A nonlinear updating algorithm captures suboptimal inference in the presence of signal-dependent noise

Seth W. Egger[1,2*], Mehrdad Jazayeri[1,2]

1 McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA
2 Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA

* swegger@mit.edu

Abstract

Bayesian models of behavior have advanced the idea that humans combine prior beliefs and sensory observations to minimize uncertainty. How the brain implements Bayes-optimal inference, however, remains poorly understood. Simple behavioral tasks suggest that the brain can flexibly represent and manipulate probability distributions. An alternative view is that brain relies on simple algorithms that can implement Bayes-optimal behavior only when the computational demands are low. To distinguish between these alternatives, we devised a task in which Bayes-optimal performance could not be matched by simple algorithms. We asked subjects to estimate and reproduce a time interval by combining prior information with one or two sequential measurements. In the domain of time, measurement noise increases with duration. This property makes the integration of multiple measurements beyond the reach of simple algorithms. We found that subjects were able to update their estimates using the second measurement but their performance was suboptimal, suggesting that they were unable to update full probability distributions. Instead, subjects' behavior was consistent with an algorithm that predicts upcoming sensory signals, and applies a nonlinear function to errors in prediction to update estimates. These results indicate that inference strategies humans deploy may deviate from Bayes-optimal integration when the computational demands are high.

Introduction

Sensorimotor control depends on accurate estimation of internal state variables [1–5]. Numerous experiments have used Bayesian estimation theory to demonstrate that humans estimate internal states by integrating multiple sources of information including prior beliefs and sensory cues from various modalities [6–16]. Bayesian estimation is typically formulated in terms of three components: prior distributions representing *a priori* beliefs about state variables, likelihood functions derived from noisy sensory measurements, and cost functions that characterize reward contingencies [17]. In this formulation, the likelihood function and prior distribution are combined to compute a posterior distribution and the cost function is used to extract an estimate that maximizes expected reward. This formulation is the basis of most psychophysical studies of Bayesian integration [9–15, 18–20].

Implicit in this formulation is the assumption that the brain has access to priors, likelihoods, and cost functions. Access to these quantities is appealing as it could

support rapid and optimal state estimation without the need to learn new policies for novel behavioral contexts [21, 22]. However, in most experiments, Bayes-optimal behavior can also be achieved by simpler algorithms that do not depend on direct access to likelihoods, priors and cost functions [21–23]. For example, optimal cue combination in the presence of Gaussian noise may be implemented by a weighted sum of measurements [6]. Similarly, integration of noisy evidence with prior beliefs may be implemented by a suitable functional mapping between measurements and estimates [24, 25]. Finally, online estimation of a variable from sequential measurements that are subject to Gaussian noise can be achieved by a *Kalman filter* that only keeps track of the mean and variance [26] without representing and updating the full posterior distribution.

In contrast to simple laboratory tasks, optimal inference in natural settings is often intractable and involves approximations that may deviate from optimality [27]. Therefore, it is critical to go beyond statements of optimality and suboptimality, and assess the inference algorithms humans use during sensorimotor and cognitive tasks [28, 29]. Further, as articulated by Marr [30], characterization of the underlying algorithms could establish a bridge between behaviorally relevant computations and neurobiological mechanisms.

We devised an experiment in which the computational demands for optimal Bayesian estimation were incompatible with simple algorithmic solutions. Subjects had to reproduce an interval by integrating their prior belief with one or two measurements of the interval. In the domain of time, the noise is signal-dependent, which causes behavioral variability to scale with duration [31]. Theoretical considerations suggest that an important consequence of scalar variability is that simple algorithms that only update certain parameters of the posterior (e.g., mean and/or variance) cannot emulate Bayes-optimal behavior. Therefore, optimal behavior in this paradigm would provide strong evidence that the underlying inference algorithm involves updating probability distributions. Conversely, suboptimal behavior would suggest that subjects rely on a simpler algorithm. We found that when subjects made two measurements their performance was suboptimal. Furthermore, comparison of behavior with various models indicated that subjects relied on an inference algorithm that used measurements to update point estimates using point nonlinearities.

Results

Subjects integrate interval measurements with prior knowledge

Subjects performed an interval reproduction task consisting of two randomly interleaved trial types (Fig 1A,B). In "1-2-Go" trials, two flashes (S1 followed by S2) demarcated a sample interval (t_s). Subjects had to reproduce t_s immediately after S2. The time between the onset of S2 and when the keyboard was pressed was designated as the production interval (t_p). In "1-2-3-Go" trials, t_s was presented twice, demarcated once by S1 and S2 and once by S2 and S3, providing the opportunity to make two measurements (Fig 1B). Similar to 1-2-Go, subjects had to match t_p (the interval between S3 and keyboard press) to t_s. Across trials, t_s was drawn from a discrete uniform distribution ranging between 600 and 1000 ms (Fig 1C). Subjects received two forms of trial-by-trial feedback based on the magnitude and sign of the error. First, a feedback stimulus was presented whose location relative to the warning stimulus reflected the magnitude and sign of the error (Fig 1A,B; see Materials and methods). Second, if the error exceeded a threshold window (Fig 1D), stimuli remained white and a tone denoting incorrect response was presented. Otherwise, the stimuli turned green and a tone denoting correct was presented. The threshold window for correct performance

was proportionally larger for the longer t_s to accommodate the scalar variability of timing due to signal-dependent noise [40–45]. The threshold was adjusted adaptively and on a trial-by-trial basis according to performance (see Materials and methods).
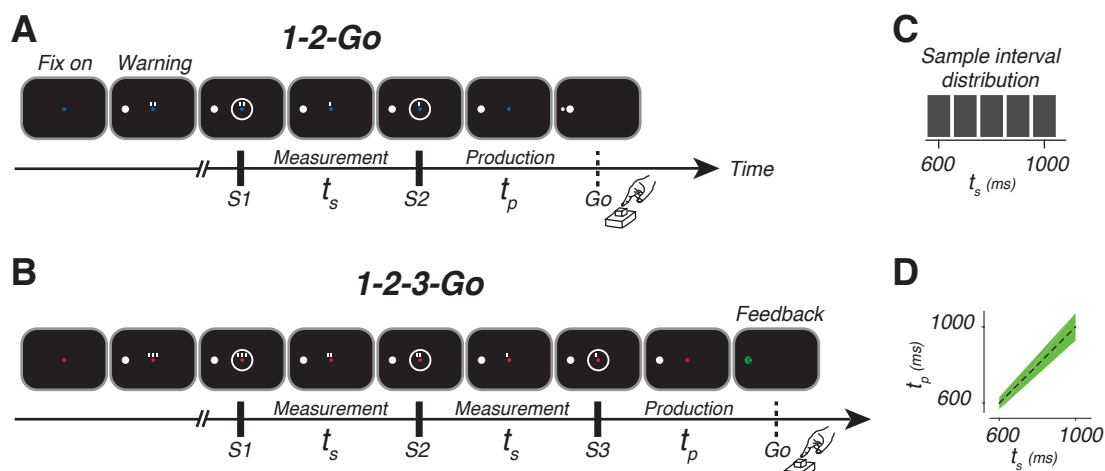

Fig 1. The 1-2-Go and 1-2-3-Go interval reproduction task. (A,B) Task design. Each trial began with the appearance of a fixation spot (Fix on). The color of the fixation spot informed the subject of the trial type: blue for 1-2-Go, and red for 1-2-3-Go. After a random delay, a warning stimulus (large white circle) appeared. Additionally, two or three small white rectangles were presented above the fixation spot. The number of rectangles was associated with the number of upcoming flashes. After another random delay, two (S1 and S2 for 1-2-Go) or three (S1, S2 and S3 for 1-2-3-Go) white annulae were flashed for 100 ms in a sequence around the fixation spot. Consecutive flashes were separated by the duration of the sample interval (t_s). With the disappearance of each flash, one of the small rectangles also disappeared (rightmost first and leftmost last). The white rectangles were provided to help subjects keep track of events during the trial. Subjects had press a button after the last flash to produce an interval (t_p) that matched t_s. Immediately after button press, subjects received analog feedback. The analog feedback was a small circle that was presented to the left or right of the warning stimulus depending on whether t_p was larger or smaller than t_s, respectively. The distance of the feedback circle to the center of the warning stimulus was proportional to the magnitude of the error ($t_p - t_s$). (C) Experimental distribution of sample intervals. (D) Feedback. Subjects were given positive feedback if their production times fell within the green region. The width of the positive feedback window was scaled with t_s.

Subjects' timing behavior exhibited three characteristic features (Fig 2). First, t_p increased monotonically with t_s. Second, t_p was systematically biased toward the mean of the prior, as evident from the tendency of responses to deviate from t_s (diagonal) and gravitate toward the mean t_s. As proposed previously [24, 34–36], this so-called regression to the mean indicated that subjects relied on their knowledge of the prior distribution of t_s. Third, performance was better in 1-2-3-Go condition in which subjects made two measurements, as evidenced by a lower root-mean-square error (RMSE) in 1-2-3-Go compared to 1-2-Go condition (Fig 2C; permutation test; p-value < 0.01 for all subjects). This observation indicates that subjects combined the two measurements to improve their estimates, corroborating reports from other behavioral paradigms [46–54]. Combined with the systematic bias toward the mean of t_s, these results indicated that subjects integrated prior information with one or two measurements to improve their performance.

A Bayesian model of behavior

Building on previous work [11, 18, 22, 55], we asked whether subjects' behavior could be accounted for by a Bayesian observer model based on the Bayes-Least Squares (BLS)
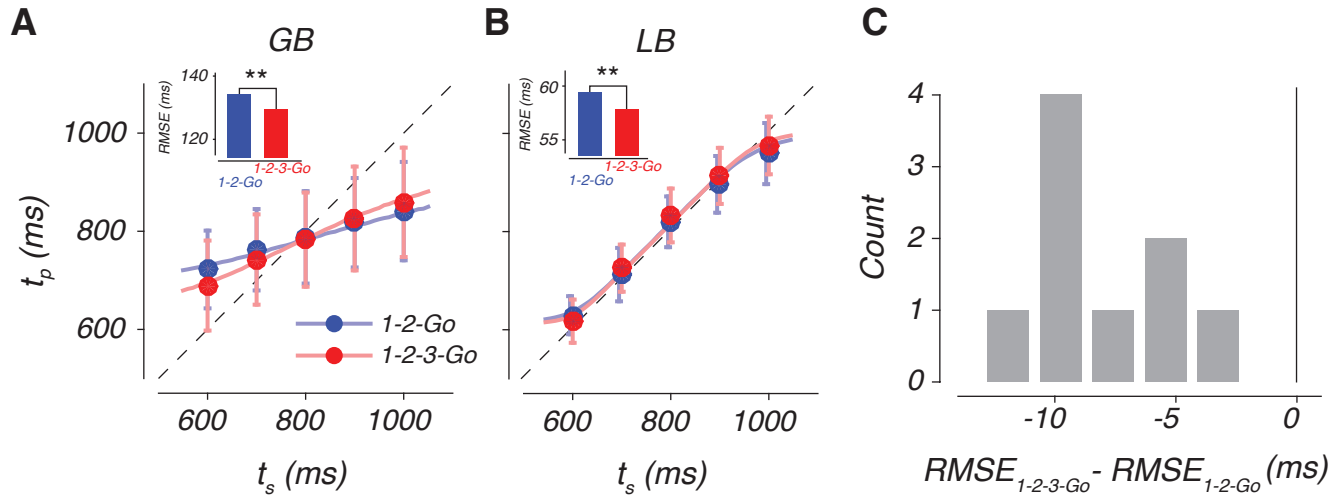
Fig 2. Performance in the interval reproduction task. (A) Production interval (t_p) as a function of sample interval (t_s) for a low sensitivity subject. Filled circles and error bars show the mean and standard deviation of t_p for each t_s in the 1-2-Go (blue) and 1-2-3-Go (red) conditions. The dotted unity line represents perfect performance and the colored lines show the expected t_p from a Bayes Least-Squares (BLS) model fit to the data. Inset: root-mean-square error (RMSE) in the 1-2-Go (blue) and 1-2-3-Go (red) conditions differed significantly (asterisk, p-value < 0.01; permutation test). (B) Same as (A) for a high sensitivity subject (LB). (C) The histogram of changes in RMSE across conditions for all subjects.

estimator. For the 1-2-Go trials, the observer model (1) makes a noisy measurement of t_s, which we denote by t_{m_1}, (2) combines the likelihood function associated with t_{m_1}, $p(t_{m_1}|t_s)$, with the prior distribution of t_s, $p(t_s)$, to compute the posterior, $p(t_s|t_{m_1})$, and (3) uses the mean of the posterior as the optimal estimate, t_{e_1}. We modeled $p(t_{m_1}|t_s)$ as a Gaussian distribution centered at t_s with standard deviation, σ_m, proportional to t_s with constant of proportionality, w_m; i.e., $\sigma_m = w_m t_s$ (Fig 3A, left box). We assumed that the production process was also perturbed by noise and modeled t_p as a sample from a Gaussian distribution centered at t_{e_1} with standard deviation, σ_p, proportional to t_{e_1} with constant of proportionality, w_p; i.e., $\sigma_p = w_p t_{e_1}$ (Fig 3A, right box). Note that the entire operation of the BLS estimator can be described in terms of a deterministic mapping of t_{m_1} to t_{e_1} using a nonlinear function, which we denote as $f_{BLS_1}(t_{m_1})$ (Fig 3B) [24].

