Leave-One-Trial-Out (LOTO): A general approach to link single-trial parameters of cognitive models to neural data

Sebastian Gluth[1]* and Nachshon Meiran[2]

[1]Department of Psychology, University of Basel, Switzerland

[2]Department of Psychology and Zlotowski Center for Neuroscience, Ben-Gurion University of the Negev, Israel

*Corresponding author: Sebastian Gluth,
Deparment of Psychology, University of Basel,
Missionsstrasse 62a, 4055 Basel, Switzerland
sebastian.gluth@unibas.ch

Acknowledgements: We thank the members of the Decision Neuroscience and Economic Psychology groups at the University of Basel for helpful discussions and Gilad Zlotkin, Yoav Kessler, and Gideon Rosenthal for suggesting related work in other domains. This work was supported by a grant from the Swiss National Science Foundation (#100014_172761) to SG and by a grant from the Israel Science Foundation (#381-15) to NM.

Competing interests: The authors declare no competing interests.

Abstract

It has become a key goal of model-based neuroscience to estimate trial-by-trial fluctuations of cognitive model parameters for linking these fluctuations to brain signals. However, previously developed methods were limited by being difficulty to implement, time-consuming, or model-specific. Here, we propose an easy, efficient and general approach to estimating trial-wise changes in parameters: Leave-One-Trial-Out (LOTO). The rationale behind LOTO is that the difference between the parameter estimates for the complete dataset and for the dataset with one omitted trial reflects the parameter value in the omitted trial. We show that LOTO is superior to estimating parameter values from single trials and compare it to previously proposed approaches. Furthermore, the method allows distinguishing true variability in a parameter from noise and from variability in other parameters. In our view, the practicability and generality of LOTO will advance research on tracking fluctuations in latent cognitive variables and linking them to neural data.

Keywords: Cognitive modeling, intra-individual variability, leave-one-out, jackknife, model-based cognitive neuroscience

I. Introduction

Model-based cognitive neuroscience attempts to link mathematical models of cognitive processes to neural data for advancing our knowledge about the mind and brain (Forstmann et al., 2011). A particularly promising but challenging approach in this regard is to derive trial-specific values for parameters of cognitive models and to relate these values to trial-specific brain data, which offers insights into cognitive and neural principles on a highly detailed level of analysis (Gluth and Rieskamp, 2017; van Maanen et al., 2011; Turner et al., 2017; Wiecki et al., 2013). In the present work, we introduce a novel technique to capturing trial-specific values of model parameters that is efficient, general but also easy to apply. In the following, we briefly summarize related work and methods before we turn to our current proposal.

Primarily, the difficulty in linking momentary fluctuations in cognitive states to neural measures lies in specifying the variability in the cognitive model. Often, cognitive models specify the distribution from which this variability is assumed to originate (e.g., a normal distribution), but remain silent about its direction and extent in single trials. Previous attempts to capture parameter variability have often been specific to a single model. For example, van Maanen and colleagues (2011) derived maximum likelihood estimates for two single-trial parameters of the Linear Ballistic Accumulator (LBA) model (Brown and Heathcote, 2008). Since the LBA model is of special interest for the present paper, we describe it in some detail here: The LBA belongs to the class of sequential sampling models of decision making, which assume that decisions arise from an evidence-accumulation process that continues until a threshold has been reached, indicating that the required amount of evidence for a decision has been gathered. Three critical parameters of the LBA are the rate of evidence accumulation (drift rate), the amount of evidence required to reach a decision (decision threshold), and the point from which the accumulation begins (start point). Importantly, sequential sampling models such as LBA predict accuracy rates and the shape of response times (RT) distributions conjointly. Wiecki, Sofer, and Frank (2013) introduced a hierarchical Bayesian modeling tool

for another sequential sampling model, the Diffusion Decision Model (DDM) (Ratcliff, 1978). Their tool allows regressing single-trial parameters of the DDM onto trial-by-trial neural measures such as electroencephalography (EEG) or event-related functional magnetic resonance imaging (fMRI) recordings (similar regression-based approaches were proposed by Hawkins et al., 2017; Nunez et al., 2017). Turner and colleagues proposed a related method, whose rationale is to jointly model behavioral and neuroimaging data under a common (hierarchical Bayesian) umbrella (Palestro et al., 2018; Turner et al., 2013, 2015). Although the approach was again tested within the DDM framework, it may be generalized to any other cognitive model. As we have argued before, however, the complexity of this approach might discourage many cognitive neuroscientists from applying it (Gluth & Rieskamp, 2017). Another but less general approach of modeling behavioral and brain data jointly was proposed by van Ravenzwaaij, Provost and Brown (2017).

