strength (Palmer et al., 2005).

Supporting our qualitative observations, these model fits indicated that the *full-*$\hat{\rho}$ model best captured behavioral performance across correlation conditions (Fig. 6a; see Appendix 1 for average best-fitting parameters for all models). Thus, the participants' behavior was consistent with an accumulate-to-bound process that was scaled on each trial to form decisions based on a normative, correlation-dependent weighing of evidence that was encoded according to a slight misestimate of the correlation. This result was true for all correlation magnitudes individually, except for the 0.2 group, for which the fit statistics were equivocal between the *full-*$\hat{\rho}$ and *bound-*$\hat{\rho}$ models (Fig. 6b; predictions of the two models are similar for low correlations and become more distinguishable as correlations increase, compare Fig. 6c to Fig. 6—figure supplement 2). The models that did not include correlation-dependent bound adjustments (the *base* and *drift* models) provided poor fits to the data (Fig. 6a, Fig. 6—figure supplement 3), underscoring the importance of changes in evidence weighing rather than simply changes in subjective stimulus strength for capturing the participants' performance.

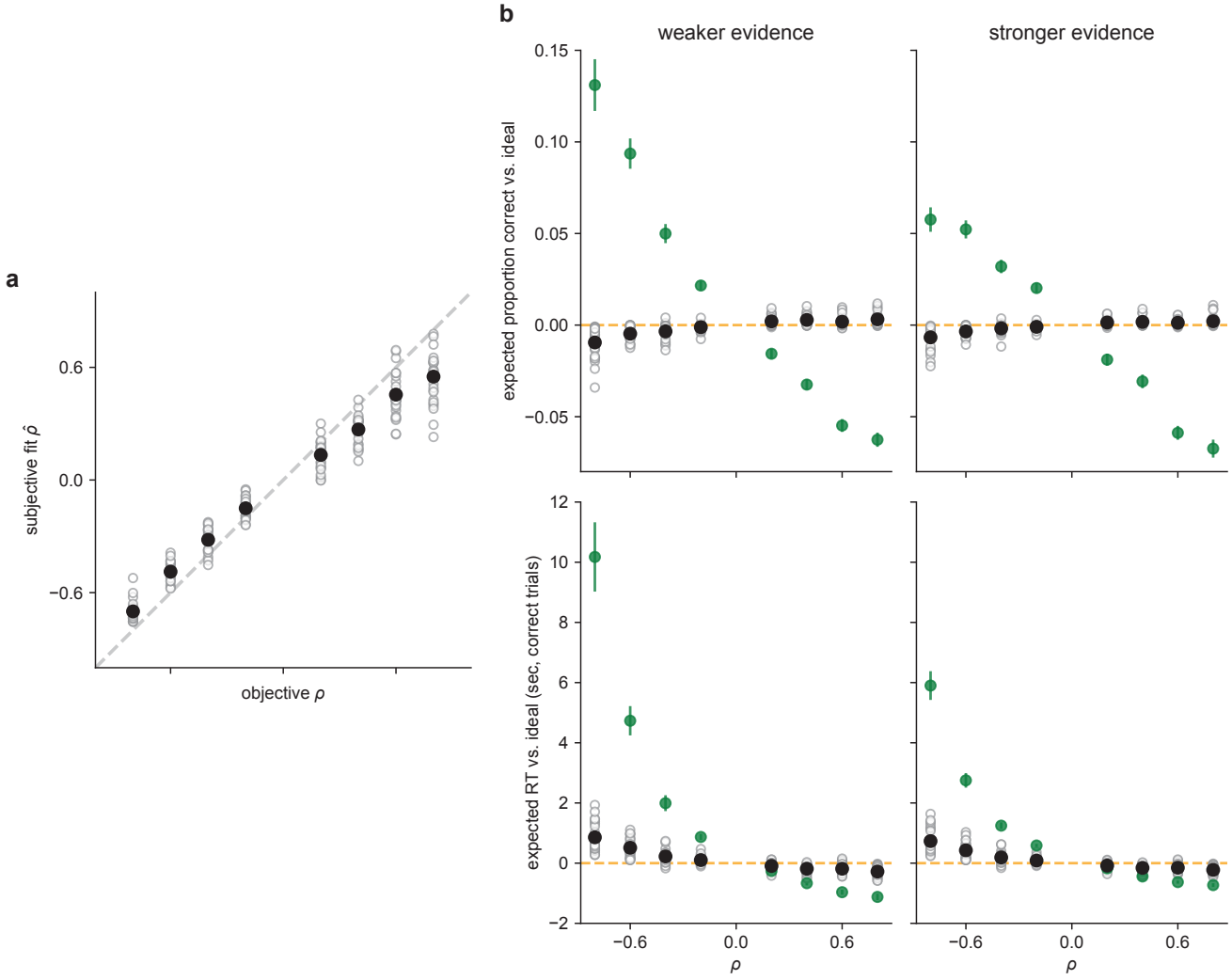

Figure 7. Participants used near-optimal correlation estimates, with slight biases away from extreme values. **a** The subjective fit correlation ($\hat{\rho}$) from the DDM as a function of the objective correlation (ρ). Open circles are the fits from individual participants. Closed circles are the means per correlation condition (means 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** Expected accuracy (top) and RT (bottom) from the DDM fits to data from each participant (open circles) for weak (left) or strong (right) evidence, relative to an ideal observer (orange line) simulated with the per-participant DDMs using the true, objective correlation ($\hat{\rho} = \rho$). Black circles are mean values across participants. Green circles are data simulated with the per-participant DDMs for a naïve observer that used the same subjective fit correlation but did not use the correlation to normatively adjust the bound (i.e., $\hat{\rho}_{SD} = \hat{\rho}$, $\hat{\rho}_B = 0$). Note that predicted performance that appears slightly better than ideal for positive correlations is an artifact of varying the correlation independently of the drift rate in our simulations (for illustrative purposes), when they would both presumably be affected by suboptimalities in encoding.

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

The *full-*$\hat{\rho}$ model, which implements normative evidence weighing given a subjective estimate of the correlation, best accounted for behavioral differences across correlation conditions. We leveraged the fits to relate subjective correlation estimates to performance.

There was a strong relationship between the objective and subjective correlations used by each participant (Fig. 7a; B = 0.71, $t(99) = 47.92$, $p < 0.001$, Fisher z-transformed scale). This result confirms that the participants were sensitive to the correlations and used them to adjust their decision process. 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, $t(199) = -13.29$, $p < 0.001$). This result is consistent with our hypothesis that their deviations from optimal behavior (i.e., unequal RTs across correlation conditions) resulted from subjective correlation estimates that were systematically lower than the true generative correlation. The fit (subjective) correlations also tended to be more variable for positive versus negative correlations (Fig 7a; mean value of the standard deviation of Fisher z-transformed estimates = 0.20 for positive correlation conditions, 0.09 for negative correlation conditions), possibly reflecting the weaker consequences of misestimating positive versus negative correlations on performance (Fig. 7b). Despite these suboptimalities, overall performance was much closer to ideal (i.e., perfect encoding and use of the correlation) than if the participants ignored the correlations when computing the weight of evidence (Fig. 7b).