For the 1-2-3-Go trials, the subject makes two measurements, t_{m_1} and t_{m_2}, and uses the mean of the posterior based on both measurements, $p(t_s|t_{m_1}, t_{m_1})$, to derive an optimal estimate, t_{e_2} (Fig 3C; see Materials and methods). In these trials, the mapping from t_{m_1} and t_{m_2} to t_{e_2} can be described in terms of a two-dimensional nonlinear function, denoted by $f_{BLS_2}(t_{m_1}, t_{m_2})$ (Fig. 3D). Note that the iso-estimate contours of $f_{BLS_2}(t_{m_1}, t_{m_2})$ are both nonlinear and convex (Fig. 3D, red). The nonlinearity indicates that the effect of t_{m_1} and t_{m_2} on t_{e_2} is non-separable, and the convexity indicates that t_{e_2} is more strongly influenced by the larger of the two measurements. These features are direct consequences of scalar noise and are not present when measurements are perturbed by Gaussian noise (see S1 Appendix).

We fit the model to each subject's data assuming that responses in both 1-2-Go and 1-2-3-Go conditions were associated with the same w_m and w_p (Materials and methods). The model was augmented in two ways to ensure that estimates of w_m and w_p were accurate. First, we included an offset parameter to absorb interval-independent biases (e.g., consistently pressing the button too early or too late). Second, trials in which t_p grossly deviated from t_s were designated as "lapse" trials (see Materials and methods).
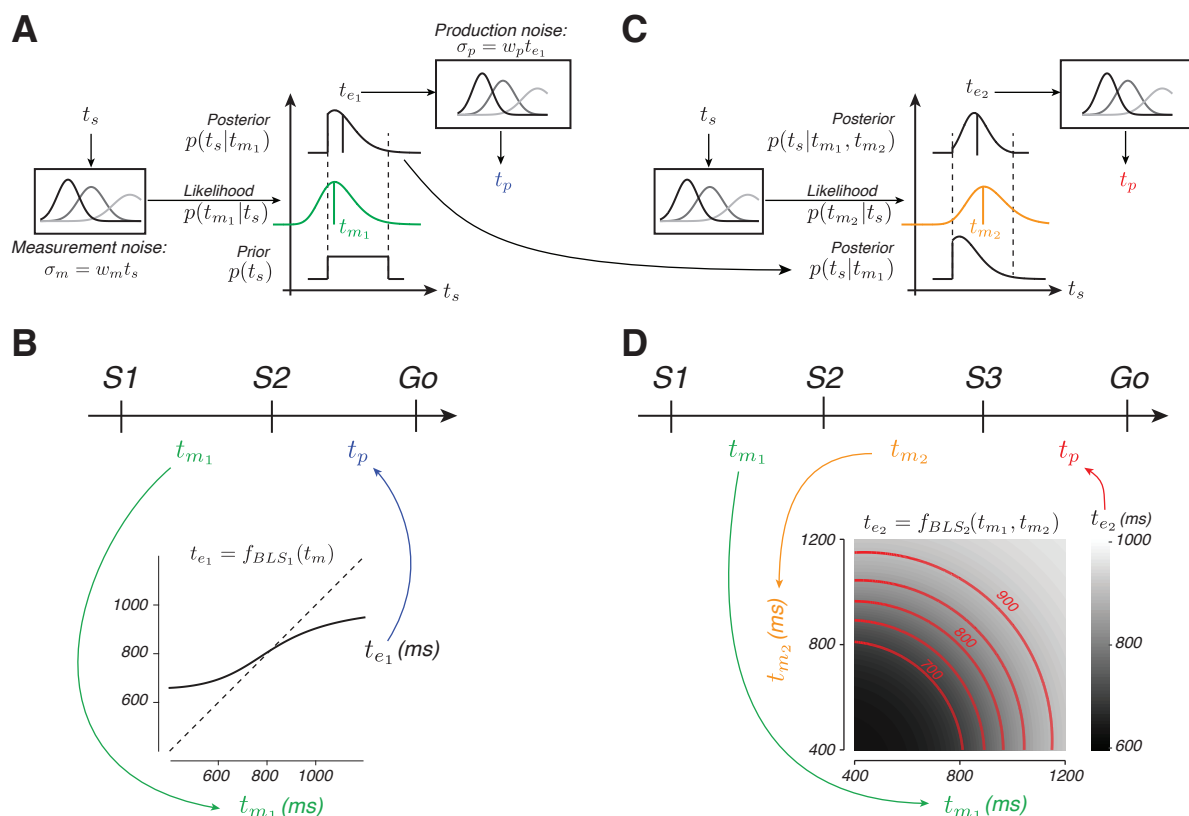
Fig 3. BLS model of interval integration. (A) BLS model for 1-2-Go trials. The left panel illustrates the measurement process. The measured interval, t_{m_1}, is perturbed by zero-mean Gaussian noise whose standard deviation is proportional to the sample interval, t_s, with constant of proportionality w_m ($\sigma_m = w_m t_s$). The middle panel illustrates the estimation process. The model multiplies the likelihood function associated with t_{m_1} (middle panel, green) with the prior (bottom), and uses the mean of the posterior (top) to derive an interval estimate (t_{e_1}, black vertical line on the posterior). The right panel illustrates the production process. The produced interval, t_p, is perturbed by zero-mean Gaussian noise with standard deviation proportional to t_{e_1}, with constant of proportionality w_p ($\sigma_p = w_p t_{e_1}$). (B) The effective mapping function (f_{BLS_1}, black curve) from the first measurement, t_{m_1}, to the optimal estimate, t_{e_1}. The dashed line indicates unity. (C) BLS model for 1-2-3-Go trials. The model uses the posterior after the first measurement, $p(t_s|t_{m_1})$, as the prior and combines it with the likelihood of the second measurement (t_{m_2}, orange) to compute an updated posterior, $p(t_s|t_{m_1}, t_{m_2})$. The mean of the updated posterior is taken as the interval estimate (t_{e_2}). (D) The effective mapping function (f_{BLS_2}, grayscale) from each combination of measurements, t_{m_1} and t_{m_2}, to the optimal the estimate, t_{e_2}. Red lines indicate combinations of measurements that lead to identical estimates (shown for $t_{e_2} = 700, 750, 800, 850,$ and 900 ms).

Model fits captured subjects' behavior for both conditions as shown by a few representative subjects (Fig 4A, 2A and 2B; see Supporting information for fits to all the subjects). Following previous work [24], we evaluated model fits using two statistics, an overall bias, BIAS, and an overall variability, $\sqrt{\text{VAR}}$ (see Materials and methods). As shown in Fig 4B, the model broadly captured the bias and variance for all subjects in both 1-2-Go and 1-2-3-Go conditions. However, the observed BIAS was significantly larger than predicted by the model fits in the 1-2-3-Go condition (Fig 4B, inset; two tailed t-test, t(8) = 4.6982, p-value = 0.0015), but not in the 1-2-Go condition (two tailed t-test, t(8) = -0.3236, p-value = 0.7546).

We quantified this observation across subjects by normalizing each subject's RMSE in the 1-2-Go and 1-2-3-Go conditions to the RMSE expected from the BLS model in the 1-2-3-Go condition (Fig 4C). We found that the observed RMSE in the 1-2-3-Go
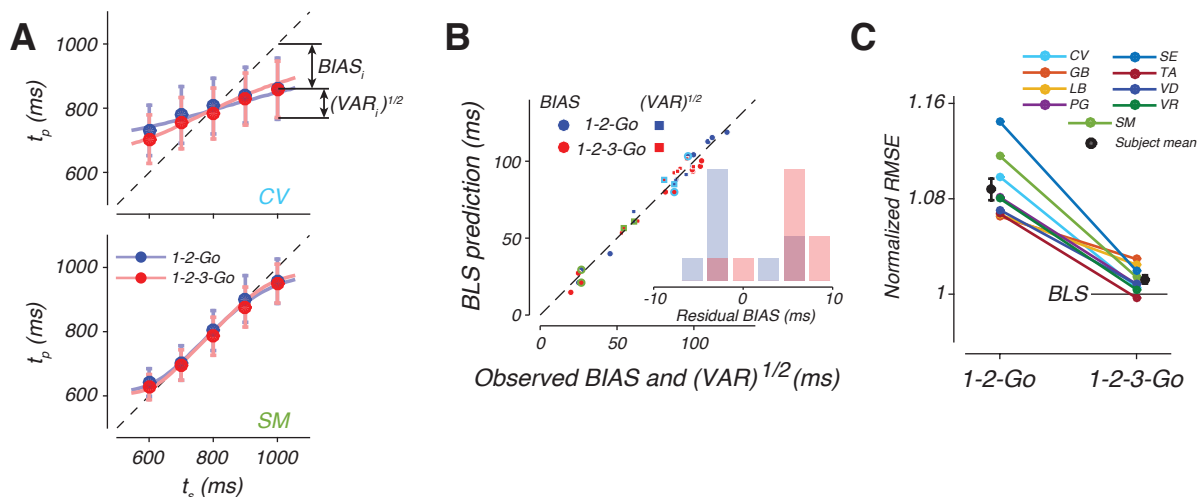
Fig 4. BLS model fits to data. (A) Behavior of two subjects and the corresponding BLS model fits with the same format as in Fig 2A and 2B. (B) BIAS (circles) and $\sqrt{\text{VAR}}$ (squares) of each subject (abscissa) and the corresponding values computed from simulations of the fitted BLS model (ordinate). Red and blue points correspond to 1-2-Go and 1-2-3-Go, respectively. The dotted line plots unity. Data points corresponding to subjects SM and CV are marked by light green and light blue, respectively. Inset: difference between the BIAS observed from data and that expected by the BLS model fit for 1-2-Go (blue) and 1-2-3-Go (red) conditions. (C) Comparison of behavioral performance to model predictions. Each line connects the RMSE for 1-2-Go (left) and 1-2-3-Go (right) conditions for one subject. To facilitate comparison across subjects, RMSE values for each subject were normalized by the RMSE of the BLS model in the 1-2-3-Go condition. The black circles and error bars correspond to the mean and standard error of the normalized RMSE across subjects. See also Supporting information.

condition was significantly larger than expected (two tailed t-test, $t(8) = 3.5484$, p-value = 0.007). Further, the drop in observed RMSE in the 1-2-3-Go was significantly less than expected by the BLS model (one tailed t-test, $t(8) = -4.4600$, p-value = 0.0011). These analyses indicate that subjects were able to integrate the two measurements but failed to optimally update the posterior by the likelihood information associated with the second measurement.

An algorithmic view of Bayesian integration

The success of the BLS model in capturing behavior in the 1-2-Go condition [24, 34, 36] and its failure in the 1-2-3-Go condition suggests that subjects were unable to update the posterior by the second measurement. We examined a number of simple inference algorithms that could account for this limitation. One of the simplest algorithms proposed for integrating sequential measurements is the Kalman filter. The Kalman filter only updates the mean and variance of the posterior [26]. This strategy is optimal when measurement noise is Gaussian because a Gaussian distribution is fully determined by its mean and variance. More generally, when integrating the likelihood function leaves the parametric form of the posterior distribution unchanged, a simple inference algorithm that updates those parameters can implement optimal integration.

First, we asked whether there exists a similarly simple and optimal updating algorithm when the noise is signal-dependent (i.e., scalar noise). For the posterior to have the same parametric form after one and two measurements, it is necessary that the product of two likelihood functions have the same parametric form as a single likelihood function. We tested this property analytically and verified that the parametric form of the likelihood function associated with scalar noise was not invariant under

multiplication (see S1 Appendix). Therefore, any inference algorithm that only updates certain statistics of the posterior (e.g., mean and variance) is expected to behave suboptimally when multiple time intervals have to be integrated. Therefore, we hypothesized that subjects might have used a simple updating algorithm analogous to the Kalman filter to integrate multiple measurements.

A linear-nonlinear estimator (LNE) model for approximate Bayesian inference

The first algorithm we tested was one in which the observer combines the last estimate $\bar{t}_{n-1}$, with the current measurement, t_{m_n}, using a linear updating strategy. If we denote the corresponding weights by $1 - k_n$ and k_n and set $k_n = 1/n$, this algorithm tracks the running average of the measurements, $\bar{t}_n$ (k_1 and k_2 are 1 and 0.5, respectively). However, such linear updating scheme would certainly fail to account for the observed nonlinearities in subjects' behavior (S2 Fig). Therefore, we constructed a linear-nonlinear estimator (LNE) that augmented the linear updating by a point nonlinearity that could account for the observed prior dependent biases in t_p (Fig 5A). The nonlinear function, $f_{BLS_1}(\bar{t}_n)$, was chosen to match the BLS estimator for a single measurement ($n = 1$), which is determined by w_m.

Simulation of LNE verified that it could indeed integrate multiple measurements and exhibit prior-dependent biases (not shown). However, the behavior of LNE was qualitatively different from BLS. The contrast between the two models was evident from a comparison of the relationship between measurements and estimates. Unlike BLS (Fig 3D), estimates derived from LNE are linear with respect to t_{m_1} and t_{m_2}, a feature that can be visualized by the linear iso-estimate contours of the LNE model (Fig 5B).