We recently proposed a comparatively simple and very general approach to capture trial-by-trial variability in cognitive model parameters (Gluth & Rieskamp, 2017). Again, the approach is based on Bayesian principles: It specifies the posterior probability of a parameter value in a specific trial, using the “average” parameter value across all trials and the behavior in the specific trial as the prior and likelihood, respectively. Thus, the method basically answers the question of how the difference between a person’s behavior in specific trials vs. in all trials can be mapped (in a Bayesian optimal way) onto changes in a specific parameter of interest. For example, a surprisingly fast decision of an otherwise very cautious decision maker might be attributed to a reduced decision threshold (in the context of sequential sampling models such as LBA or DDM), which then might be linked to altered neural activation in a specific region of the brain such as the pre-supplementary motor area (Gluth et al., 2012; van Maanen et al., 2011). While this approach is both general and comparatively simple, it can require high amounts of computation time, in particular when one seeks to simultaneously estimate variability in more than one parameter.

In the current work, we propose a novel approach to capturing trial-by-trial variability in cognitive model parameters that tries to overcome the aforementioned shortcomings of specificity, complexity, and inefficiency: Leave-One-Trial-Out (LOTO). Briefly, the idea of LOTO is that the difference in the parameter estimates that are based on all trials vs. based on all trials except a single trial provides a reflection of the "true" parameter value in the left-out trial. The rest of the article is structured as follows: First, we introduce the rationale of LOTO and exemplify this rationale with a simple toy model. Second, we show under which circumstances LOTO can be expected to capture trial-by-trial variability appropriately, and explain how LOTO's performance can be improved by adapting its application, the underlying cognitive model or the experimental design. Third, we illustrate how one can test the assumption of the presence of systematic parameter variability. Fourth, we compare LOTO to previously proposed methods. Fifth, we present a simulation-based power analysis that aims at testing whether LOTO can be expected to identify the neural correlates of trial-by-trial fluctuations of model parameters with sufficient power and specificity. Finally, we provide an example analysis, in which we apply LOTO to link variability in a cognitive process (encoding of episodic memory) to brain signals (fMRI activation in the hippocampus). Although we consider it being essential for the potential user to read the entire article (for understanding when and why LOTO can be expected to provide acceptable results), we note that some parts of the article are rather technical and may be difficult to understand. Therefore, our suggestion for readers with limited statistical and mathematical knowledge is to read Section II, which introduces the LOTO principle, and Section XI, which exemplifies the application of LOTO, and then to proceed directly to the discussion, in which we provide a "LOTO recipe" that provides a step-by-step summary of how LOTO should be applied. The recipe then refers the reader back to the specific sections of the article.

II. The LOTO principle

The rationale of LOTO is simple and intuitive and can be summarized as follows:

- When fitting a (cognitive) model to data, all *n* trials contribute to the model's goodness-of-fit statistic (e.g., log-likelihood) and thus to the estimation of parameter values.
- A single trial *t* shifts the parameter estimation to the direction that accords with the behavior in trial *t*.
- Therefore, if trial *t* is taken out and the model is fitted again to the *n*-1 dataset, the new parameter estimate will shift (slightly) to the opposite direction.
- Therefore, the (true and unknown) parameter value in *t* is reflected in the difference between the parameter estimate of all trials and the parameter estimate of all trials except *t*.

Accordingly, the application of LOTO works as follows:

1. Estimate all parameters of your model by fitting the model to the entire dataset (e.g., all *n* trials of a subject).
2. Re-estimate the parameter(s) of interest by fitting the model to the dataset without trial *t*; keep all other parameters fixed to their estimates derived in step 1.
3. Take the difference between the parameter estimate(s) of step 1 and step 2, which will provide a reflection of the parameter value in *t* (i.e., the difference should be correlated positively with the true parameter value).
4. Repeat the Steps 2 and 3 for all *n* trials.
5. You may then use the obtained vector of trial-by-trial parameter values to link it to any external measure that was recorded at the same time and that has an appropriate temporal resolution (e.g., event-related fMRI data, EEG data, single-unit recordings, eye-tracking data, skin conductance responses; see also Gluth & Rieskamp, 2017).