We do not know how the participants formed their subjective estimates of the correlation or why in many cases these estimates tended to be biased toward zero. One possibility is that they formed an independent, empirical estimate on each trial, based on the observed samples.

Consistent with this idea, these empirical correlations tended to be biased toward zero when based on a very limited number of samples, as expected (Fig. 7—figure supplement 1; Zimmerman et al., 2003). These estimates might also have been biased toward their across-trial mean of zero, which could have served as a prior given that the estimates were based on limited data and thus were highly uncertain.

However, two lines of evidence argue against the idea that the participants computed an empirical estimate of the correlation on each trial. First, this idea predicts that the largest biases toward zero should occur when the estimates are based on the fewest number of samples. Contrary to this prediction, we found no reliable evidence that participants with the shortest average RTs (and thus who observed the fewest samples) tended to have subjective estimates that were most strongly biased towards zero (Fig. 7—figure supplement 2). Second, there was no systematic relationship between the best-fitting $\hat{\rho}$ and the mean empirical ρ computed per participant for each correlation condition (Fig. 7—figure supplement 3). These results suggest that the participants did not compute the correlation explicitly on each trial and thus instead might have used pattern matching (e.g., quickly assessing if the first few samples seem consistent with other observations from the negative-, zero-, or positive-correlation stimuli they already observed) or other heuristics to make correlation-dependent adjustments to the decision process.

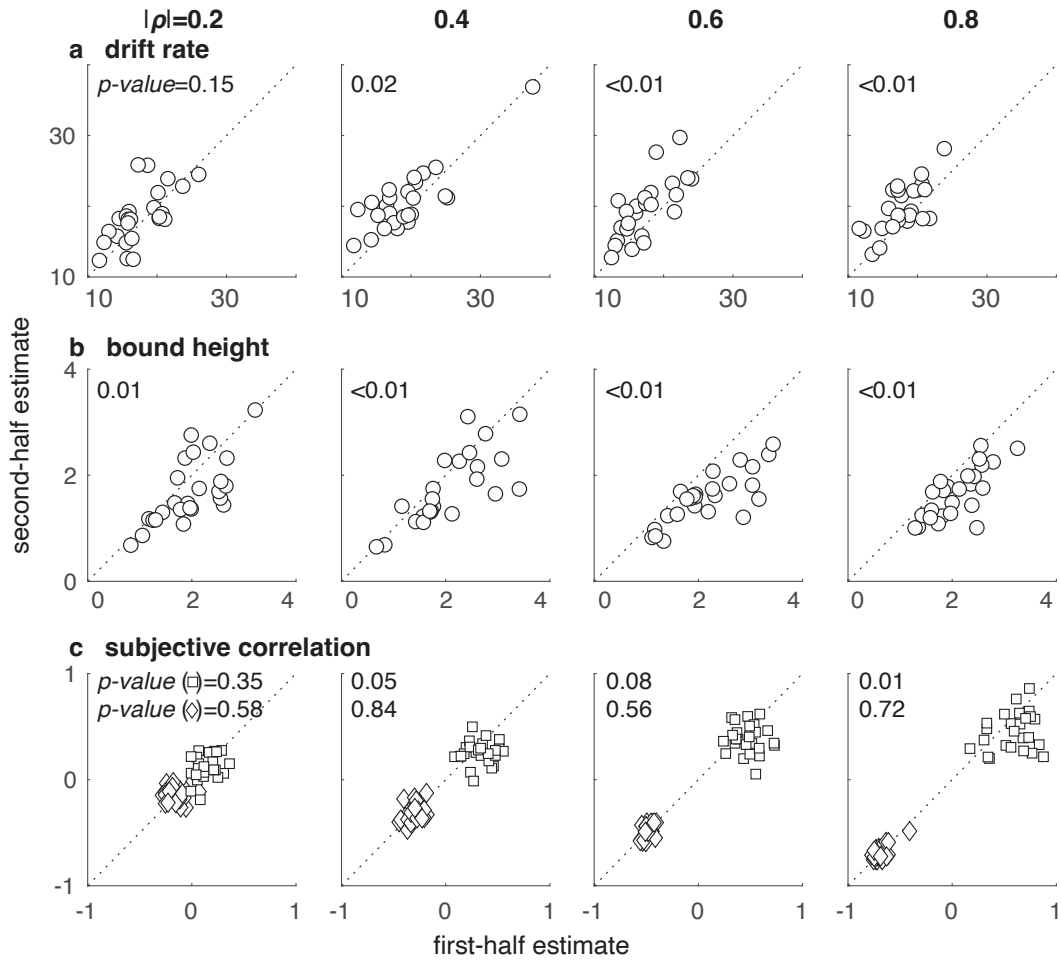

Figure 8. Participants used stable estimates of the correlations, even as they adjusted other components of the decision process over the course of a session. Each panel shows a scatterplot of DDM parameters estimated using the first (abscissa) versus second (ordinate) half of trials from a given participant. Points are data from individual participants. Columns are correlation-magnitude group, and rows are: (**a**) drift rate, k_0; (**b**) bound height, B_0; (**c**) 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 (only effects labeled as $p < 0.01$ survived Bonferroni correction).

These correlation-dependent differences in the decision process also 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.

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 weighing 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 changes in decision speed and/or accuracy 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 and combine multiple pieces of evidence following normative principles has come from two influential 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 the observed stimulus strength to obtain the weight of evidence associated with each observation. 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 subjective 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 observations to arrive at a weight of evidence. In the simplest evidence-accumulation paradigms, objective stimulus strength is held constant across decisions within a block. In this case, the same scale factor can be applied 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 objective stimulus strength. Results from studies using these paradigms have been mixed, including bounds that do (Malhotra et al., 2017; Starns and Ratcliff, 2012) or do not (Balci et al., 2011) vary across blocks. Interpretation of these studies is complicated by the fact that the participants are 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 the appropriate 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 can 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 (objective evidence strength) across conditions, we avoided many of these potential confounds and isolated the effects of correlations on behavior.

Objective stimulus strength is also often 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). These 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 (or near-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 stimulus-strength levels is 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 evidence is weak. Under certain conditions, human decision-making behavior is consistent with making such adjustments, for example by using decision bounds that "collapse" over time (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, correlations between pairs of observations can complicate the relationship between the strength of evidence provided by individual observations and elapsed time. For example, for our task negative correlations lead to slower decisions than positive correlations if the observations are treated as uncorrelated, when in fact the objective evidence 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, to determine an appropriate weight of evidence. In support of this idea, our data are consistent with decisions that used collapsing bounds, which helped adjust the decision process when the objective evidence strength changed within each correlation condition (i.e., the low and high values of evidence strength that corresponded to substantial modulations of the psychometric and chronometric functions). However, collapsing bounds alone, without additional correlation-dependent

computations, could not account for our behavioral results, which by design used values of objective evidence strength that did not change across correlation conditions.