We fitted LNE to each subject independently and asked how well it accounted for the observed statistics. The LNE model broadly captured the observed regression to the mean (Fig 5C,D; see S4 Fig for fits to all subjects), but had a qualitative failure: fits exhibited significantly more BIAS in 1-2-Go condition (Fig 5D, inset; two tailed t-test, t(8) = 4.9304, p-value = 0.001) and significantly less BIAS in 1-2-3-Go condition (Fig 5D, inset; two tailed t-test, t(8) = -2.3782, p-value = 0.045) than the biases present in the data. This failure can be readily explained in terms of how LNE functions. Since the static nonlinearity in LNE is the same for one and two measurements, the bias LNE generates is the same for the 1-2-Go and 1-2-3-Go conditions. Therefore, when we fitted LNE to data from both conditions, the model consistently overestimated BIAS for the 1-2-Go condition, and underestimated BIAS for the 1-2-3-Go condition (Fig 5C, red and blue lines nearly overlap).

We further evaluated LNE by asking how it accounted for the observed performance improvement in the 1-2-3-Go condition compared to the 1-2-Go condition. We normalized each subject's RMSE from the 1-2-Go and 1-2-3-Go conditions to the RMSE expected from the behavior of the fitted LNE model in the 1-2-3-Go condition (Fig 5E). Most subjects surpassed the predictions of the LNE model (horizontal line) for the 1-2-3-Go condition, and average RMSE reached values that were significantly lower than expected (0.990; two tailed t-test, t(8) = -2.463, p-value = 0.039). Based on these results, we concluded that LNE fails to capture subjects' behavior both qualitatively and quantitatively.

An extended Kalman filter (EKF) model for approximate Bayesian inference

We considered a moderately more sophisticated algorithm inspired by the extended Kalman filter (EKF) [56]. This algorithm is shown in Fig 6A. Upon each new
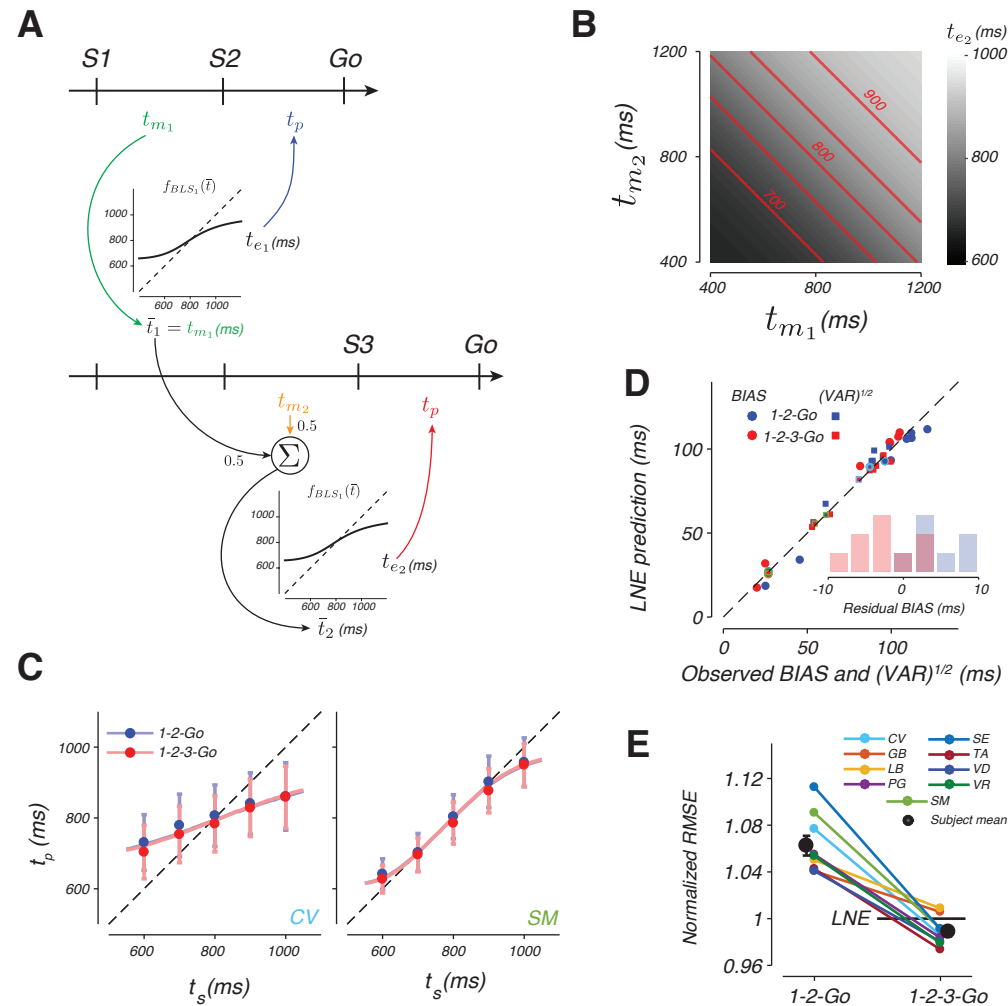
Fig 5. A linear-nonlinear estimator (LNE) model and its fits to the data. (A) LNE algorithm. LNE derives an estimate by applying a nonlinear function, f_{BLS_1}, to the average of the measurements. In the 1-2-Go trials (top), the average, $\bar{t}_1$, is the same as the first measurement, t_{m_1}, and the estimate, t_{e_1}, is $f_{BLS_1}(\bar{t}_1)$. In 1-2-3-Go trials (bottom), the average, $\bar{t}_2$, is updated by the second measurement, t_{m_2} ($\bar{t}_2 = 0.5(\bar{t}_1 + t_{m_2})$), and the estimate, t_{e_2}, is $f_{BLS_1}(\bar{t}_2)$. In both conditions, the produced interval, t_p, is perturbed by zero-mean Gaussian noise with standard deviation proportional to the final estimate (t_{e_1} for 1-2-Go and t_{e_2} for 1-2-3-Go) with the constant of proportionality w_p, as in the BLS model. (B) The mapping from measurements to estimates (grayscale) for the LNE estimator in the 1-2-3-Go trials. Red lines indicate combinations of measurements that lead to identical estimates (shown for t_{e_2} = 700, 750, 800, 850, and 900 ms). (C) Mean and standard deviation of t_p as a function of t_s for two example subjects (circles and error bars) along with the corresponding fits of the LNE model (lines). (D) BIAS (circles) and $\sqrt{\text{VAR}}$ (squares) of each subject (abscissa) and the corresponding values computed from simulations of the fitted LNE model (ordinate). Conventions match Fig 4B. (E) The RMSE in the 1-2-Go and 1-2-3-Go conditions relative to the corresponding predictions from the LNE model (conventions as in Fig 4C). See also S4 Fig.

measurement, EKF uses the error between the previous estimate and the current measurement to generate a new estimate. The difference between EKF and the Kalman filter is that the error is subjected to a nonlinear function before being used to update the previous estimate. This nonlinearity is necessary for the algorithm to be able to account for the nonlinear prior-dependent biases observed in behavior.

In our experiment, immediately after the first flash, the only information about the

sample interval, t_s, comes from the prior distribution. Accordingly, we set the initial estimate, t_{e_0}, to the mean of the prior distribution. After the first measurement, EKF computes an "innovation" term by applying a static nonlinearity, $f^\star(x)$ to the error, x_1 between t_{m_1} and t_{e_0}. This innovation is multiplied by a gain, k_1 and added to t_{e_0} to compute the new estimate, t_{e_1}. In the 1-2-Go condition in which only one measurement is available, t_{e_1} serves as the final estimate that the model aims to reproduce.

For the 1-2-3-Go condition, EKF repeats the updating procedure after the second measurement t_{m_2}. It computes the difference between t_{m_2} and t_{e_1} to derive a prediction error, x_2, which is subjected to the same nonlinear function, $f^\star(x)$, to yield a second innovation. This innovation is then scaled by an appropriate gain, k_2, and added to t_{e_1} to generate an updated estimate, t_{e_2}, which the model aims to reproduce.

The two important elements that determine the overall behavior of EKF are the nonlinear function $f^\star(x)$ and the gain factor(s) applied to the innovation(s) (k_1 and k_2) to update the estimate(s). We set the form of the nonlinear function $f^\star(x)$ such that biases in t_{e_1} after one measurement are the same between EKF and BLS models. This ensures that EKF and BLS behave identically in the 1-2-Go condition. Note that our implementation of EKF assumes that the same nonlinear function is applied after every measurement. If one allows this nonlinear function to be optimized separately for each measurement, EKF would be able to replicate the behavior of BLS exactly (S3 Fig).

For the gain factors, we reasoned that the most rational choice is to set the weight of each innovation based on the expected reliability of the corresponding estimate, $t_{e_{n-1}}$, relative to the new measurement, t_{m_n}, as in the Kalman filter (see Materials and methods). This causes the gain factor to decrease with the number of measurements, and ensures that the influence of each new measurement is appropriately titrated. With these assumptions, EKF remains suboptimal for the 1-2-3-Go condition. However, it captures certain aspects of the nonlinearities associated with the optimal BLS estimator as shown by Fig 6B (compare to Fig 3D).

The algorithm implemented by EKF is appealing as it uses a simple updating strategy that can be straightforwardly extended to multiple sequential measurements. Furthermore, EKF captures important features of human behavior. First, integration of each new measurement causes a reduction in RMSE, as seen in 1-2-3-Go compared to 1-2-Go condition. Second, the nonlinear function applied to innovations allows EKF to incorporate prior information and capture prior-dependent biases. Third, since the nonlinearity is applied to each innovation (as opposed to the final estimate), EKF, unlike LNE, is able to capture the reduction in BIAS in 1-2-3-Go compared to 1-2-Go condition.

We fitted EKF to each subject independently and asked how well it accounted for the observed statistics. Similar to BLS and LNE, EKF broadly captured the observed regression to the mean in the 1-2-Go trials (Fig 6C,D, blue). This is not surprising since the EKF algorithm is identical to BLS when the prior is integrated with a single measurement. EKF was also able to capture the mean t_p as a function of t_s in the 1-2-3-Go trials (Fig 6C,D red). Importantly, unlike BLS, there was no significant difference between the BIAS observed in subject behavior and the EKF model (Fig 6D, inset; two tailed t-test, t(8) = 4.6055, p-value = 0.02639).

We also asked if EKF could account for the observed RMSEs. To do so, we performed the same analysis we used to evaluate the BLS and LNE models. We normalized each subject's RMSE from the 1-2-Go and 1-2-3-Go conditions to the RMSE expected from the EKF model for 1-2-3-Go (Fig 6E). We found no significant difference between observed and predicted RMSEs for the 1-2-3-Go condition (two-tailed t-test, t(8) = 1.5506, p-value = 0.160), and no significant difference between the observed and predicted change in RMSE from the 1-2-Go to the 1-2-3-Go condition (one tailed t-test, t(8) = -1.1090, p-value = 0.150). These results indicate that subjects' suboptimal
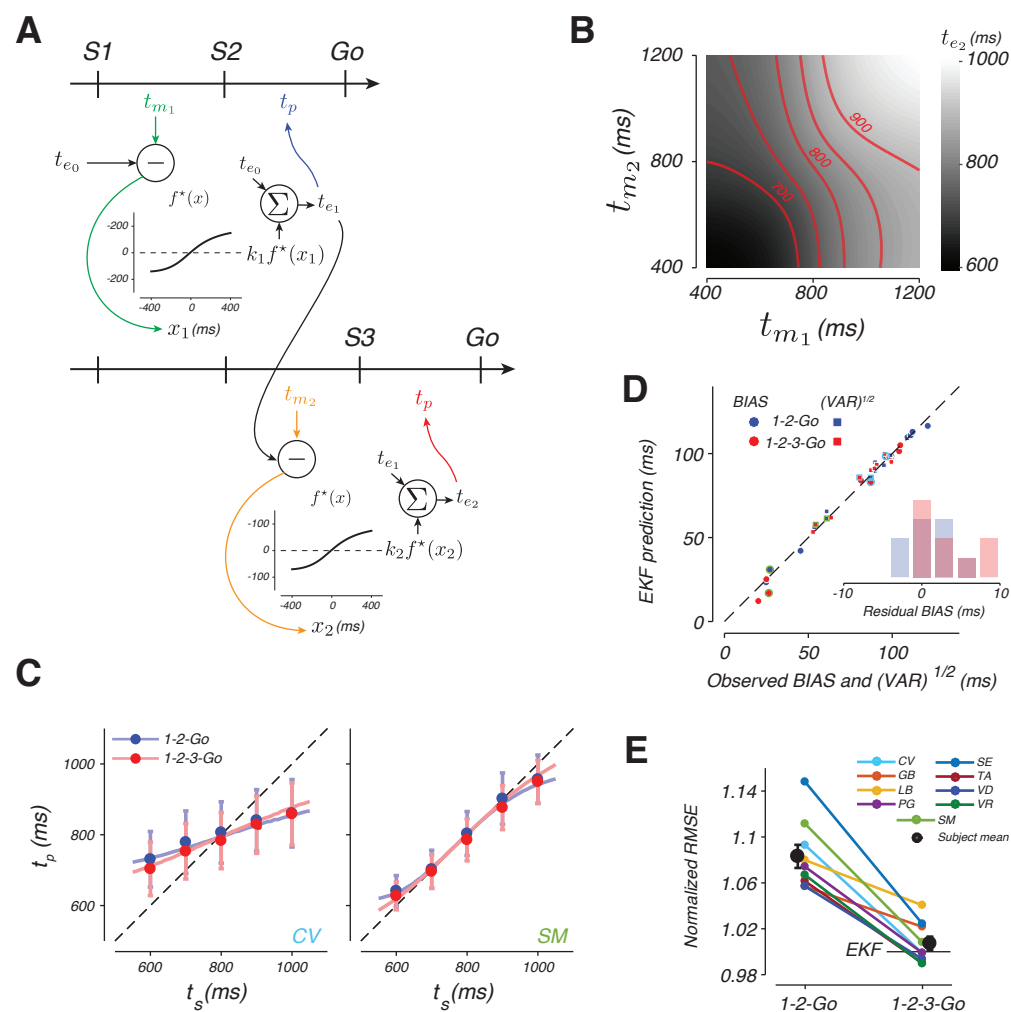
Fig 6. An extended Kalman filter (EKF) model and its fits to the data. (A) EKF algorithm. EKF is a real-time inference algorithm that uses each measurement to update the estimate. After the first flash, EKF uses the mean of the prior as its initial estimate, t_{e_0}. The second flash furnishes the first measurement, t_{m_1}. EKF computes a new estimate, t_{e_1}, using the following procedure: (1) it measures the difference between t_{m_1} and t_{e_0} to compute an error, x_1, (2) it applies a nonlinear function, $f^\star(x)$, to x_1, (3) it scales $f^\star(x_1)$ by a gain factor, k_1, whose magnitude depends on the relative reliability of t_{m_1} and t_{e_0}, and (4) it adds $k_1 f^\star(x_1)$ to t_{e_0} to compute t_{e_1}. In the 1-2-Go condition (top), t_{e_1} is the final estimate used for the production of t_p. In the 1-2-3-Go condition (bottom), the updating procedure is repeated to compute a new estimate t_{e_2} by adding t_{e_1} to $k_2 f^\star(x_2)$ where x_2 is the difference between the second measurement, t_{m_2}, and t_{e_1}, and k_2 is the scale factor determined by the relative reliability of t_{e_1} and t_{m_2}. t_{e_2} is then used as the final estimate for the production of t_p. We assumed that the produced interval, t_p, is perturbed by zero-mean Gaussian noise with standard deviation proportional to the final estimate (t_{e_1} for 1-2-Go and t_{e_2} for 1-2-3-Go) with the constant of proportionality w_p, as in the BLS model. (B) The mapping from measurements to estimates (grayscale) for the EKF estimator in the 1-2-3-Go condition. Red lines indicate combinations of measurements that lead to identical estimates (shown for $t_{e_2} = 700, 750, 800, 850,$ and 900 ms). (C) Mean and standard deviation of t_p as a function of t_s for two example subjects (circles and error bars) along with the corresponding fits of the EKF model (lines). (D) BIAS (circles) and $\sqrt{\text{VAR}}$ (squares) of each subject (abscissa) and the corresponding values computed from simulations of the fitted EKF model (ordinate). Conventions match Fig 4B. (E) The RMSE in the 1-2-Go and 1-2-3-Go conditions relative to the corresponding predictions from the EKF model (conventions as in Fig 4C). See also S5 Fig.