Note that instead of taking the difference between the estimates for *n* and *n*-1 trials in step 3, one could simply multiply the *n*-1 trials estimates by -1 to obtain LOTO estimates that are

positively correlated with the true parameter values. However, we prefer taking the difference, as this nicely separates above-average from below-average parameter values which will be positive and negative, respectively. In general, it should be obvious that LOTO does not provide single-trial parameter estimates in an absolute sense. Rather, the method provides information about the direction and (relative) amount of deviation of single-trial parameter values from the average. Note that in most cases, cognitive neuroscientists will be interested in the correlation between trial-by-trial parameter values and neural data, for which absolute values do not matter (Gluth and Rieskamp, 2017). In the rest of the article, we sometimes refer to "LOTO estimates of a parameter", but it should be clear that this refers to estimates in a relative sense only.

III. A "toy" model example

In the following, we illustrate the LOTO principle using the binomial distribution as a toy model example. In general, the binomial distributions specifies the probability of observing k successes out of n trials given probability θ, which specifies the probability of a success in each trial. Let us assume that we seek to find the estimate of θ that provides the best account of the observation of k successes in n trials. The Maximum Likelihood Estimate (MLE) of this is simply (e.g., Farrell and Lewandowsky, 2018):

$$\mathrm{MLE}(\theta) = \frac{k}{n} \quad . \tag{1}$$

Let us now assume that we believe that θ is not stable but may vary from trial to trial, and that our goal is to estimate this trial-by-trial variability using LOTO (note that for this and the following sections, we simply assume that such trial-by-trial variability exists; we address the issue of testing for the presence of trial-by-trial variability in Section VI). As explained above, LOTO's estimate for trial t is the difference between the n and the n-1 estimates of θ. For the binomial model, we thus obtain:

$$\mathrm{LOTO}_t = \mathrm{MLE}(\theta) - \mathrm{MLE}(\theta)_{\neg t} = \frac{k}{n} - \frac{k-k_t}{n-1} \ , \tag{2}$$

where k_t indicates whether a success was observed in trial t ($k_t = 1$) or not ($k_t = 0$). The critical question is whether we can expect the LOTO approach to capture changes in observations meaningfully. To answer this question, we take the derivative of Equation 2 with respect to k_t to see how LOTO's estimate changes as a function of k_t:

$$\frac{\partial \mathrm{LOTO}_t}{\partial k_t} = \frac{1}{n-1} \ . \tag{3}$$

Equation 3 implies that changes in single-trial observations map linearly and positively onto LOTO's estimates, which is what we desired. To give an example, let us assume $k = 5$, $n = 10$, $k_{t=1} = 1$, and $k_{t=2} = 0$. Then LOTO's estimates for $t = 1$ and $t = 2$ are 1/18 and -1/18, respectively, with the difference being $1/(n–1) = 1/9$. Equation 3 also reveals a general property of LOTO, namely that the change in LOTO is inversely related to the number of trials: Naturally, the influence of a single trial on the estimation of a parameter decreases as the number of trials increases. We will further elaborate on this issue in Section VIII. Also, we see that LOTO_t is not an estimate of θ for trial t in an absolute sense (e.g., a θ of -1/18 is impossible), but that positive and negative LOTO_t values indicate above- and below-average single-trial values, respectively.

So far, we have only established a sensible relationship between LOTO's estimate and the observation k_t, but ultimately we are interested in whether LOTO is also related to the underlying trial-wise (data-generating) parameter θ_t. This depends on how a change in θ_t is mapped onto a change in k_t, which is given by the *Fisher information*, the variance of the first derivative of the log-likelihood function with respect to θ_t. For the binomial distribution with $n = 1$ (i.e., the Bernoulli distribution), the Fisher information is (e.g., Ly et al., 2017):

$$I(\theta) = \frac{1}{\theta * (1-\theta)} \ . \tag{4}$$

As can be seen in the left panel of Figure 1A, the Fisher information of the Bernoulli distribution indicates that observations are more informative about the underlying parameter

at the extremes. Accordingly, LOTO's performance (i.e., the correlation between the true θ_t and LOTO's estimates of it) will also depend on the range of θ_t. To illustrate this, we ran simulations by drawing 300 values of θ_t from a uniform distribution with a range of .2 (e.g., between .4 and .6) and repeated this for every range from [0, .2] to [.8, 1] in steps of .01 (detailed specifications of all simulations are provided in Appendix A). LOTO was then used to recover θ_t. The middle panel of Figure 1A depicts the average correlation between θ_t and its LOTO estimate as a function of the average θ. As predicted by the Fisher information, the correlation is higher when θ is close to 0 or close 1.