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. For example, 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 simple 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 can 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. Given our finding that suboptimal evidence weighing reflected subjective misestimates of the correlation at the level of encoding, these putative encoding mechanisms appear to be tightly coupled to those that convert observations into a weight of evidence.

The impact of correlations on the weight of evidence depends ultimately 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 for a particular form of correlation, 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, chosen randomly with equal probability on each 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 each, with brief breaks between blocks. Within each block, there were 12 different stimulus conditions varied pseudo-randomly from trial-to-trial, per participant: 2 sources (left, right) x 2 evidence strengths (low, high) x 3 correlations (ρ_-, 0.0, ρ_+). 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 (their total points could never go below zero). 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, 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 based on 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 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) multiplied by 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., $logLR_{S_1,S_0}(x_1 + x_2)$, where $(x_1 + x_2) \sim N(2\mu_g, \sqrt{2\sigma_g^2(1+\rho)})$; 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 objective evidence strength for a given condition as the expected value of the logLR for a single (paired) observation:

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

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 objective 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 via changes in only signal, μ_g, and not noise, σ_g. In that case, the scaling is constant across signal strengths and thus typically simply assumed to be captured by the drift and bound parameters.

However, our task included correlation-dependent effects on both signal and noise. To specify the stimulus strength in the model when noise varies, both the drift-rate and bound-height terms

must be scaled by the noise, which we implemented by scaling both terms by the correlation-dependent component of the noise (i.e., the standard deviation, or SD, of the generative process), $\sqrt{1+\hat{\rho}_{SD}}$. Here $\hat{\rho}_{SD}$ is a fit parameter corresponding to the subjective correlation experienced by the observer, which can deviate from the objective correlation. Unlike subjective scaling of signal strength, subjective deviations in the noise cannot be assumed to be absorbed by the drift-rate (i.e., *k)* or bound-height parameters because of the nonlinear effects of the noise. Note also that this formulation assumes that σ_g, the generative standard deviation in the zero-correlation condition, is absorbed into the drift-rate and bound-height parameters, which serve as an important baseline for the correlation-based adjustments. Therefore, the drift rate in our model was formulated as:

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

where $k_0$is a fit parameter corresponding to the drift rate in the zero-correlation condition, μ_0 is the true (objective) generative mean in the zero-correlation condition, and $\sqrt{1+\rho}$ is the scale factor we used to scale the generative mean to equate objective evidence strength across correlation conditions.

To implement normative evidence weighing in the DDM, we started by assuming that the evidence distribution was based on the logLR estimated for each observed sample pair (Fig. 2c). Converting the observation distribution to the evidence distribution entails scaling the observations by the signal and noise characteristics of the observation distribution, as derived above. In the DDM, this scaling of the evidence is equivalent to: 1) assuming that the decision variable accumulates momentary evidence of the form $(x_1 + x_2)$, and then 2) dividing the bound height by the appropriate scale factor. An alternative approach would be to scale both the signal

and noise components of the DDM by the scale factor. However, scaling the bound is simpler and maintains the conventional interpretation of the DDM parameters in which the bound reflects the decision-related components of the evidence accumulation process, and the drift rate represents sensory-related components. Accordingly, we scaled the bound height as:

$$B_\rho = B\,\frac{\sigma_g^2(1+\hat{\rho}_B)}{2\mu_0\sqrt{1+\rho}}\frac{1}{\sqrt{1+\hat{\rho}_{SD}}} = B_0\,\frac{(1+\hat{\rho}_B)}{\sqrt{1+\rho}}\frac{1}{\sqrt{1+\hat{\rho}_{SD}}}.$$

where the first scale factor is the reciprocal of the normative evidence-weighing term that converts observations to logLR, as derived above; the second scale factor is the scaling by the correlation-dependent component of the noise; and B_0 is a fit parameter corresponding to the bound height in the zero-correlation condition (that absorbs the constant terms $\sigma_g^2/2\mu_0$). We include a subjective fit correlation $\hat{\rho}_B$ in the correlation-dependent evidence-weighing term that was different than the one we used in the noise term ($\hat{\rho}_{SD}$), because in principle an observer may fail to appropriately scale the evidence by the observed correlation (e.g., a naïve observer who ignores the correlation will have $\hat{\rho}_B = 0$), even though the correlation can still influence performance through its effect on the noise.

We focus on two special cases of this model. When $\hat{\rho}_B = \hat{\rho}_{SD} = \hat{\rho}$,

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

We refer to this as the "*full-$\hat{\rho}$*" model because the subjective correlation estimate $\hat{\rho}$ is the same in both the drift and bound terms. This formulation is equivalent to assuming that misestimates of correlation magnitude are reflected in how the brain converts the observations to evidence, such that the correlation reflected in the internal observation distribution (which in the model is used to scale both the drift rate and bound height) is also used to compute logLR (which in the model

is used to scale the bound height). This formulation predicts that misestimated correlations (i.e., $\hat{\rho} \neq \rho$) affect the chronometric function (because changes in the subjective scaling of the drift rate and bound height result in correlation-dependent differences in the evidence per time step, and thus the time taken to reach the bound) but not the psychometric function (which depends only on the product of the drift rate and bound height, and thus the ρ- and $\hat{\rho}$-dependent terms cancel and accuracy is constant across correlations; Palmer et al., 2005; also see Fig. 5).

In the second case, when $\hat{\rho}_{SD} = \rho$,

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

This "*bound-$\hat{\rho}$*" model assumed that the statistics of the internal observation distribution followed objective ρ, and only the conversion of the observations to a weight of evidence was based on subjective $\hat{\rho}_B$ (i.e., the drift rate simplifies to $k_0\mu_0$, and the bound height was set as B_ρ, defined above). This formulation predicts that misestimated correlations affect both the chronometric and psychometric function (because the mis-specified weight of evidence results in correlation-dependent differences in the bound height and thus changes in the speed-accuracy trade-off; see Fig. 5–figure supplement 1).

We tested four additional variants of the DDM. The *base* model assumed no correlation-dependent scaling of the bound height. The *drift* model allowed for unconstrained correlation-dependent changes to the drift rate via three drift parameters fit separately for the three correlation conditions: k_-, k_0, k_+. The *scaled-$\hat{\rho}$* model fit $\hat{\rho}_{SD}$ and $\hat{\rho}_B$ separately, allowing subjective evidence weighing to diverge from the observed correlation. The *bound-$\hat{\rho}$ + drift* model used the form of the bound from the *bound-$\hat{\rho}$* model and included unconstrained, correlation-dependent adjustments in the drift rate as in the *drift* model.