behavior is consistent with the approximate Bayesian integration implemented by the

EKF algorithm.

To further validate the superiority of the EKF model, we directly compared various models to BLS using log likelihood ratio. Specifically, we computed the ratio of the log likelihood of each model given t_s and t_p ($log\mathcal{L}(\mathcal{M}_i|t_s, t_p)$ see Materials and methods) to the log likelihood of the BLS model, $log\mathcal{L}(\mathcal{M}_{BLS}|t_s, t_p)$, for each subject. We found that EKF provided the best fit for 8 out of 9 subjects (Table 1). For one of the subjects, the fits were poor for all models but LNE provided the best fit.

Table 1. Predictive log likelihood ratio for each model and subject.

Model	Subject								
	CV	GB	LB	PG	SM	SE	TA	VR	VD
LNE	1.1678	-1.8267	-0.3094	0.1516	**0.4896**	-1.7151	-0.0365	0.3456	0.3793
	(0.5012)	(0.2976)	(0.2020)	(0.3407)	**(0.3427)**	(0.3493)	(0.2806)	(0.3593)	(0.2175)
EKF	**1.1900**	**0.0380**	**0.7168**	**0.3828**	-1.4262	**1.0909**	**0.2358**	**0.3623**	**0.6537**
	(0.1998)	**(0.1326)**	**(0.2234)**	**(0.1428)**	(0.2755)	**(0.1543)**	**(0.1250)**	**(0.1581)**	**(0.1002)**
BLS$_{mem}$	0.4621	-0.0821	-0.2445	0.0086	-0.0637	0.0440	-0.0191	-0.0359	-0.0421
	(0.1832)	(0.0267)	(0.0743)	(0.0518)	(0.2714)	(0.0876)	(0.0195)	(0.0249)	(0.0268)

The predictive log likelihood of each model was calculated as the likelihood of the model given 100 trials of validation data. The model was fit to the remaining data not used for validation. This was iterated until all the data was used for validation (e.g. leave 100 out cross validation). Each entry is the mean of the predictive log likelihood ratio (standard error) between a given model and the BLS model (i.e. $log[\mathcal{L}(\mathcal{M}_i|t_s, t_p)/\mathcal{L}(\mathcal{M}_{BLS}|t_s, t_p)]$,. For each subject, the model with the highest log likelihood is shown in bold. See also S6 Fig.

Discussion

The neural systems implementing sensorimotor transformations must rapidly compute state estimates to effectively implement online control of behavior. Behavioral studies indicate that, at a computational level, state estimation may be described in terms of Bayesian integration [6–15]. However, describing behavior with a Bayesian model does not necessarily indicate that the brain implements these computations by representing probability distributions [21, 22, 57]. Here, we focused on integration of multiple time intervals and found evidence that the brain relies on simpler algorithms that approximate optimal Bayesian inference.

We demonstrated that humans integrate prior knowledge with one or two measurements to improve their performance. A key observation was that the integration was nearly optimal for one measurement but not for two measurements. In particular, when two measurements were provided, subjects systematically exhibited more BIAS toward the mean of the prior than expected from an optimal Bayesian model. This observation motivated us to investigate various algorithms that could lead to similar patterns of behavior.

Analytical and numerical analyses suggested that simple inference algorithms that update certain parameters of the posterior instead of the full distribution can not integrate multiple measurements optimally when the noise is signal-dependent. We then systematically explored simple inference algorithms that could perform sequential updating and account for the behavioral observations. One of the simplest updating algorithms is the Kalman filter [58]. However, this algorithm updates estimates linearly and thus cannot account for the nonlinearities in subjects' behavior, even for a single measurement (S2 Fig). The LNE model augmented the Kalman filter such that the final estimate was subjected to a point-nonlinearity. This allowed LNE to generate nonlinear biases but since the nonlinearity was applied to the final estimate, LNE failed to capture the decrease in bias observed in the 1-2-3-Go compared to 1-2-Go condition.

Finally, we developed EKF, which is a more sophisticated variant of the Kalman filter that applies a static nonlinearity to the errors in estimation at every stage of updating. This algorithm accounted for optimal behavior in the 1-2-Go condition and exhibited the same patterns of suboptimality observed in humans in the 1-2-3-Go condition. Therefore, EKF provides a good characterization of the algorithm brain uses when there is need to integrate multiple pieces of information presented sequentially. This finding implies that subjects may only rely on the first few moments of a distribution and use nonlinear updating strategy to track those instead of updating the entire posterior. This strategy is simple and in many scenarios could lead to optimal behavior with little computational cost. Moreover, the recursive nature of EKF's updating strategy allows it to readily generalize to scenarios when it is necessary to update estimates in real time, even when the number of available samples is not known *a priori*, which extensions of our experiment could test.

Maintaining and updating probability distributions is computationally expensive. Moreover, it is not currently known how neural networks might implement such operations [22]. In contrast, EKF is relatively simple to implement. The only requirement is to use the current estimate to predict the next sample, and use a nonlinear function of the error in prediction to update estimates sequentially. Predictive mechanisms that EKF relies on are thought to be an integral part of how brain circuits support perception and sensorimotor function [1, 5, 8, 59–62]. As such, the relative success of EKF may be in part due to its compatibility with predictive coding mechanisms that the brain uses to perform sequential updating. This observation makes the following intriguing prediction: when performing 1-2-3-Go task, subjects do not make two measurements; instead, they use the prior to predict the time of the second flash, use the prediction error to update their estimate, and the new estimate to predict the third flash. In other words, according to the EKF model, the underlying neural signals encode intervals prospectively, which is consistent with electrophysiological experiments in nonhuman primates [63, 64] and rodents [?, ?, ?].

While EKF provides a better account of the observed data in our experiments, it may be that our specific formulation of the Bayesian model did not capture the underlying process. Our BLS model was based on three assumptions: (1) that likelihood function is characterized by signal dependent noise, (2) that the subjective prior matches the experimentally imposed uniform prior distribution, and (3) that the final estimates are derived from the mean of the posterior, which implicitly assumes that subject rely on a quadratic cost function, as was previously demonstrated [24]. Our formulation of the likelihood function is particularly important, as it is the key factor that prohibits simple algorithms such as EKF to optimally integrate multiple measurements. The inherent signal-dependent noise in timing causes the likelihood function to be skewed toward longer intervals (see S1 Appendix). This characteristic feature was particularly important for explaining human behavior in a task requiring interval estimation following several measurements [52]. Moreover, it has been shown that subjects exhibit larger biases for longer intervals within the domain of the prior indicating that the brain has an internal model of this signal-dependent noise [24, 36, 65, 66]. These results support our formulation of the likelihood function. However, one aspect of our formulation that deserves further scrutiny is the assumption that noise perturbing the two measurements was independent. This seems unlikely given the long-range positive autocorrelations in behavioral variability [37, 67, 68], and because S2 is shared between the two measurements, which may lead to correlations between t_{m_1} and t_{m_2}.

Our formulation of the prior and cost function should also be further evaluated. For example, humans may not be able to correctly encode a uniform prior probability distribution for interval estimation [34, 36]. Similarly, the cost function may not be quadratic [69]. However, since priors and the cost functions impact both 1-2-Go and

1-2-3-Go conditions, moderate inaccuracies in modeling these components may not be able to explain optimal behavior in 1-2-Go and suboptimal behavior in the 1-2-3-Go condition simultaneously. Finally, recent results suggest the performance may be limited by imperfect integration [70, 71] and imperfect memory [48, 72], which future models of sequential updating should incorporate.

Materials and methods

Subjects and apparatus

All experiments were performed with the approval of the Committee on the Use of Humans as Experimental Subjects at MIT after receiving informed consent. Eleven human subjects (6 male and 5 female) between 18 and 33 years of age participated in the interval reproduction experiment. Of the 11 subjects, 10 were naive to the purpose of the study.

Subjects sat in a dark, quiet room at a distance of approximately 50 cm from a display monitor. The display monitor had a refresh rate of 60 Hz, a resolution of 1920 by 1200, and was controlled by a custom software (MWorks; http://mworks-project.org/) on an Apple Macintosh platform.

Interval reproduction task

Experiment consisted of several 1 hour sessions in which subjects performed an interval reproduction task (Fig 1). The task consisted of two randomly interleaved trial types − "1-2-Go" and "1-2-3-Go". On 1-2-Go trials, two flashes (S1 followed by S2) demarcated a sample interval (t_s) that subjects had to measure [24]. On 1-2-3-Go trials, t_s was presented twice, once demarcated by S1 and S2 flashes, and once by S2 and S3 flashes. For both trial types, subjects had to reproduce t_s immediately after the last flash (S2 for 1-2-Go and S3 for 1-2-3-Go) by pressing a button on a standard Apple keyboard. On all trials, subjects had to initiate their response proactively and without any additional cue (no explicit Go cue was presented). Subjects received graded feedback on their accuracy.

Each trial began with the presentation of a 0.5 deg circular fixation point at the center of a monitor display. The color of the fixation was blue or red for the 1-2-Go and 1-2-3-Go trials, respectively. Subjects were asked to shift their gaze to the fixation point and maintain fixation throughout the trial. Eye movements were not monitored. After a random delay with a uniform hazard (100 ms minimum plus and interval drawn from an exponential distribution with a mean of 300 ms), a warning stimulus and a trial cue were presented. The warning stimulus was a white circle that subtended 1.5 deg and was presented 10 deg to the left of the fixation point. The trial cue consisted of 2 or 3 small rectangles 0.6 deg above the fixation point (subtending 0.2 x 0.4 deg, 0.5 deg apart) for the 1-2-Go and 1-2-3-Go trials, respectively. After a random delay with a uniform hazard (250 ms minimum plus an interval drawn from an exponential distribution with mean of 500 ms), flashes demarcating t_s were presented. Each flash (S1 and S2 for 1-2-Go, and S1, S2 and S3 for 1-2-3-Go) lasted for 6 frames (100 ms) and was presented as an annulus around the fixation point with an inside and outside diameter of 2.5 and 3 deg, respectively (Fig 1A,B). The time between consecutive flashes, which determined t_s, was sampled from a discrete uniform distribution ranging between 600 and 1000 ms with a 5 samples (Fig 1C). To help subjects track the progression of events throughout the trial, after each flash, one rectangle from the trial cue would disappear (starting from the rightmost).

Produced interval (t_p) was measured as the interval between the time of the last flash and the time when the subject pressed a designated key on the keyboard (Fig 1A,B).