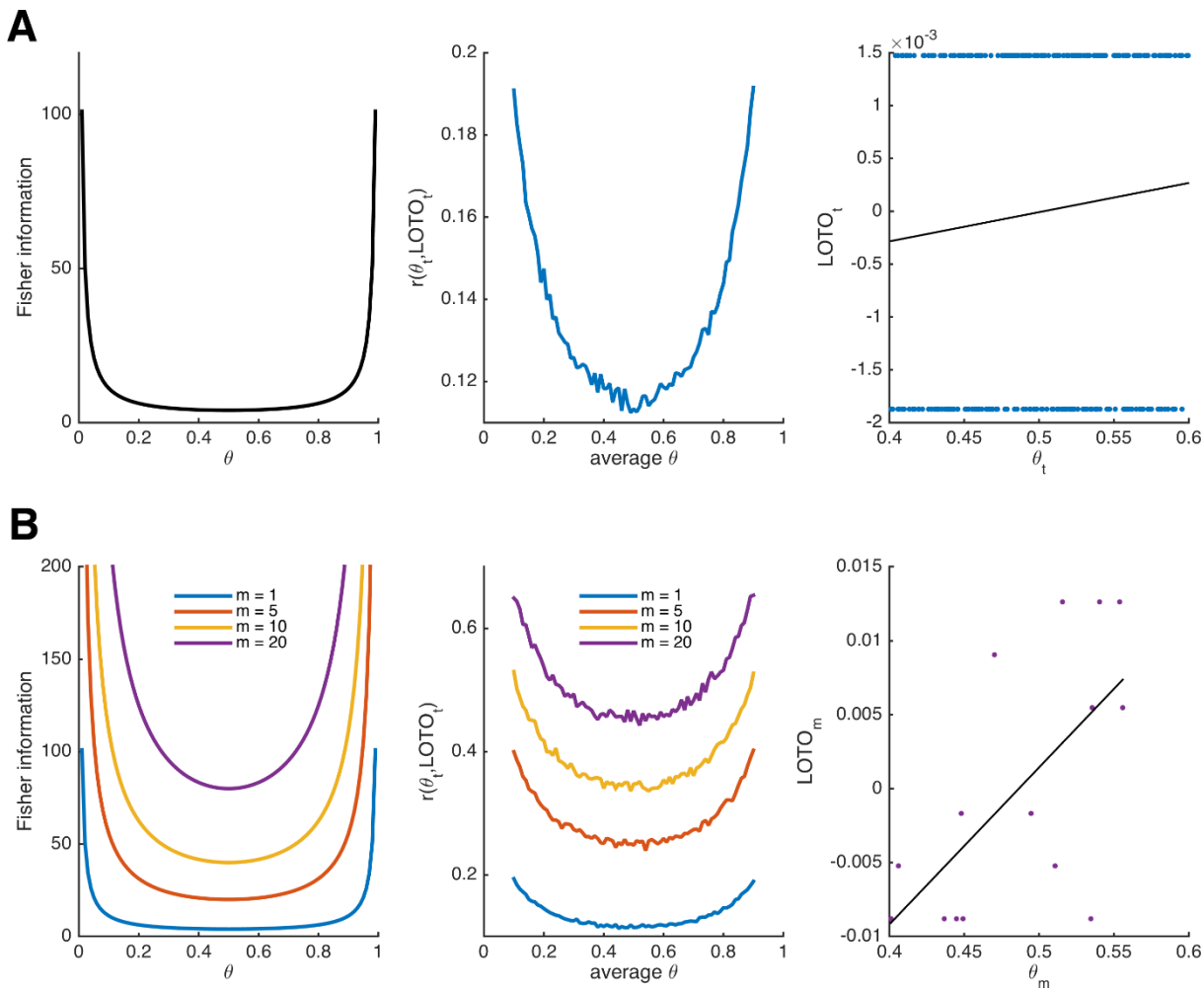

Figure 1. Fisher information and performance of LOTO for the binomial distribution. **A**) Fisher information (left panel), average correlation between and LOTO's estimate (middle panel), and example correlation for values of θ_t in the range [.4, .6] (right panel) for the Bernoulli distribution. **B**) The same as in **A** but generalized to the binomial distribution, that is, assuming

that θ is stable over *m* trials, for different levels of *m* (the example correlation in the right panel uses *m* = 20).