For all models that fit subjective correlation estimates, separate parameters were used for the positive and negative correlation conditions ($\hat{\rho}_+$, $\hat{\rho}_-$). Additionally, for all models that did not fit $\hat{\rho}_{SD}$, we set $\hat{\rho}_{SD} = \rho$, which is equivalent to assuming that the objective, correlation-dependent stimulus strength was encoded correctly (but then could also be scaled linearly by k_0, as in the standard DDM; Palmer et al., 2005). 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 (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 (Fig. 6—figure supplement 1). Therefore, all models in the main analyses included a linear collapsing bound, using parameter t_B to determine the rate of linear collapse. For the *full-*$\hat{\rho}$ and *bound-*$\hat{\rho}$ models, the bound height at time *t* is then:

$$B_\rho^t = \frac{\sqrt{1+\hat{\rho}}}{\sqrt{1+\rho}}(B_0 - t_B t),$$

such that the correlation-dependent bound adjustment is applied to the instantaneous, and not the initial, bound (in a pilot analysis of data from 22 participants in the 0.6 correlation-magnitude

group, we found that the choice of whether to apply this adjustment to the instantaneous or initial bound had a negligible effect on model goodness-of-fit: ΔAIC = −0.9, protected exceedance probability = 0.63, in favor of scaling the instantaneous bound over the initial bound). The bounds are symmetric about the starting point, such that B_ρ^t is the distance between the starting point and either bound. Choice commitment occurs when one of the bounds is reached, which happens when $|x(t)| \geq B_\rho^t$, where $x(t)$ is the value of the decision variable at time *t*.

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 used the model fits to generate predicted performance for each participant for their actual fit correlation parameters, as well as for the true correlation values and correlations of zero, holding all other parameters fixed at their fit values. 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 objective 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 objective 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, 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., 2014). 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, NY: 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

Kang YH, Löffler A, Jeurissen D, Zylberberg A, Wolpert DM, Shadlen MN. 2021. Multiple decisions about one object involve parallel sensory acquisition but time-multiplexed evidence incorporation. *Elife* **10**. doi:10.7554/eLife.63721

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

Luyckx F, Spitzer B, Blangero A, Tsetsos K, Summerfield C. 2020. Selective integration during sequential sampling in posterior neural signals. *Cerebral Cortex* **30**:4454–4464. doi:10.1093/cercor/bhaa039

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.

Rangelov D, Fellrath J, Mattingley JB. 2024. Integrated perceptual decisions rely on parallel evidence accumulation. *The Journal of Neuroscience* **44**:e2368232024. doi:10.1523/JNEUROSCI.2368-23.2024

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. 2014. 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*.

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

Usher M, Tsetsos K, Glickman M, Chater N. 2019. Selective Integration: An Attentional Theory of Choice Biases and Adaptive Choice. *Curr Dir Psychol Sci* **28**:552–559. doi:10.1177/096372141986227

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

Wyart V, Myers NE, Summerfield C. 2015. Neural mechanisms of human perceptual choice under focused and divided attention. *The Journal of Neuroscience* **35**:3485–3498. doi:10.1523/JNEUROSCI.3276-14.2015

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

Zimmerman DW, Zumbo BD, Williams RH. 2003. Bias in estimation and hypothesis testing of correlation. *Psicológica* **24**:133–158.

Supplementary Figures

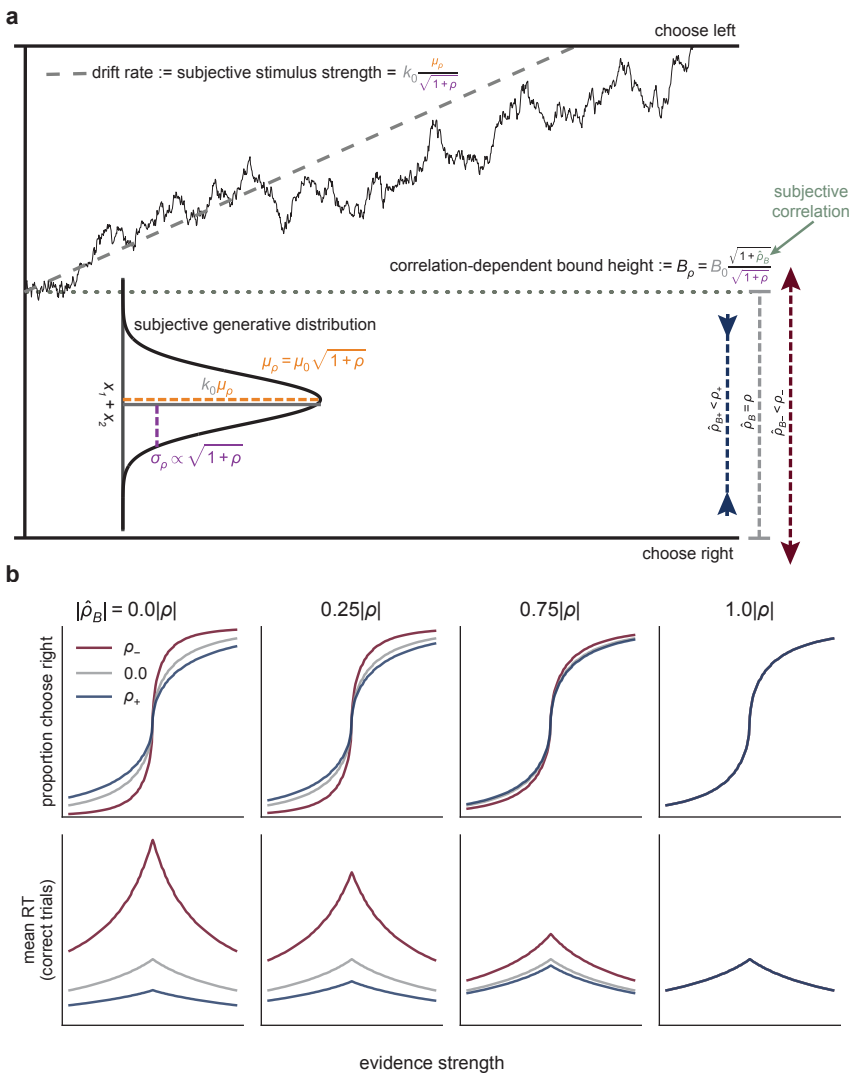

Figure 5—figure supplement 1. A drift-diffusion model (DDM) based on suboptimal evidence weighing. **a** For this DDM, $\hat{\rho}_{SD} = \rho$, but $\hat{\rho}_B$ is free to vary, leading to suboptimal evidence weighing. See Fig. 5 and text for a full description of the DDM. **b** Predictions from the 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**) but with a different subjective correlation $\hat{\rho}_B$ (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 ($|\hat{\rho}_B| < |\rho|$, first three columns) leads to differences in choices and RTs across the conditions, where the magnitude of the differences is a function of the degree of underestimation. Only $\hat{\rho}_B = \rho$ (rightmost column) produces equal predicted performance across conditions.