Subjects received trial-by-trial visual feedback based on the magnitude and sign of the relative error, $(t_p - t_s)/t_s$. A 0.5 deg circle ("analog feedback") was presented to the right (for error < 0) or left (error > 0) of the the warning stimulus at a distance that scaled with the magnitude of the error. Additionally, when the error was smaller than a threshold, both the warning stimulus and the analog feedback turned green and a tone denoting "correct" was presented. If the production error was larger than the threshold, the warning stimulus and analog feedback remained white and a tone denoting "incorrect" was presented. The error threshold was scaled with the sample interval to accommodate the scalar variability of timing that leads to more variable production intervals for longer sample intervals (Fig 1D). The scaling factor was initialized at 0.15 at the start of every session and adjusted adaptively using a one-up, one-down scheme that added or subtracted 0.001 to the scaling factor for incorrect or correct responses, respectively. These manipulations ensured that the performance across conditions, subjects, and trials remained approximately at a steady state of 50% correct trials.

To ensure subjects understood the task design, the first session included a number of training blocks. Training blocks were conducted with the supervision of an experimenter. Training trials were arranged in 25 trial blocks. In the first block, the subjects performed the 1-2-Go condition with the sample interval fixed at 600 ms. In the second block, we fixed the interval to be 1000 ms. In the third block, the subject performed the 1-2-3-Go task with the interval fixed at 1000 ms. In the fourth block, the subject continued to perform the 1-2-3-Go task, but with the intervals chosen at random from the experimental distribution. In the final training block, the task condition and sample intervals were fully randomized, as in the main experiment. The subject then performed 400 trials of the main experiment. Subjects completed 10 sessions total, performing 800 trials in each of the remaining 9 experimental sessions. To ensure subjects were adapted to the statistics of the prior [24], we discarded the first 99 trials of each session. We also discarded any trial when the subject responded before S2 (for 1-2-Go) or S3 (for 1-2-3-Go) or 1000 ms after the veridical t_s. S1 Table summarizes the number of completed trials for each subject. Data from two subjects were not included in the analyses because they were not sensitive to the range of sample intervals we tested and their production interval distributions were not significantly different for the longest and shortest sample intervals.

Models

We considered several models for the interval estimation: (1) an optimal Bayes Least-Squares model (BLS), (2) an optimal Bayes Least-Squares model that allowed different noise levels for the two measurements in 1-2-3-Go trials, (3) an extended Kalman filter model (EKF), and (4) a linear-nonlinear estimation model (LNE). All models were designed to be identical for the 1-2-Go task where only one measurement was available but differed in their prediction for the 1-2-3-Go trials.

BLS model

We used the Bayesian integration model that was previously shown to capture behavior in the 1-2-Go task [24]. This model assumes that subjects combine the measurements and the prior distribution probabilistically according to the Bayes' rule:

$$p(t_s|t_m) = \frac{\lambda(t_m|t_s)p(t_s)}{p(t_m)}, \tag{1}$$

$$\lambda(t_m|t_s) = \frac{1}{\sqrt{2\pi w_m^2 t_s^2}} e^{-\frac{(t_m - t_s)^2}{2w_m^2 t_s^2}}, \tag{2}$$

where $p(t_s)$ represents the prior distribution of the sample intervals and $p(t_m)$ the probability distribution of the measurements. The likelihood function, $\lambda(t_m|t_s)$, was formulated based on the assumption that measurement noise was Gaussian and had zero mean. To incorporate scalar variability into our model, we further assumed that the the standard deviation of noise scales with t_s with constant of proportionality w_m representing the Weber fraction for measurement.

Following previous work [24], we further assumed that subjects' behavior can be described by a BLS estimator that minimizes the expected squared error, and uses the expected value of the posterior distribution as the optimal estimate:

$$t_{e_1} = f_{BLS_1}(t_{m_1}) = E[t_s|t_{m_1}], \tag{3}$$

where f_{BLS_1} denotes the BLS function that maps the measurement (t_{m_1}) to the Bayesian estimate after one measurement (t_{e_1}). The subscript 1 is added to clarify that this equation corresponds to the condition with a single measurement (i.e., 1-2-Go). The notation $E[\bullet]$ denotes expected value. Given a uniform prior distribution with a range from $t_s^{\min}$ to $t_s^{\max}$, the BLS estimator can be written as:

$$f_{BLS_1}(t_{m_1}) = \frac{\int_{t_s^{\min}}^{t_s^{\max}} t_s \lambda(t_{m_1}|t_s)dt_s}{\int_{t_s^{\min}}^{t_s^{\max}} \lambda(t_{m_1}|t_s)dt_s}, \tag{4}$$

We assume that $t_s^{\min}$ and $t_s^{\max}$ match the minimum and maximum of the experimentally imposed sample interval distribution. We extended this model for the 1-2-3-Go task to two measurements. To do so, we incorporated two likelihood functions in the derivation of the posterior. Assuming that the two measurements are conditionally independent, the posterior can be written as:

$$p(t_s|t_{m_1}, t_{m_2}) = \frac{\lambda(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)p(t_s)}{\int \lambda(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)p(t_s)dt_s}, \tag{5}$$

where t_{m_1} and t_{m_2} denote the first and second measurements, respectively, and the likelihood function, λ, is from Eq (2). Because measurements are taken in a sequence, we can rewrite Eq (5) in a recursive form

$$p(t_s|t_{m_1}, t_{m_2}) = \frac{1}{Z}\lambda(t_{m_2}|t_s)p(t_s|t_{m_1}), \tag{6}$$

Where $p(t_s|t_{m_1})$ the posterior as specified in Eq (1) and Z is a normalization factor equal to the denominator of Eq (5). Although the posterior for Eqs (5) and (6) are identical, specifying the posterior in this way allows for the algorithm to be updated following each measurement.

The corresponding BLS estimator can again be written as the expected value of the posterior:

$$t_{e_2} = f_{BLS_2}(t_{m_1}, t_{m_2}) = E[t_s|t_{m_1}, t_{m_2}], \tag{7}$$

$$f_{BLS_2}(t_{m_1}, t_{m_2}) = \frac{\int_{t_s^{\min}}^{t_s^{\max}} t_s \lambda(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)dt_s}{\int_{t_s^{\min}}^{t_s^{\max}} \lambda(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)dt_s}, \tag{8}$$

where f_{BLS_2} denotes the BLS function that uses two measurements (t_{m_1} and t_{m_2} to compute t_{e_2}. The subscript 2 represents the optimal mapping function for two measurements (i.e., 1-2-3-Go). We performed the integrations for the BLS model numerically using Simpson's quadrature.

BLS$_{\mathrm{mem}}$ model

We also considered the possibility that the brain may not be able to hold representations of the first measurement or the associated posterior perfectly over time until the time for integration. To model this we assumed two Weber fractions $- w_m$ as formulated in the BLS model and w_{mem} which adjusts the Weber fraction of the first measurement in 1-2-3-Go trials to account for noisy memory or inference processes. In 1-2-Go trials, the posterior was set according to Eq (1) with w_m controlling the signal dependent noise. In 1-2-3-Go trials, the posterior was set according to

$$p(t_s|t_{m_1}, t_{m_2}) = \frac{\lambda_{mem}(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)p(t_s)}{\int \lambda_{mem}(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)p(t_s)dt_s}, \tag{9}$$

With the likelihood function associated with the first measurement defined as

$$\lambda_{mem}(t_m|t_s) = \frac{1}{\sqrt{2\pi w_{mem}^2 t_s^2}} e^{-\frac{(t_m - t_s)^2}{2w_{mem}^2 t_s^2}}, \tag{10}$$

This formulation allows the measurement noise to be different for the two measurements. The optimal estimator was then calculated as

$$f_{BLS_2}(t_{m_1}, t_{m_2}) = \frac{\int_{t_s^{\min}}^{t_s^{\max}} t_s \lambda_{mem}(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)dt_s}{\int_{t_s^{\min}}^{t_s^{\max}} \lambda_{mem}(t_{m_1}|t_s)\lambda(t_{m_2}|t_s)dt_s}, \tag{11}$$

EKF model

EKF implements an updating algorithm in which after each flash, the observer updates the estimate, t_{e_n}, based on the previous estimate, $t_{e_{n-1}}$, and the current measurement, t_{m_n}. The updating rule changes $t_{e_{n-1}}$ by a nonlinear function of the error between $t_{e_{n-1}}$ and t_{m_n}, which we denote by x_n.

$$t_{e_n} = t_{e_{n-1}} + k_n f^\star(x_n), \tag{12}$$

$$x_n = t_{m_n} - t_{e_{n-1}}, \tag{13}$$

$f^\star$ is a nonlinear function based on the BLS estimator, f_{BLS_1}:

$$f^\star(x_n) = f_{\mathrm{BLS}_1}(x_n + t_{e_0}) - t_{e_0}, \tag{14}$$

k_n is a gain factor the controls the magnitude of the update and is set by the relative reliability of $t_{e_{n-1}}$ and t_{m_n}, which were formulated in terms of two Weber fractions, w_{n-1} and w_m, respectively:

$$k_n = \frac{w_{n-1}^2}{w_{n-1}^2 + w_m^2}, \tag{15}$$

To track the reliability of t_{e_n} we used a formulation based on optimal cue combination under Gaussian noise. For Gaussian likelihoods, the reliability of the estimate is related to the inverse of the variance of the posterior. Similarly, the reliability of the interval estimate is related to the inverse of the Weber fraction. Therefore, we used the following algorithm to track the Weber fraction of the estimate, w_n:

$$w_n = \frac{w_{n-1} w_m}{\sqrt{w_{n-1}^2 + w_m^2}}, \tag{16}$$

As in the case of Gaussians, this algorithm ensures that w_n decreases with each additional measurement, reflecting the increased reliability of the estimate relative to the measurement. This ensures that the weight of the innovation respects information already integrated into the estimate by previous iterations of the EKF algorithm.

At S1, no measurements are available. Therefore, we set the initial estimate, t_{e_0} to the mean of the prior, and its reliability, w_0, to ∞. After S2 (one measurement), the EKF estimate is identical to the BLS model. For two measurements, the process is repeated to compute t_{e_2}, but the estimate is suboptimal. This formulation can be readily extended to more than two measurements.

LNE model

LNE uses a linear updating strategy similar to a Kalman filter to update estimates by measurements as follows:

$$\bar{t}_n = (1 - k_n)\bar{t}_{n-1} + k_n t_{m_n}, \tag{17}$$

The algorithm is initialized such that $\bar{t}_1 = t_{m_1}$ and we chose the weighting to be $k_n = 1/n$. This choice minimizes the squared errors in 1-2-3-Go trials. Note that any other choice for k_n would deteriorate LNE's performance. Following this sequential and linear updating scheme, LNE passes the final estimate through a nonlinear transfer function specified by the BLS model for one measurement (f_{BLS_1}):

$$f_{LNE_n}(t_{m_1}, ..., t_{m_n}) = f_{BLS_1}(\bar{t}_n), \tag{18}$$

where f_{LNE_n} denotes the linear-nonlinear estimator after n measurements. This formulation ensures that LNE is identical to the BLS in 1-2-Go trials.

Interval production model

In all models, the final estimate is used for the production phase. Following previous work [24], we assumed that the production of an interval is perturbed by Gaussian noise whose standard deviation scales with the estimated interval. The model was additionally augmented by an offset term to account for stimulus-independent biases observed in responses:

$$p(t_p|t_e) = \frac{1}{\sqrt{2\pi w_p^2 t_e^2}} e^{-\frac{(t_p - t_e - b)^2}{2w_p^2 t_e^2}}, \tag{19}$$

where w_p is the Weber fraction for production, b is the offset term, and t_e can refer to the estimate for either 1-2-Go and 1-2-3-Go trials.

All models accommodated "lapse trials" in which the produced interval was outside the mass of the production interval distribution. The lapse trials were modeled as trials in which the production interval was sampled from a fixed uniform distribution, $p(t_p|\text{lapse})$, independent of t_s. With this modification, the production interval distribution can be written as:

$$p(t_p|t_e, \gamma) = (1 - \gamma)p(t_p|t_e) + \gamma p(t_p|\text{lapse}), \tag{20}$$

where γ represents the lapse rate. With this formulation, we could identify lapse trials as those for which the likelihood of lapse exceeded the likelihood of a nonlapse. To limit cases of falsely identified lapse trials, we set the width of this uniform distribution conservatively to the range of possible production intervals (between 0 and 2000 ms).

Using simulations, we verified that our model was able to detect lapses for the range of w_m, w_p, and γ values inferred from the behavior of our subject pool. Most subjects had a small probability of a lapse trial that was consistent with previous reports [24]. Two subjects had relatively unreliable performance with a larger number of lapse trials. However, our conclusions do not depend on the inclusion of these two subjects.

Analysis and model fitting

All analyses were performed using MATLAB R2014b or MATLAB R2017a, The MathWorks, Inc., Natick, Massachusetts, United States. We used a predictive maximum likelihood procedure to fit each model to the data. Assuming that production intervals are conditionally independent across trials, the log likelihood of model parameters can be formulated as:

$$\log p(t_p^1, t_p^2, ..., t_p^N|t_s, w_m, w_p, b, \gamma) = \sum_{i=1}^{N} \log p(t_p^i|t_s, w_m, w_p, b, \gamma), \tag{21}$$

where the superscripts denote trial number. Maximum likelihood fits were derived from N-100 trials and cross validated on the remaining 100 trials. This process was performed iteratively until all the data was fit. The final model parameters were taken as the average of parameter values across all the fits to the data. Fits were robust to changes in the amount of left out data. See S1, S4, and S5 Fig for a summary of maximum likelihood parameters and predictions of each model fit to our subjects.