Notably, the (average) correlations between θ_t and LOTO's estimate are quite low (i.e., < .2). The reason for this can be seen in the right panel of Figure 1A, which shows the correlation for an example of 300 trials in the range [.4, .6]: The estimate of LOTO can only take two values, depending in whether a success was observed in *t* or not. Fortunately, the performance can be improved by assuming that θ does not change from trial to trial but remains constant over a certain number of trials, which we denote *m*. The larger *m*, the more different values LOTO can take (and at the same time our confidence in the estimate of θ increases). LOTO's estimate, its derivative with respect to *m*, and the Fisher information for the Bernoulli distribution generalized to $m \geq 1$ (i.e., the binomial distribution) are:

$$\mathrm{LOTO}_m = \mathrm{MLE}(\theta) - \mathrm{MLE}(\theta)_{\neg m} = \frac{k}{n} - \frac{k-k_m}{n-m} \ , \tag{5a}$$

$$\frac{\partial \mathrm{LOTO}_m}{\partial k_m} = \frac{1}{n-m} \ \text{, and} \tag{5b}$$

$$I(\theta) = \frac{m}{\theta*(1-\theta)} \ \text{, respectively.} \tag{5c}$$

Figure 1B illustrates the increases in Fisher information and LOTO's performance as *m* increases. This is true, even though we kept the number of draws to 300 (implying that when *m* = 20, there are only *n*/*m* = 300/20 = 15 values estimated by LOTO; see Figure 1B, right panel). Thus, it might be more promising to track fluctuations over multiple trials (e.g., short blocks or sessions) than trying to capture parameter values in each and every trial. For example, Gluth et al. (2015) investigated memory-based decisions and inferred whether an item had been remembered based on three trials that included the item as a choice option (see also Section XI, in which we use LOTO to reproduce this analysis).

We see that LOTO can be applied to the Bernoulli and binomial distributions, but it should be noted that LOTO is not particularly helpful in these cases. This is because LOTO will not provide us with any novel information over and above what is already known from

the observations. In other words, LOTO's estimates are perfectly correlated with k_t or k_m, which is mirrored in the derivatives (Equations 3 and 5b) that depend only on constants n or m. In principle, there are two ways to make LOTO (and any other method for tracking trial-by-trial variability in parameters) more useful: First, one could introduce variations in the task, with these variations being directly linked to the predictions of the cognitive model. Second, one could apply cognitive models that predict a "richer" set of data (e.g., sequential sampling models that predict accuracy rates and RT distributions conjointly; see above). We first turn to introducing variations in the task in the next section.

IV. Task variations and comparison with "single-trial fitting"

We demonstrate the importance of using variations in the experimental design that are systematically linked to the prediction of the cognitive model within the context of studying intertemporal choices. Intertemporal choices are decisions between a smaller reward available now or in near future vs. a larger reward available only in the more distant future (an example trial of a typical intertemporal choice task could be: *"Do you want to receive $20 now or $50 in 20 days?"*). Traditionally, these decisions have been modeled using temporal discounting models that describe the decrease in utility u_i of an option i as a function of the delay d_i of its pay-out (Frederick et al., 2002). We apply the hyperbolic discounting (HD) model (Mazur, 1987), according to which:

$$u_i = \frac{x_i}{1+\kappa * d_i} \ , \qquad (6)$$

where x_i refers to the amount of reward and κ is the discount rate (i.e., a free parameter modeling the degree to which utilities are lowered by delays; $0 \leq \kappa \leq 1$). To derive choice probabilities (that can then be subjected to MLE for estimating κ), the HD model is usually combined with the logistic (or soft-max) choice function (e.g., Gluth et al., 2017; Peters et al.,

2012), according to which the probability p_i of choosing option i from a set of options including i and j is:

$$p(i|\{i,j\}) = p_i = \frac{1}{1+\exp(-\beta*[u_i-u_j])} \ , \tag{7}$$

where β is a second free parameter modeling the sensitivity of the decision maker to utility differences (β ≥ 0).

Our goal here is to show that by applying LOTO, we can extract more information about parameter κ (which we assume to vary from trial to trial) than what is already provided by the observations. We do this by looking at the first derivatives of the log-likelihood of the HD model with respect to κ. Let us assume that option i is an "immediate" option, offering amount x_i at delay $d_i = 0$, so that $u_i = x_i$, and that option j is a "delayed" option, offering amount x_j ($x_j > x_i$) at delay d_j ($d_j > 0$). It turns out that the first derivative of the log-likelihood for choosing the immediate option is (for details, see Appendix B):

$$\frac{\partial LL_i}{\partial \kappa} = \frac{\beta*x_j*d_j}{[1+\kappa*d_j]^2} * (1-p_i) \ , \tag{8a}$$