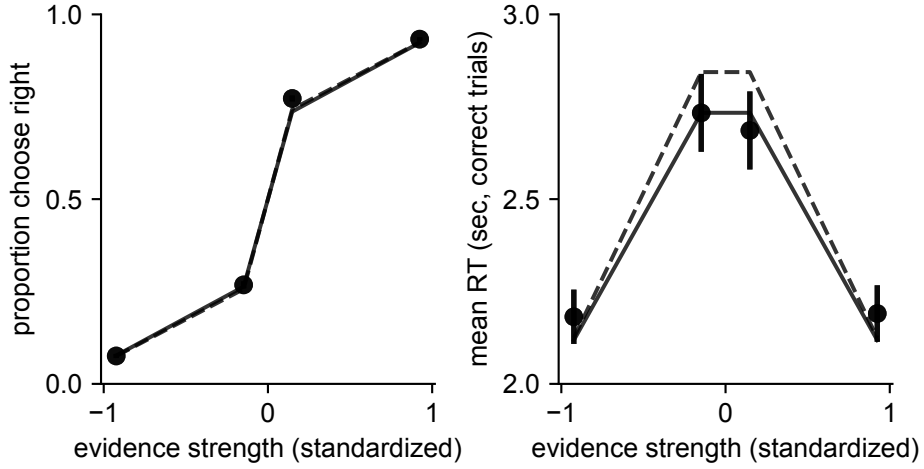

Figure 6—figure supplement 1. A collapsing-bound model accounts for behavior. Predictions from the DDM (lines) plotted against participant data (points) for choice (left) and RT (right) for the zero-correlation condition from all participants. Predictions and data are averaged across participants. Error bars are SEM. We fit two variants of a basic DDM. The first (dashed line) had 4 parameters: drift rate, bound height, non-decision time, and lapse rate. The second model (solid line; *base* model in main text) included an additional parameter governing the slope of a linear collapsing bound, for a total of 5 parameters. The collapsing bound model was a better fit to the data (mean ΔAIC = −8.47, protected exceedance probability [PEP] = 1.0 in favor of the collapsing bound model).

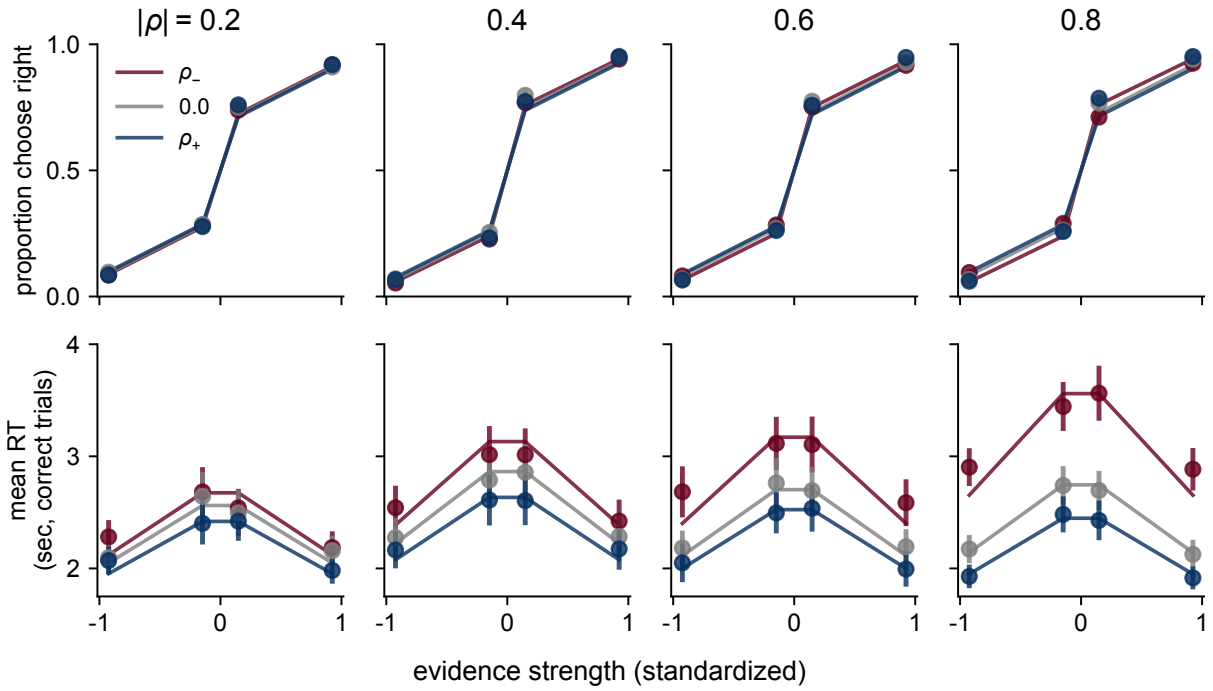

Figure 6—figure supplement 2. Predictions from the *bound-$\hat{\rho}$* DDM (lines) plotted against participant data (points) for choice (top) and RT (bottom) for each correlation-magnitude group (columns, labels at top). Predictions and data are averaged across participants. Colors correspond to the three correlation conditions (see legend). Error bars are SEM.