We evaluated model fits by generating simulated data from that model and comparing various summary statistics (BIAS, $\sqrt{\text{VAR}}$, and RMSE) observed for each subject to those generated by model simulations. For the observed data, summary

statistics were computed for non-lapse trials and after removing the offset (b). Model simulations were performed without the lapse term and after setting the offset to zero. The summary statistics were computed as follows:

$$\text{BIAS}^2 = \frac{1}{N} \sum_{i=1}^{N} (\bar{t}_{p_i} - t_{s_i})^2, \tag{22}$$

$$\text{VAR} = \frac{1}{N} \sum_{i=1}^{N} \sigma_i^2, \tag{23}$$

$$\text{RMSE} = \sqrt{\text{BIAS}^2 + \text{VAR}}, \tag{24}$$

BIAS^2 and VAR represent the average squared bias and average variance over the N distinct t_s's of the prior distribution. The terms $\bar{t}_{p_i}$, σ_i^2 represent the mean and variance of production intervals for the i-th sample interval (t_{s_i}). The overall RMSE was computed as the square root of the sum of BIAS^2 and VAR. To find the BIAS^2 and VAR of each model we took the mean value of each after 1000 simulations of the model with the trial number matched to each subject. This ensured an accurate estimate of these quantities that includes the systematic deviations from the true model behavior due to a finite number of trials.

To perform model comparison, we measured the likelihood of each model, given the maximum likelihood model parameters and the data, $\mathcal{L}(\mathcal{M}_i|t_s, t_p)$. We then computed the ratio $\mathcal{L}(\mathcal{M}_i|t_s, t_p)$ and $\mathcal{L}(\mathcal{M}_{BLS}|t_s, t_p)$, the likelihood of the BLS model, and computed the logarithm of that value to measure the log likelihood ratio. To generate confidence intervals, we evaluated the likelihoods using 100 trials of test data that were left out of model fitting. We iterated this process until all the data was used as training data, allowing us to measure the variability of the log likelihood ratio for each subject.
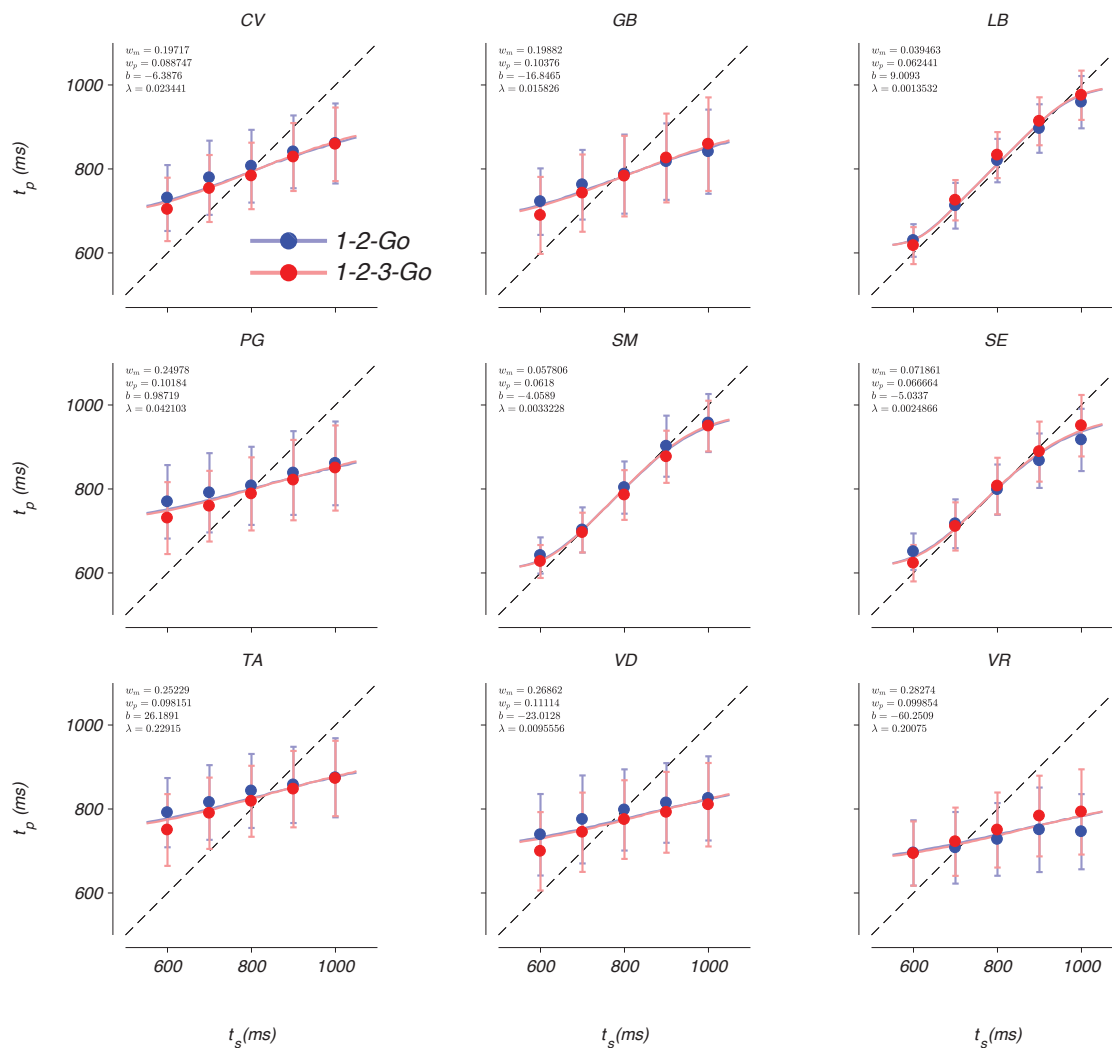
Supporting information
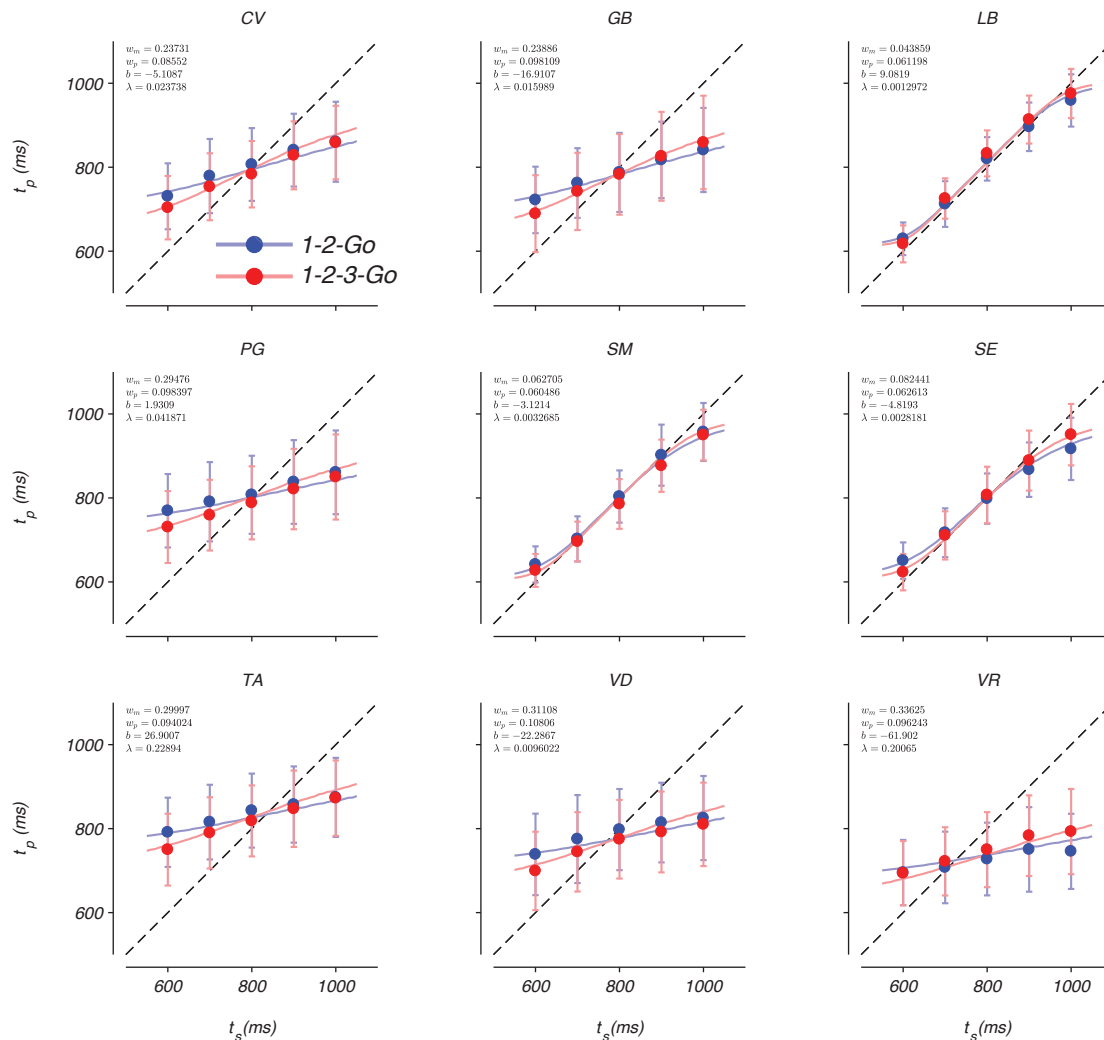
S1 Fig.   Fit of the BLS model to each subject. Mean and standard deviation (circles and error bars) of the produced interval, t_p, as a function of the sample interval, t_s, for each subject in 1-2-Go (blue) and 1-2-3-Go (red) trials. The dashed line indicates unity and the solid lines depict the BLS model for each subject. Insets: maximum likelihood estimate of model parameters for each subject.
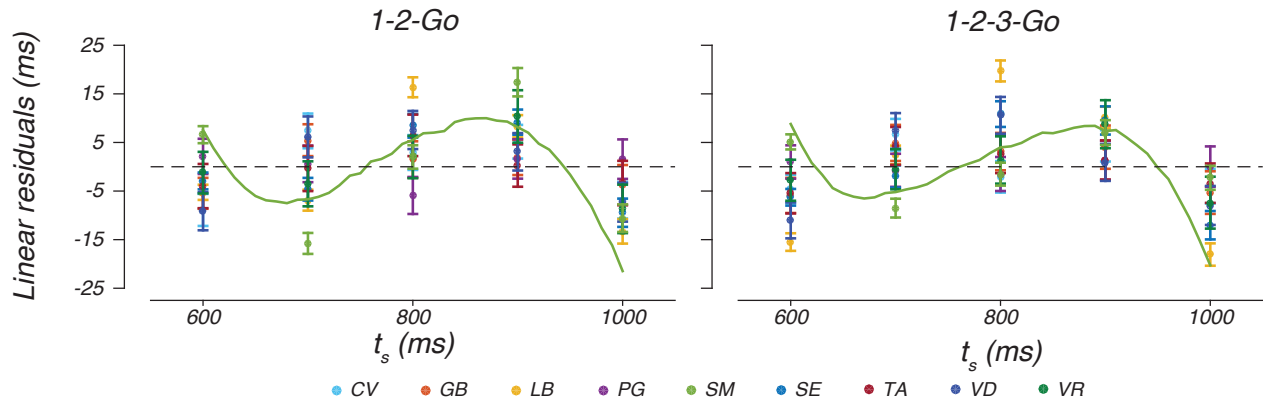
S2 Fig. Nonlinear processing of intervals. Mean residuals of the best fitting linear regression model ($\hat{t}_p = \beta t_s + \gamma$) relating production interval (t_p) to sample interval (t_s) for all subjects (different colors). Each data point represents the mean residual (e.g. $\frac{1}{M}\sum_{i=1}^{M}(t_{p_i} - \hat{t}_p)$, with i indexing trials). The error bars represent the standard error of the residuals. Residuals should be near zero (dashed line) if t_p was linearly related to t_s. The solid green line depicts the expected residuals from the BLS model fits to data from one subject (SM). The predictions for other subjects are qualitatively similar, but the exact shape depends on the measurement noise for each subject. The deviation of the data from the linear model indicates that subjects utilized a nonlinear transfer function to estimate t_s. See also Section 2.4.
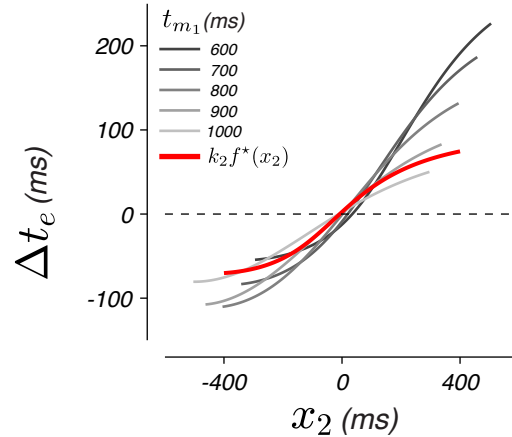
S3 Fig. The update function which achieves the BLS after two measurements. The BLS estimator for the 1-2-3-Go task uses the mean of the posterior, t_{e_2} after updating the posterior with the likelihood function associated with the second measurement, t_{m_2}. We characterized the function of the BLS estimator in terms of a linear updating term Δt_e. This term quantified the magnitude by which t_{e_1} has to be updated after the second measurement. A plot of Δt_e as a function of the error between the first estimate and the second measurement ($x_2 = t_{m_2} - t_{e_1}$) reveals two notables features. First the updating is a nonlinear function of x_2, and second, the form of the nonlinearity depends on t_{m_1} (different grayscales). In other words, BLS can be implemented with a linear updating scheme based on errors only if the update is derived from a nonlinear function whose form of nonlinearity varies with t_{m_1}. The red curve corresponds to the nonlinear updating scheme implemented by the EKF model, $k_2 f^\star(x_2)$. In this case, the updating relies on a nonlinear function of error but the form of the nonlinearity does not vary with t_{m_1}. See also Sections 2.3 and 2.5.
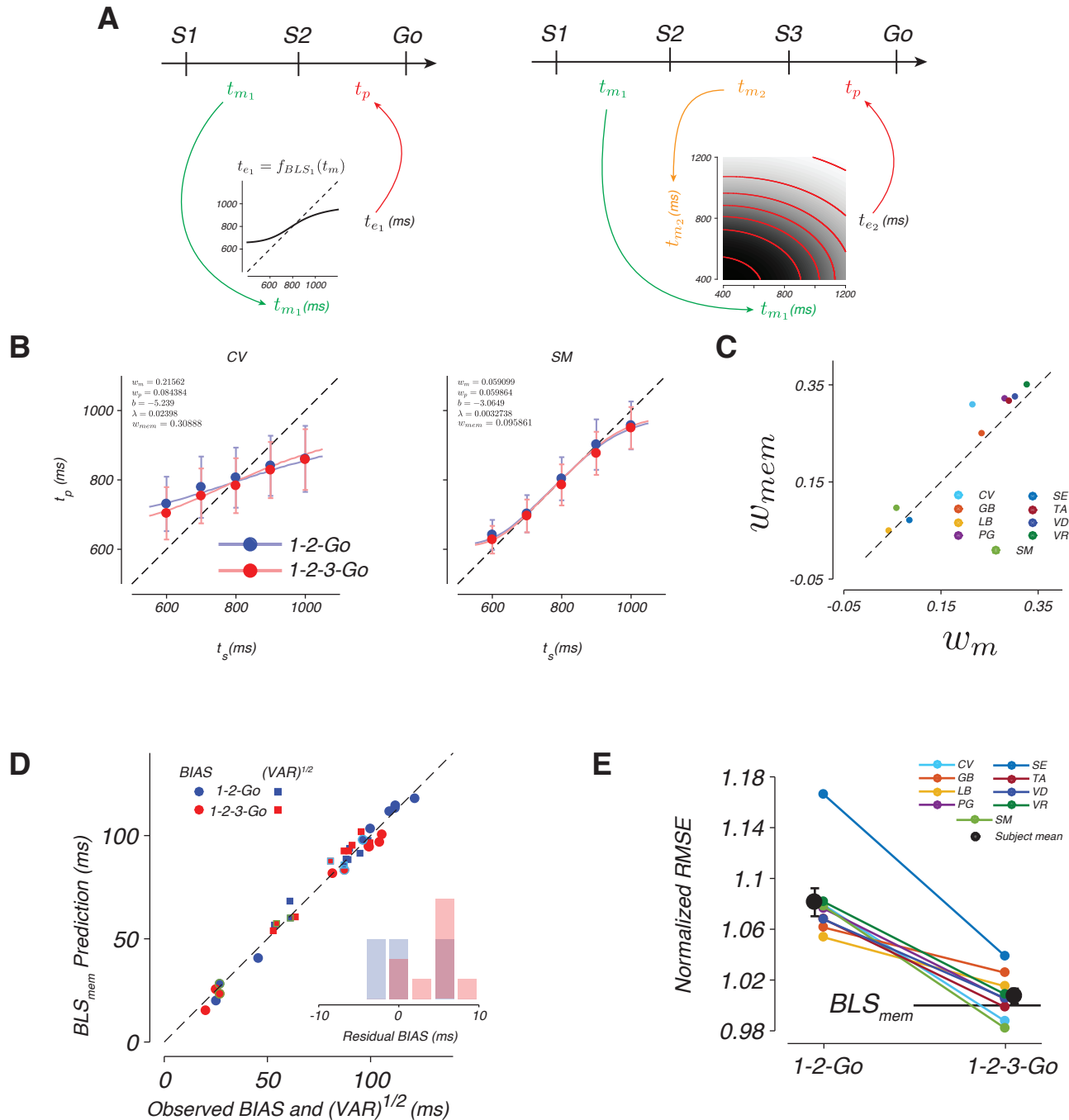
S4 Fig. Fit of the LNE model to each subject. Conventions are the same as Supporting information but for the LNE model.
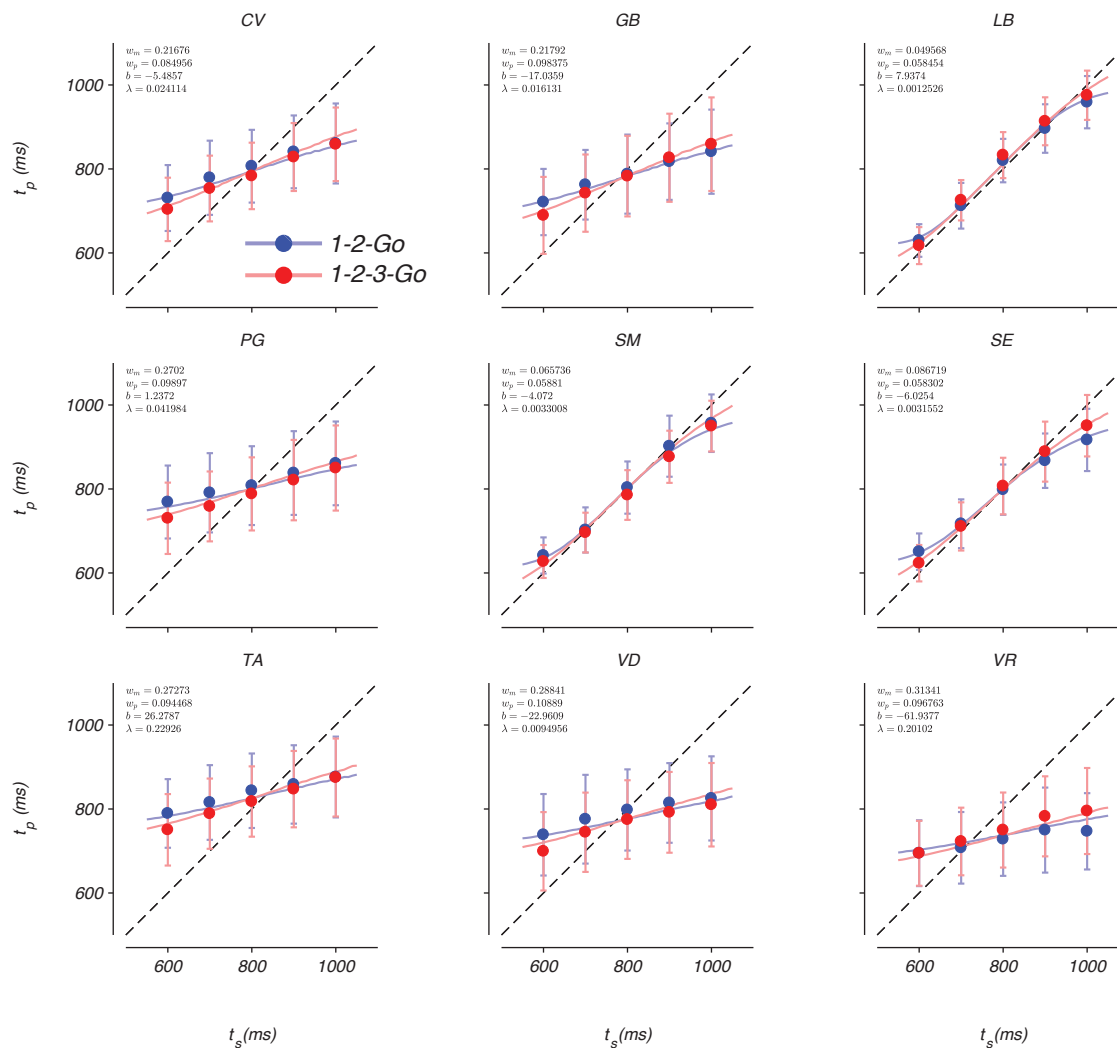
S5 Fig.   Fit of the EKF model to each subject. Conventions are the same as Supporting information but for the EKF model.