Similarly, the first derivative of the log-likelihood for choosing the delayed option is:

$$\frac{\partial LL_j}{\partial \kappa} = -\frac{\beta*x_j*d_j}{[1+\kappa*d_j]^2} * p_i \ . \tag{8b}$$

Note that the MLE can be obtained by setting these derivatives to 0. Importantly, we see that in contrast to the binomial distribution the derivatives of the HD model depend on features of the task, namely on the amount and delay of the delayed option (i.e., x_j and d_j) and on the amount of the immediate option (i.e., x_i), which is hidden in p_i. At first glance, this already looks as if the estimation of single-trial values of κ_t will vary with the task's features, which is what we desired. Note, however, that both of the two terms on the right-hand side of the derivatives are ≥ 0. Thus, by setting the derivatives in Equation 8a and 8b to 0 for obtaining the MLEs, parameter κ_t will take the highest possible value (i.e., $\kappa_t = 1$) whenever the immediate option is chosen and the lowest possible value (i.e., $\kappa_t = 0$) whenever the delayed

option is chosen (intuitively speaking: the best account for choosing the immediate option is to assume maximal impatience, the best account for choosing the delayed option is to assume is maximal patience). This is illustrated in Figure 2A and B, which depict the log-likelihoods of choosing *i* (left panels) or *j* (middle panels) and their derivatives as a function of κ for two different trials with $x_{i,1} = 20$, $x_{j,1} = 50$, and $d_{j,1} = 20$ in trial 1 (bright colors) and $x_{i,2} = 18$, $x_{j,2} = 100$, and $d_{j,2} = 25$ in trial 2 (dark colors): Independent of the features of the task (i.e., the trial-specific amount and delays), the log-likelihoods of choosing *i* are monotonically increasing, and their derivatives are always positive, whereas the log-likelihoods of choosing *j* are monotonically decreasing, and their derivatives are always negative.