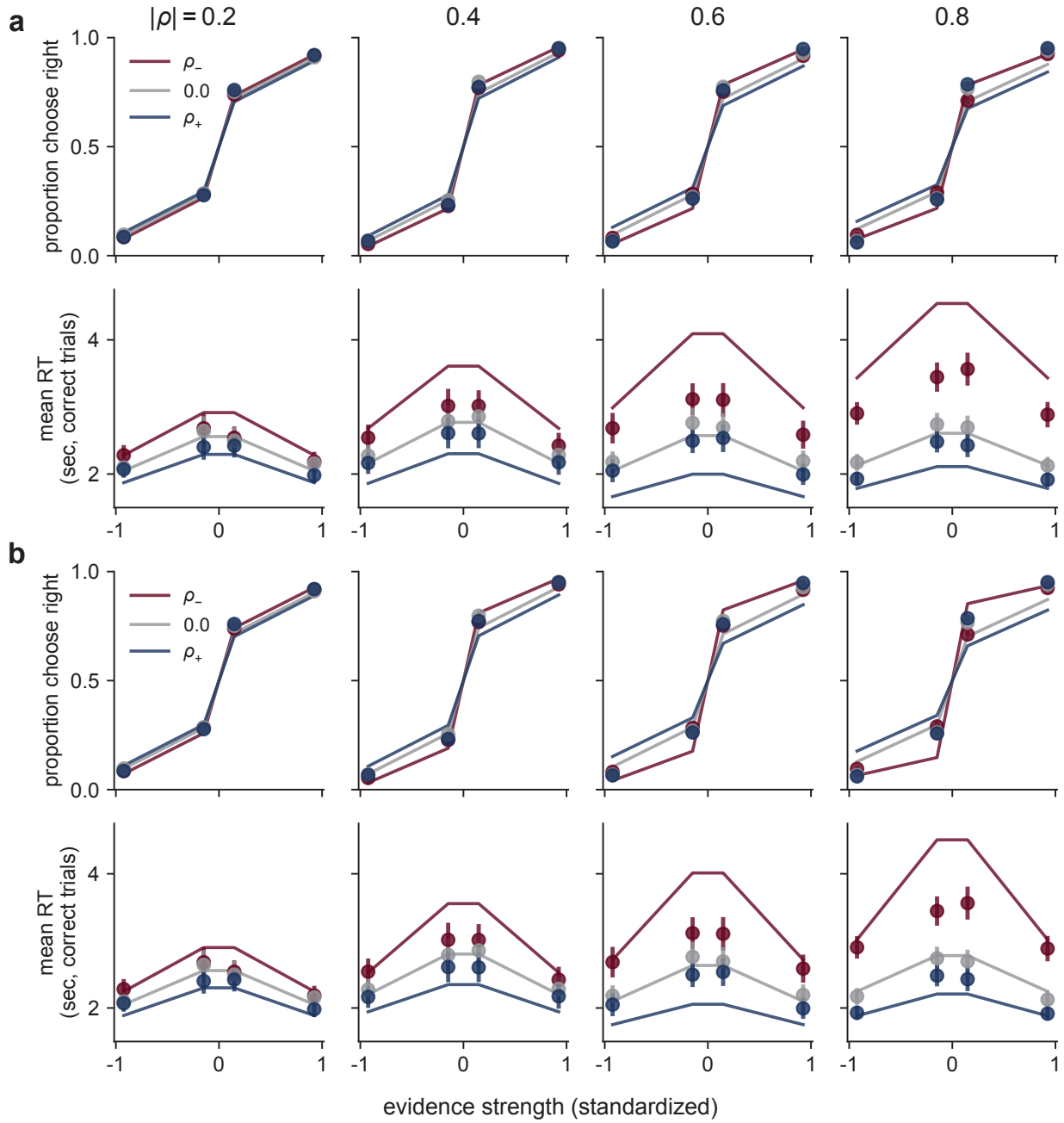

Figure 6—figure supplement 3. Predictions (lines) from the *base* model (**a**) and the *drift* model (**b**) plotted against participant data (points) for choice (top) and RT (bottom) for each correlation-magnitude group (columns, labels at top). Predictions and data are averaged across participants. Colors correspond to the three correlation conditions (see legend). Error bars are SEM.

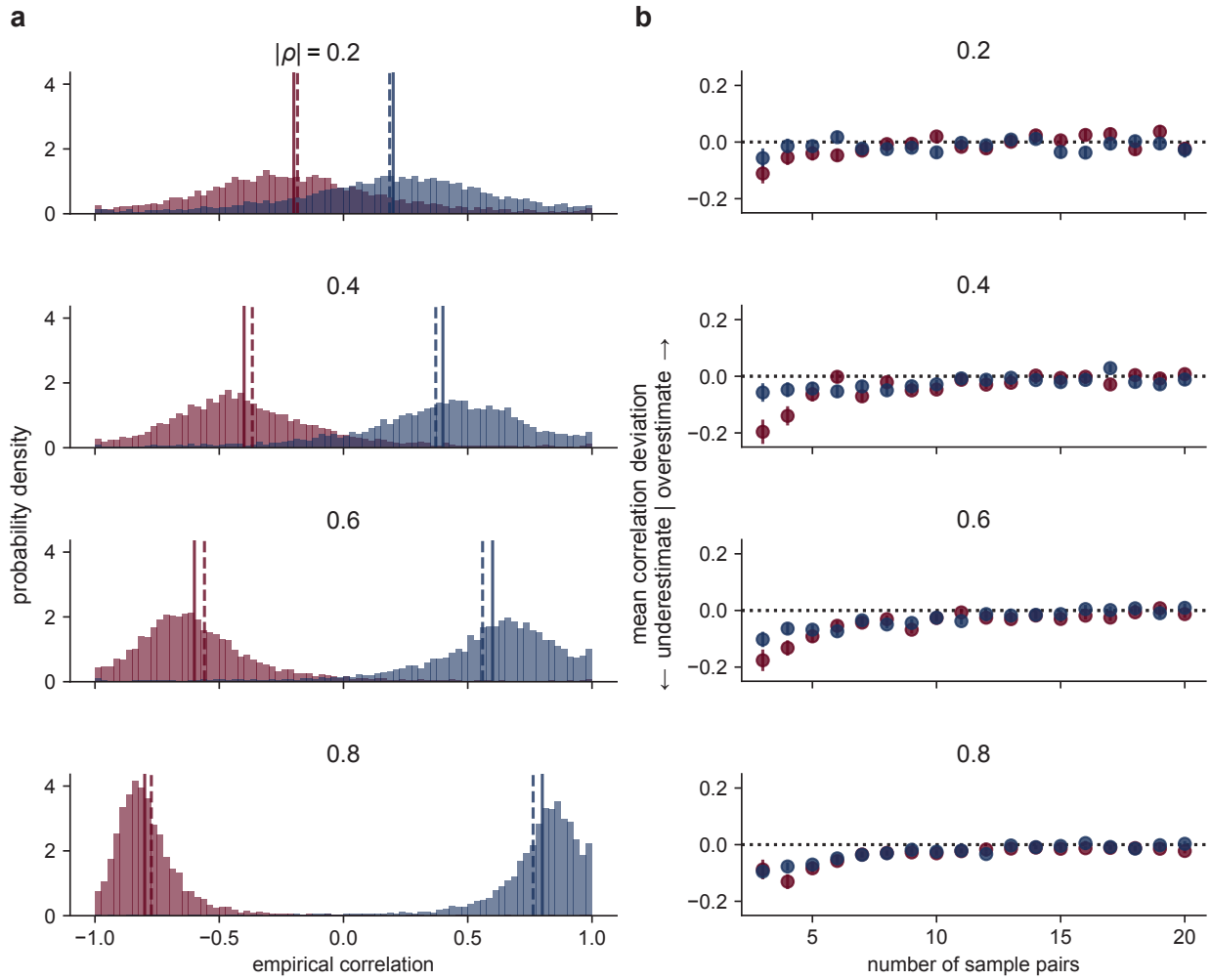

Figure 7—figure supplement 1. Empirical correlation estimates. **a** Distributions of correlation coefficients computed from the observed star positions on each trial, computed across all participants and trials, separately for each correlation magnitude (rows, as indicated) and sign (red for ρ_-, blue for ρ_+). Solid and dashed vertical lines are the true generative ρ and mean estimate, respectively, per condition. **b** Systematic misestimates of ρ as a function of the number of sample pairs used for the estimate, separated by correlation condition as in **a**. Negative/positive values are under-/over-estimates. Points and errorbars are mean±sem computed across all participants and trials. These analyses included only trials with at least three sample pairs to estimate a non-degenerate correlation.