S6 Fig. BLS with different levels of noise for the two measurements (BLS_mem). (A) Left: For the 1-2-Go condition, BLS_{mem} is identical to BLS. Right: For the 1-2-3-Go condition, BLS_{mem} assumes that the Weber fraction for the measurement immediately before production, w_m is the same as the 1-2-Go condition, but the Weber fraction of the first measurement, denoted by w_{mem}, could be different from w_m. Therefore, the BLS_{mem} model has one more parameter than the BLS model. The bottom grayscale shows the estimate, t_{e_2}, as a function of t_{m_1} and t_{m_2} for $w_{mem} = 0.225$ and $w_m = 0.15$. Comparison of the iso-estimate contours (red lines) to the BLS mapping function (Fig 3D), indicates an increased reliance on t_{m_2}. (B) Predictions of the BLS_{mem} model (lines) and mean t_p +/- the standard deviation of example subjects (circles and error bars) as a function of t_s. Conventions are the same as panels shown in Supporting information. (C) Maximum

likelihood w_{mem} plot against w_m fit to each subject. Dashed line plots unity. In only one subject w_m exceeded w_{mem}. (D) Subject BIAS and $\sqrt{\text{VAR}}$ plotted against the predictions of the BLS_{mem} model fit to each subject. Inset: difference between the expected BIAS based on the BLS_{mem} model and that observed. (E) Each subject's RMSE in the 1-2-Go and 1-2-3-Go conditions to normalized by RMSE expected from the BLS_{mem} model in the 1-2-3-Go condition. See also Table 1.

Total trials	Trial type	Subject								
		CV	GB	LB	PG	SM	SE	TA	VD	VR
	1-2-Go (lapse)	3096 (68)	3156 (36)	3138 (5)	3130 (117)	3120 (10)	3143 (11)	2565 (587)	3099 (18)	2103 (371)
	1-2-3-Go (lapse)	3132 (38)	3120 (32)	3169 (1)	3084 (72)	3067 (5)	3159 (3)	2775 (415)	3182 (16)	2108 (316)

S1 Table. Trials completed by condition and subject in the interval reproduction task. Summary of the number of trials for the interval reproduction task in each condition and the number of trials identified as lapses in parentheses.

S1 Appendix

We modeled the likelihood of a sample, s, after making a measurement, m_i, perturbed by scalar noise as follows:

$$p(m_i|s) = \frac{1}{\sqrt{2\pi w_m^2 s^2}} e^{-\frac{(m_i-s)^2}{2w_m^2 s^2}}, \tag{1}$$

where the parameter w_m scales the standard deviation of the noise with the magnitude of the signal. For N conditionally independent measurements with the same w_m, the likelihood can be written as follows:

$$p(m_1, m_2, ..., m_N|s) = \left(\frac{1}{\sqrt{2\pi w_m^2 s^2}}\right)^N e^{-\frac{\sum_{i=1}^{N}(m_i-s)^2}{2w_m^2 s^2}}, \tag{2}$$

In Eq (2), because the term in the parentheses, which depends on s, is exponentiated to the power of N, we can conclude that the form of the likelihood function is not invariant with respect to the number of measurements. Therefore, in the presence of scalar noise, to perform exact estimation, it is necessary to update the full distribution after each measurement.

The maximum likelihood estimator (MLE) associated with Eq (2) can be derived analytically by setting the derivative of the logarithm of the likelihood function to zero, which leads to the following solution:

$$f_{MLE}(m_1, m_2, ..., m_N) = \overline{m}\left[\frac{-1 + \sqrt{1 + 4w_m^2 \frac{\overline{m^2}}{\overline{m}^2}}}{2w_m^2}\right], \tag{3}$$