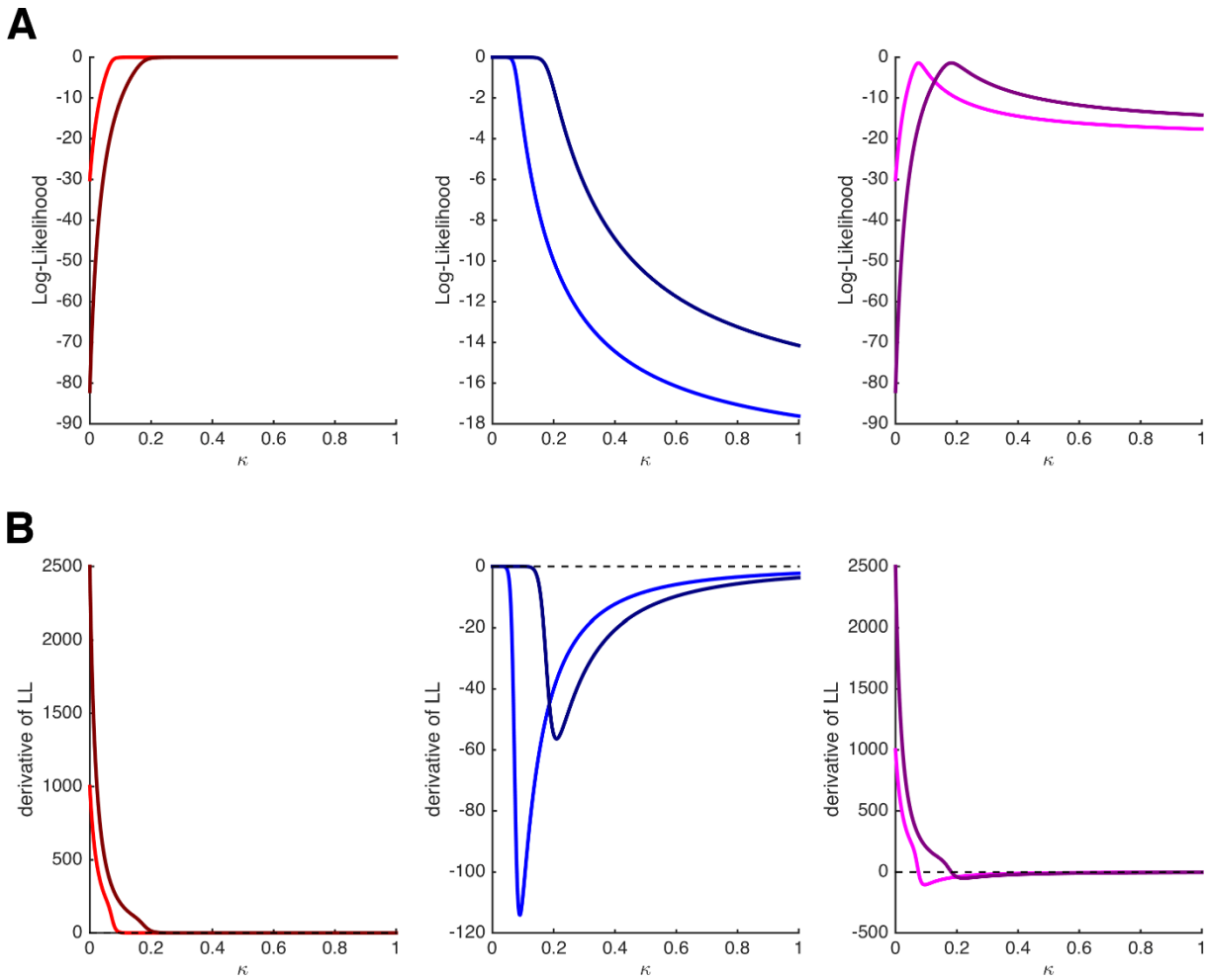

Figure 2. Log-likelihoods and derivatives for the hyperbolic discounting model. **A**) Log-likelihoods for choosing the immediate option (left panel), choosing the delayed option (middle panel), and choosing once the immediate and once the delayed option (right panel) as a function of parameter κ. The bright and dark colors refer to two different trials with different amounts

and delays. **B**) First derivatives of the log-likelihoods shown in **A**. Note that only when both options are chosen at least once, the estimation of κ will depend on the features of the task.

Critically, however, this caveat disappears as soon as each option (*i* and *j*) is chosen at least once. For instance, when keeping x_i, x_j, and d_j constant over two trials in which both *i* and *j* are chosen once, the derivative of the log-likelihood becomes:

$$\frac{\partial LL_{\{i,j\}}}{\partial \kappa} = \frac{\partial LL_i}{\partial \kappa} + \frac{\partial LL_j}{\partial \kappa} = \frac{\beta * x_j * d_j}{[1+\kappa * d_j]^2} * (1 - 2 * p_i) \,, \tag{8c}$$

which is positive if $p_i < 0.5$ but negative if $p_i > 0.5$. The right panels of Figure 2A and B show the log-likelihoods and their derivatives when assuming that trial 1 or trial 2 were presented twice, and *i* was chosen in one occasion and *j* was chosen in the other occasion. Here, the derivatives cross the 0-line, and even more importantly, they do so at different values of κ for the different trials (the MLE of κ for trial 2 is higher because the immediate option is comparatively less attractive than in trial 1).

Why does this allow LOTO to extract more information about κ_t than what is given by the observations? Recall that the estimate of LOTO is the difference between the MLEs of κ for all trials and for all trials except *t*. As explicated above, these two MLEs will depend on the task features in the *n* / *n*-1 trials as long as both choice options (e.g., immediate, delayed) have been chosen at least once in the *n* / *n*-1 trials (so that we get non-extreme estimates of κ from the *n* / *n*-1 trials). In principle, LOTO can therefore be applied as soon as both options have been chosen ≥ 2 times over the course of an experiment (if one option has been chosen in only one trial, then the estimate of κ when excluding this particular trial will be extreme as in the single-trial case). We illustrate this using the two example trials from Figure 2 in Appendix C.

It follows that LOTO should outperform the arguably most elementary method of capturing trial-by-trial parameter variability, which we denote "single-trial fitting". With single-trial fitting, we mean that after estimating all parameters of a model once for all trials, a

specific parameter of interest (e.g., κ of the HD model) is fitted again to each trial individually (while all other parameters are fixed to their "all-trial" estimates). Here, we show that LOTO but not single-trial fitting of parameter κ of the HD model allows to extract additional information about the true κ_t over and above what is already provided by the observations. Thereto, we generated 160 trials with a constant immediate choice option and variable delayed choice options, and simulated decisions of a participant with κ_t varying from trial to trial (Appendix A). Then, we applied LOTO and the single-trial fitting method to recover κ_t. As can be seen in Figure 3, only the estimates of LOTO (left panel) but not of single-trial fitting (right panel) are correlated with the true κ_t after accounting for the observations (i.e., whether the immediate or the delayed option was chosen). In other words, the positive correlation between κ_t and the single-trial estimates (dashed magenta line) can be attributed entirely to the fact that higher values of κ_t lead to more choices of the immediate option, but the positive correlation between κ_t and the LOTO estimates is further driven by how expectable/surprising a decision is (given the trial features $x_{i,t}$, $x_{j,t}$, and $d_{j,t}$).