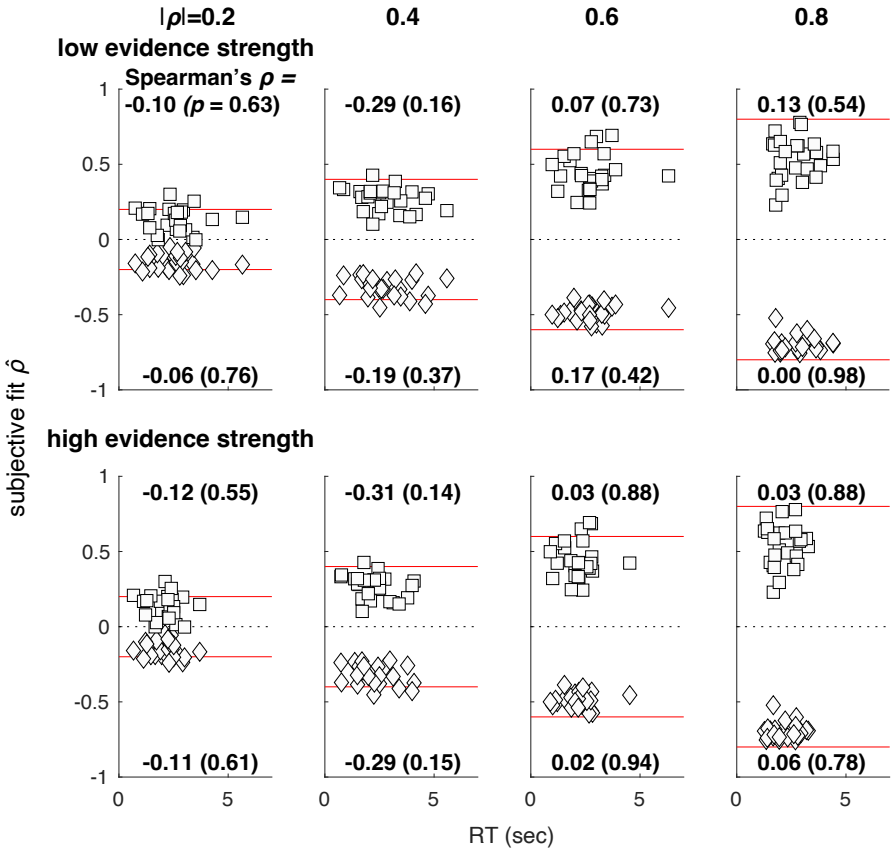

Figure 7—figure supplement 2. Subjective correlation (best-fitting $\hat{\rho}$) versus mean response time (RT) for each participant. Rows are data separated by low (top) and high (bottom) objective evidence strength. Columns are data separated by correlation-magnitude group, as indicated at the top. Each panel shows data plotted separately for positive (squares) and negative (diamonds) objective ρ, along with the associated Spearman's correlation coefficient of subjective correlation versus RT (p-value in parentheses, uncorrected for multiple comparisons). Red lines are objective correlation values. For each participant, there was a single, best-fitting value of $\hat{\rho}_+$ and $\hat{\rho}_-$ for the positive and negative correlation conditions, respectively, fit jointly to all data from both evidence strengths, but different mean RTs per evidence strength.

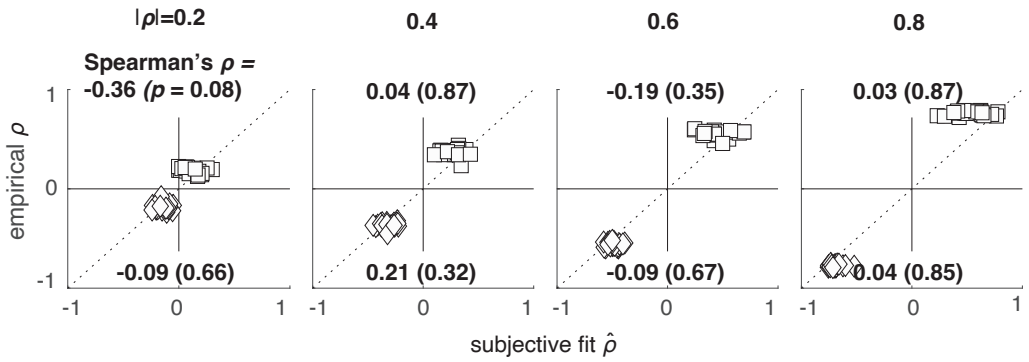

Figure 7—figure supplement 3. No systematic relationship between the mean of the empirical correlation computed on each trial (ordinate) and the best-fitting subjective correlation (abscissa) for positive (squares) and negative (diamonds) correlations. Points are data from individual participants. Columns are data separated by correlation-magnitude group, as indicated. Each panel shows the Spearman correlation coefficient comparing the empirical and subjective correlations and the associated *p*-value (uncorrected for multiple comparisons), for each correlation condition.

Appendix 1

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.775	0.513	0.103
	t_B	0.128	0.123	0.025
	k_0	17.458	3.405	0.681
	ndt	0.328	0.162	0.032
	λ	0.011	0.008	0.002
0.4	B_0	1.911	0.644	0.129
	t_B	0.131	0.109	0.022
	k_0	19.543	4.723	0.945
	ndt	0.399	0.284	0.057
	λ	0.013	0.018	0.004
0.6	B_0	1.792	0.556	0.111
	t_B	0.129	0.108	0.022
	k_0	18.021	3.302	0.66
	ndt	0.264	0.11	0.022
	λ	0.049	0.044	0.009
0.8	B_0	1.779	0.49	0.098
	t_B	0.214	0.13	0.026
	k_0	20.304	3.089	0.618
	ndt	0.359	0.171	0.034
	λ	0.119	0.049	0.01

Appendix 1—table 1. *Base* model average best-fitting parameter values.

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.772	0.504	0.101
	t_B	0.128	0.121	0.024
	k_0	17.382	3.73	0.746
	k_-	18.122	3.551	0.71
	k_+	17.169	3.529	0.706
	ndt	0.332	0.17	0.034
	λ	0.011	0.009	0.002
0.4	B_0	1.942	0.669	0.134
	t_B	0.135	0.116	0.023
	k_0	19.081	4.693	0.939
	k_-	22.008	5.515	1.103
	k_+	17.75	4.109	0.822
	ndt	0.392	0.273	0.055
	λ	0.013	0.017	0.003
0.6	B_0	1.818	0.566	0.113
	t_B	0.127	0.108	0.022
	k_0	17.122	3.123	0.625
	k_-	21.592	4.125	0.825
	k_+	15.932	2.926	0.585
	ndt	0.248	0.115	0.023
	λ	0.05	0.045	0.009
0.8	B_0	1.843	0.477	0.095
	t_B	0.19	0.127	0.025
	k_0	18.43	2.662	0.532
	k_-	29.082	8.611	1.722
	k_+	17.301	2.703	0.541
	ndt	0.316	0.19	0.038
	λ	0.115	0.047	0.009

Appendix 1—table 2. *Drift* model average best-fitting parameter values.