where

$$\overline{m} = \frac{1}{N}\sum_{i=1}^{N} m_i, \tag{4}$$

and

$$\overline{m^2} = \frac{1}{N}\sum_{i=1}^{N} m_i^2, \tag{5}$$

f_{MLE} is a nonlinear function that depends on the measurements and their squares (e.g. the right side of the equation). Fig 1 demonstrates f_{MLE} for N = 2. Evidently, MLE for different combinations of measurements is convex. This indicates a larger values of m_i contribute more to f_{MLE}. This is somewhat counterintuitive because larger m_i is typically associated with larger s whose measurement is less reliable. Nonetheless, in the presence of scalar noise, larger measurements contribute more to f_{MLE}.
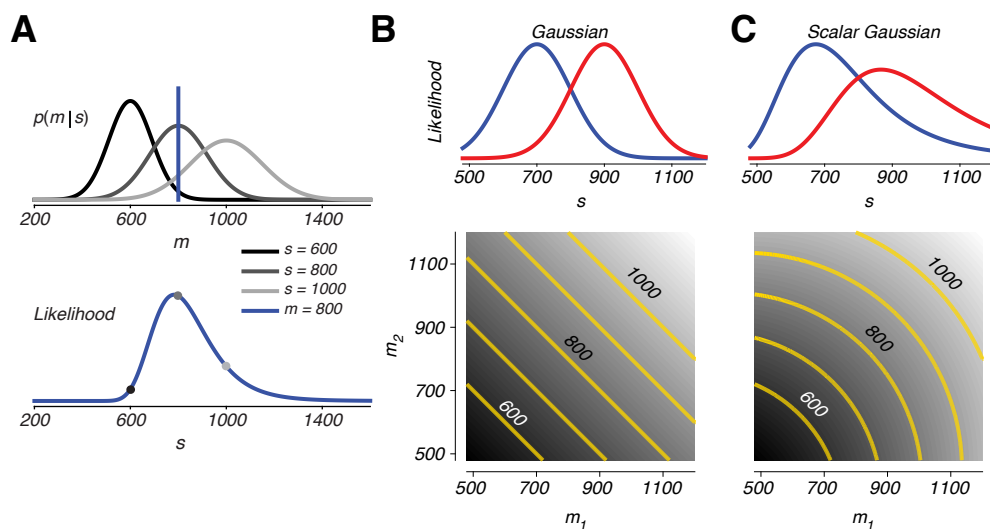
A1 Fig. Impact of scalar variability on magnitude estimation. (A) Origin of the skew in the likelihood function for a scalar Gaussian noise process. Top: the distribution of measurements (m) for a sample of 600, 800, or 1000 with the Weber fraction for measurement, w_m, set to 0.15. Scalar variability results in a Gaussian of increasing variance for increasing sample intervals. Vertical line plots a hypothetical measurement of 800 which could, with differing likelihoods, come from any of the three distributions. Notice that the large value of s is more likely than the small value of s. Bottom: The likelihood of each sample interval for a measurement of 800 (blue). Black, dark gray, and light gray circles plot the relative likelihood that the sample interval was 600, 800, or 1000, respectively. Extrapolation from these examples provides intuition for the long tail (skew) in the likelihood of a scalar Gaussian. (B-C) Consequences of skew in the likelihood function for measurement integration. Top row: likelihood functions for two measurements, $m_1 = 700$ and $m_2 = 900$ (blue and red, respectively) for (B) Gaussian and (C) scalar Gaussian likelihood functions. Bottom row: mapping functions for maximum likelihood estimators (MLE) corresponding to each likelihood function. Grayscale indicates the magnitude estimate where whiter points indicate greater magnitudes; contours plot loci of equal estimates for different combinations of measurements for estimates from 600 to 1000, spaced by 100. Measurements are given equal weight under the Gaussian noise model which has no skew. This linearity appears in the mapping function as straight contours in the (m_1, m_2) plane (panel B). Skew in the likelihood function leads to a discounting of the smaller measurement in the MLE, as revealed by the curvature of the contours of the associated mapping function (panel C).

Acknowledgments

References

1. Wolpert DM, Ghahramani Z, Jordan MI. An internal model for sensorimotor integration. Science. 1995;269(5232):1880–1882.

2. Ariff G, Donchin O, Nanayakkara T, Shadmehr R. A real-time state predictor in motor control: study of saccadic eye movements during unseen reaching movements. J Neurosci. 2002;22(17):7721–7729.

3. Shadmehr R, Smith MA, Krakauer JW. Error correction, sensory prediction, and adaptation in motor control. Annu Rev Neurosci. 2010;33:89–108.

4. Franklin DW, Wolpert DM. Computational mechanisms of sensorimotor control. Neuron. 2011;72(3):425–442.

5. Todorov E, Jordan MI. Optimal feedback control as a theory of motor coordination. Nat Neurosci. 2002;5(11):1226–1235.

6. Landy MS, Maloney LT, Johnston EB, Young M. Measurement and modeling of depth cue combination - in defense of weak fusion. Vision Res. 1995;35(3):389–412.

7. Mamassian P, Landy MS. Observer biases in the 3D interpretation of line drawings. Vision Res. 1998;38(18):2817–2832.

8. van Beers RJ, Sittig AC, Gon JJ. Integration of proprioceptive and visual position-information: An experimentally supported model. J Neurophysiol. 1999;81(3):1355–1364.

9. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002;415(6870):429–433.

10. Battaglia PW, Jacobs RA, Aslin RN. Bayesian integration of visual and auditory signals for spatial localization. J Opt Soc Am A Opt Image Sci Vis. 2003;20(7):1391–1397.

11. Körding KP, Wolpert DM. Bayesian integration in sensorimotor learning. Nature. 2004;427(6971):244–247.

12. Adams WJ, Graf EW, Ernst MO. Experience can change the 'light-from-above' prior. Nat Neurosci. 2004;7(10):1057–1058.

13. Alais D, Burr D. The ventriloquist effect results from near-optimal bimodal integration. Curr Biol. 2004;14(3):257–262.

14. Oruç İ, Maloney LT, Landy MS. Weighted linear cue combination with possibly correlated error. Vision Res. 2003;43(23):2451–2468.

15. Fetsch CR, Pouget A, DeAngelis GC, Angelaki DE. Neural correlates of reliability-based cue weighting during multisensory integration. Nat Neurosci. 2012;15(1):146–154.

16. Knill DC, Richards W. Perception as Bayesian Inference. Cambridge University Press; 1996.

17. Blackwell DA, Girshick MA. Theory of Games and Statistical Decisions. Courier Corporation; 1979.

18. Stocker AA, Simoncelli EP. Noise characteristics and prior expectations in human visual speed perception. Nat Neurosci. 2006;9(4):578–585.

19. Narain D, van Beers RJ, Smeets JBJ, Brenner E. Sensorimotor priors in nonstationary environments. J Neurophysiol. 2013;109(5):1259–1267.

20. Brainard DH, Longère P, Delahunt PB, Freeman WT, Kraft JM, Xiao B. Bayesian model of human color constancy. J Vis. 2006;6(11):1267–1281.

21. Maloney LT, Mamassian P. Bayesian decision theory as a model of human visual perception: testing Bayesian transfer. Vis Neurosci. 2009;26(1):147–155.

22. Ma WJ, Jazayeri M. Neural coding of uncertainty and probability. Annu Rev Neurosci. 2014;37:205–220.

23. Raphan M, Simoncelli EP, Scholkopf B, Platt J, Hoffman T. Learning to be Bayesian without Supervision. In: Neural Information Processing Systems; 2006. p. 1145–1152.

24. Jazayeri M, Shadlen MN. Temporal context calibrates interval timing. Nat Neurosci. 2010;13(8):1020–1026.

25. Simoncelli EP. Optimal estimation in sensory systems. The Cognitive Neurosciences, IV. 2009; p. 525–535.

26. Kalman RE. A New Approach to Linear Filtering and Prediction Problems. J Basic Eng. 1960;82(1):35–45.

27. Simon HA. A Behavioral Model of Rational Choice. Q J Econ. 1955;69(1):99–118.

28. Griffiths TL, Vul E, Sanborn AN. Bridging Levels of Analysis for Probabilistic Models of Cognition. Curr Dir Psychol Sci. 2012;21(4):263–268.

29. Griffiths TL, Lieder F, Goodman ND. Rational use of cognitive resources: levels of analysis between the computational and the algorithmic. Top Cogn Sci. 2015;7(2):217–229.

30. Marr D. Vision: A computational approach; 1982.

31. Gallistel CR. Mental Magnitudes. In: Space, Time and Number in the Brain. Elsevier; 2011. p. 3–12.

32. Eisler H. Experiments on Subjective Duration 1868-1975 A Collection of Power Function Exponents. Psychological Bulletin. 1976;83(6):1154–1171.

33. Stevens SS, Greenbaum HB. Regression effect In psychophysical judgment. Perception & Psychophysics. 1966;1:439–446.

34. Acerbi L, Wolpert DM, Vijayakumar S. Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. PLoS Comput Biol. 2012;8(11):e1002771.

35. Miyazaki M, Nozaki D, Nakajima Y. Testing Bayesian models of human coincidence timing. J Neurophysiol. 2005;94(1):395–399.

36. Cicchini GM, Arrighi R, Cecchetti L, Giusti M, Burr DC. Optimal encoding of interval timing in expert percussionists. J Neurosci. 2012;32(3):1056–1060.

37. Kwon OS, Knill DC. The brain uses adaptive internal models of scene statistics for sensorimotor estimation and planning. Proc Natl Acad Sci U S A. 2013;110(11):E1064–73.

38. Tassinari H, Hudson TE, Landy MS. Combining priors and noisy visual cues in a rapid pointing task. J Neurosci. 2006;26(40):10154–10163.

39. Berniker M, Voss M, Kording K. Learning priors for Bayesian computations in the nervous system. PLoS One. 2010;5(9).

40. Gallistel CR, Gibbon J. Time, rate, and conditioning. Psychol Rev. 2000;107(2):289–344.

41. Ivry RB, Hazeltine RE. Perception and production of temporal intervals across a range of durations: evidence for a common timing mechanism. J Exp Psychol Hum Percept Perform. 1995;21(1):3–18.

42. Merchant H, Zarco W, Prado L. Do we have a common mechanism for measuring time in the hundreds of millisecond range? Evidence from multiple-interval timing tasks. J Neurophysiol. 2008;99(2):939–949.

43. Rakitin BC, Gibbon J, Penney TB, Malapani C, Hinton SC, Meck WH. Scalar expectancy theory and peak-interval timing in humans. J Exp Psychol Anim Behav Process. 1998;24(1):15–33.

44. Gibbon J. Scalar expectancy theory and Weber's law in animal timing. Psychol Rev. 1977;84(3):279–325.

45. Getty DJ. Discrimination of short temporal intervals: A comparison of two models. Percept Psychophys. 1975;18(1):1–8.

46. Drake C, Botte MC. Tempo sensitivity in auditory sequences: evidence for a multiple-look model. Percept Psychophys. 1993;54(3):277–286.

47. Miller NS, McAuley JD. Tempo sensitivity in isochronous tone sequences: the multiple-look model revisited. Percept Psychophys. 2005;67(7):1150–1160.

48. Cai MB, Eagleman DM. Duration estimates within a modality are integrated sub-optimally. Front Psychol. 2015;6.

49. Burr D, Banks MS, Morrone MC. Auditory dominance over vision in the perception of interval duration. Exp Brain Res. 2009;198(1):49–57.

50. Hartcher-O'Brien J, Di Luca M, Ernst MO. The duration of uncertain times: audiovisual information about intervals is integrated in a statistically optimal fashion. PLoS One. 2014;9(3):e89339.

51. Elliott MT, Wing AM, Welchman AE. Moving in time: Bayesian causal inference explains movement coordination to auditory beats. Proceedings of the Royal Society of London B: Biological Sciences. 2014;281(1786):20140751.

52. Di Luca M, Rhodes D. Optimal Perceived Timing: Integrating Sensory Information with Dynamically Updated Expectations. Sci Rep. 2016;6:28563.

53. Shi Z, Ganzenmüller S, Müller HJ. Reducing bias in auditory duration reproduction by integrating the reproduced signal. PLoS One. 2013;8(4):e62065.

54. Ogden RS, Jones LA. More is still not better: testing the perturbation model of temporal reference memory across different modalities and tasks. Q J Exp Psychol. 2009;62(5):909–924.

55. Knill DC, Pouget A. The Bayesian brain: the role of uncertainty in neural coding and computation. Trends Neurosci. 2004;27(12):712–719.

56. Stengel RF. Optimal control and estimation. Dover Publications Inc.; 1994.

57. Bowers JS, Davis CJ. Bayesian just-so stories in psychology and neuroscience. Psychol Bull. 2012;138(3):389–414.

58. Shi Z, Burr D. Predictive coding of multisensory timing. Curr Opin Behav Sci. 2016;8:200–206.

59. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999;2(1):79–87.

60. Vaziri S, Diedrichsen J, Shadmehr R. Why does the brain predict sensory consequences of oculomotor commands? Optimal integration of the predicted and the actual sensory feedback. J Neurosci. 2006;26(16):4188–4197.

61. Stevenson IH, Fernandes HL, Vilares I, Wei K, Körding KP. Bayesian integration and non-linear feedback control in a full-body motor task. PLoS Comput Biol. 2009;5(12):e1000629.

62. Friston K. The free-energy principle: a unified brain theory? Nat Rev Neurosci. 2010;11(2):127–138.

63. Wang J, Narain D, Hosseini EA, Jazayeri M. Flexible timing by temporal scaling of cortical responses. Nat Neurosci. 2018;21(1):102–110.

64. Merchant H, Averbeck BB. The Computational and Neural Basis of Rhythmic Timing in Medial Premotor Cortex. J Neurosci. 2017;37(17):4552–4564.

65. Petzschner FH, Glasauer S. Iterative Bayesian Estimation as an Explanation for Range and Regression Effects: A Study on Human Path Integration. J Neurosci. 2011;31(47):17220–17229.

66. Hudson TE, Maloney LT, Landy MS. Optimal compensation for temporal uncertainty in movement planning. PLoS Comput Biol. 2008;4(7):e1000130.

67. Gilden DL, Thornton T, Mallon MW. 1/f noise in human cognition. Science. 1995;267(5205):1837–1839.

68. Farrell S, Wagenmakers EJ, Ratcliff R. 1/f noise in human cognition: is it ubiquitous, and what does it mean? Psychon Bull Rev. 2006;13(4):737–741.

69. Körding KP, Wolpert DM. The loss function of sensorimotor learning. Proc Natl Acad Sci U S A. 2004;101(26):9839–9842.

70. Acerbi L, Vijayakumar S, Wolpert DM. On the origins of suboptimality in human probabilistic inference. PLoS Comput Biol. 2014;10(6):e1003661.

71. Drugowitsch J, Wyart V, Devauchelle AD, Koechlin E. Computational Precision of Mental Inference as Critical Source of Human Choice Suboptimality. Neuron. 2016;.

72. Ma WJ, Husain M, Bays PM. Changing concepts of working memory. Nat Neurosci. 2014;17(3):347–356.