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.766	0.528	0.106
	$\hat{\rho}_{B-}$	−0.131	0.085	0.017
	$\hat{\rho}_{B+}$	0.108	0.111	0.022
	t_B	0.127	0.131	0.026
	k_0	17.66	3.69	0.738
	ndt	0.336	0.164	0.033
	λ	0.012	0.009	0.002
0.4	B_0	1.948	0.702	0.14
	$\hat{\rho}_{B-}$	−0.285	0.088	0.018
	$\hat{\rho}_{B+}$	0.245	0.149	0.03
	t_B	0.124	0.114	0.023
	k_0	19.244	4.762	0.952
	ndt	0.402	0.282	0.056
	λ	0.01	0.011	0.002
0.6	B_0	1.796	0.51	0.102
	$\hat{\rho}_{B-}$	−0.484	0.07	0.014
	$\hat{\rho}_{B+}$	0.443	0.158	0.032
	t_B	0.087	0.093	0.019
	k_0	16.972	3.263	0.653
	ndt	0.267	0.126	0.025
	λ	0.009	0.007	0.001
0.8	B_0	1.856	0.48	0.096
	$\hat{\rho}_{B-}$	−0.685	0.065	0.013
	$\hat{\rho}_{B+}$	0.493	0.137	0.027
	t_B	0.106	0.099	0.02
	k_0	17.1	2.77	0.554
	ndt	0.285	0.179	0.036
	λ	0.013	0.014	0.003

Appendix 1—table 3. *Bound-$\hat{\rho}$* model average best-fitting parameter values.

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.777	0.53	0.106
	$\hat{\rho}_-$	−0.15	0.055	0.011
	$\hat{\rho}_+$	0.132	0.08	0.016
	t_B	0.126	0.124	0.025
	k_0	17.488	3.522	0.704
	ndt	0.326	0.164	0.033
	λ	0.011	0.01	0.002
0.4	B_0	1.967	0.692	0.138
	$\hat{\rho}_-$	−0.317	0.066	0.013
	$\hat{\rho}_+$	0.268	0.08	0.016
	t_B	0.128	0.112	0.022
	k_0	19.401	4.981	0.996
	ndt	0.39	0.29	0.058
	λ	0.008	0.007	0.001
0.6	B_0	1.803	0.505	0.101
	$\hat{\rho}_-$	−0.487	0.05	0.01
	$\hat{\rho}_+$	0.445	0.123	0.025
	t_B	0.089	0.098	0.02
	k_0	17.62	3.639	0.728
	ndt	0.272	0.111	0.022
	λ	0.011	0.016	0.003
0.8	B_0	1.822	0.429	0.086
	$\hat{\rho}_-$	−0.697	0.052	0.01
	$\hat{\rho}_+$	0.536	0.139	0.028
	t_B	0.097	0.095	0.019
	k_0	18.182	3.094	0.619
	ndt	0.305	0.175	0.035
	λ	0.011	0.014	0.003

Appendix 1—table 4. *Full-$\hat{\rho}$* model average best-fitting parameter values.

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.795	0.529	0.106
	$\hat{\rho}_{B-}$	−0.098	0.088	0.018
	$\hat{\rho}_{B+}$	0.161	0.14	0.028
	t_B	0.13	0.126	0.025
	k_0	17.898	3.808	0.762
	$\hat{\rho}_{SD-}$	−0.013	0.256	0.051
	$\hat{\rho}_{SD+}$	0.222	0.284	0.057
	ndt	0.324	0.162	0.032
	λ	0.013	0.011	0.002
0.4	B_0	1.959	0.701	0.14
	$\hat{\rho}_{B-}$	−0.298	0.083	0.017
	$\hat{\rho}_{B+}$	0.351	0.121	0.024
	t_B	0.13	0.118	0.024
	k_0	19.717	4.745	0.949
	$\hat{\rho}_{SD-}$	−0.276	0.155	0.031
	$\hat{\rho}_{SD+}$	0.443	0.277	0.055
	ndt	0.4	0.279	0.056
	λ	0.011	0.012	0.002
0.6	B_0	1.82	0.535	0.107
	$\hat{\rho}_{B-}$	−0.431	0.078	0.016
	$\hat{\rho}_{B+}$	0.452	0.101	0.02
	t_B	0.091	0.092	0.018
	k_0	17.918	3.651	0.73
	$\hat{\rho}_{SD-}$	−0.328	0.161	0.032
	$\hat{\rho}_{SD+}$	0.464	0.24	0.048
	ndt	0.26	0.109	0.022
	λ	0.012	0.014	0.003
0.8	B_0	1.87	0.471	0.094
	$\hat{\rho}_{B-}$	−0.677	0.07	0.014
	$\hat{\rho}_{B+}$	0.483	0.16	0.032
	t_B	0.109	0.1	0.02
	k_0	18.066	3.262	0.652
	$\hat{\rho}_{SD-}$	−0.623	0.162	0.032
	$\hat{\rho}_{SD+}$	0.41	0.262	0.052
	ndt	0.286	0.177	0.035
	λ	0.013	0.015	0.003

Appendix 1—table 5. *Scaled-$\hat{\rho}$* model average best-fitting parameter values.

|ρ|	parameter	mean	SD	SEM
0.2	B_0	1.766	0.525	0.105
	$\hat{\rho}_{B-}$	−0.126	0.112	0.022
	$\hat{\rho}_{B+}$	0.145	0.127	0.025
	t_B	0.127	0.126	0.025
	k_0	17.149	3.28	0.656
	k_-	17.36	3.666	0.733
	k_+	18.186	3.631	0.726
	ndt	0.341	0.167	0.033
	λ	0.012	0.009	0.002
0.4	B_0	1.931	0.705	0.141
	$\hat{\rho}_{B-}$	−0.315	0.095	0.019
	$\hat{\rho}_{B+}$	0.275	0.146	0.029
	t_B	0.123	0.116	0.023
	k_0	19.551	4.923	0.985
	k_-	18.644	4.833	0.967
	k_+	19.676	4.531	0.906
	ndt	0.419	0.268	0.054
	λ	0.01	0.008	0.002
0.6	B_0	1.829	0.547	0.109
	$\hat{\rho}_{B-}$	−0.518	0.072	0.014
	$\hat{\rho}_{B+}$	0.419	0.193	0.039
	t_B	0.092	0.098	0.02
	k_0	17.547	3.283	0.657
	k_-	14.911	3.372	0.674
	k_+	18.446	4.074	0.815
	ndt	0.27	0.104	0.021
	λ	0.011	0.01	0.002
0.8	B_0	1.856	0.484	0.097
	$\hat{\rho}_{B-}$	−0.703	0.05	0.01
	$\hat{\rho}_{B+}$	0.554	0.154	0.031
	t_B	0.106	0.098	0.02
	k_0	18.572	3.459	0.692
	k_-	14.138	2.881	0.576
	k_+	20.279	3.632	0.726
	ndt	0.298	0.166	0.033
	λ	0.013	0.015	0.003

Appendix 1—table 6. *Bound-$\hat{\rho}$ + drift* model average best-fitting parameter values.