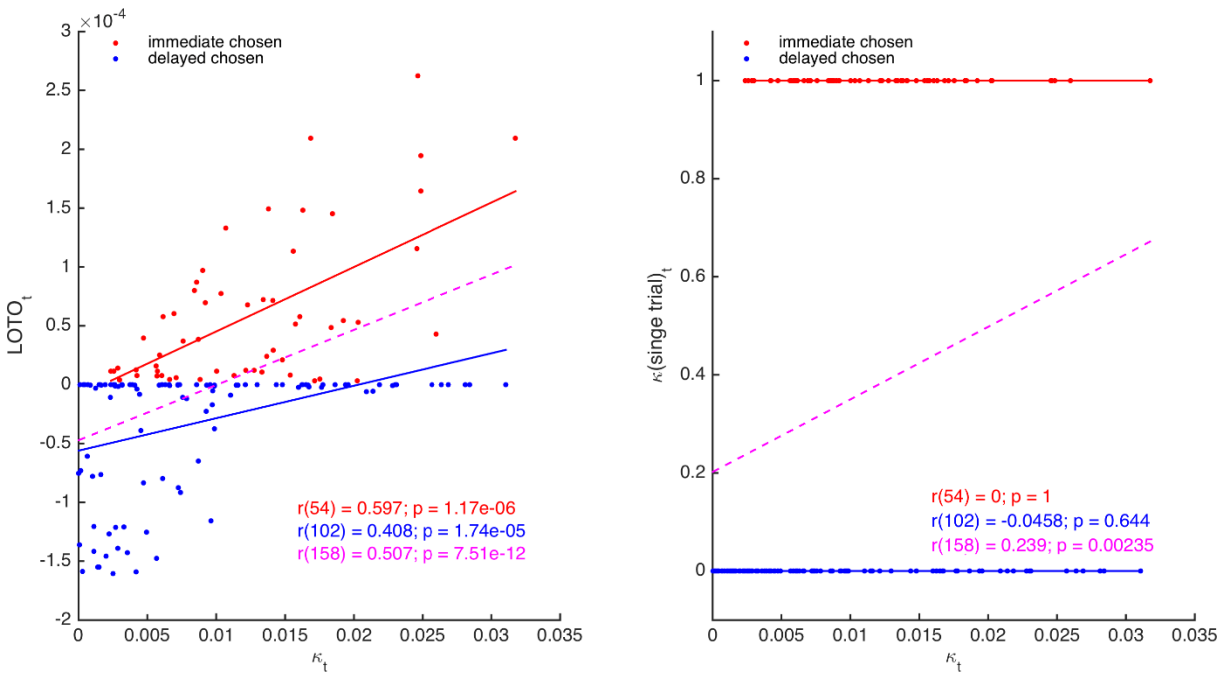

Figure 3. Comparison of LOTO and single-trial fitting. The left panel shows the correlation of the trial-specific parameter κ_t of the HD model with LOTO's estimates, separately for choosing the immediate (red) and choosing the delayed (blue) option. The dashed line (magenta) shows the regression slope when collapsing over all trials. The right panel shows the same for the single-trial fitting approach.

To summarize, one approach to extract information about trial-by-trial changes in a parameter–over and above what is already provided by the observations–is to introduce variations in the experiment that are linked in a systematic manner to the cognitive model. Because LOTO's estimate are not solely dependent on the behavior in a single trial but also take all remaining trials into account, it allows to capture this additional information. Note, however, that for specific models such as sequential sampling models of decision making, LOTO can be useful even without introducing task variations. In the next section, we will further improve LOTO's ability to recover trial-specific parameter values by extending the HD model to allow predicting choice and RT conjointly.

V. LOTO benefits from modeling an enriched set of data

Using LOTO to recover trial-specific parameter values of models that predict only binary outcomes such as choices or accuracy rates is possible but somewhat limited. This can be seen when looking back at the left panel of Figure 3, which suggests that LOTO's estimates are always higher when the immediate option was chosen. The reason for this is that taking one trial out in which the immediate/delayed option was chosen will always de-/increase the estimation of the discount factor κ, because–as stated above–the single-trial log-likelihood is monotonically in-/decreasing given the choice (the magnitude of this change varies as a function of the trial-specific choice options, but the sign of this change is determined by the choice). However, this strict choice-dependent separation might not be plausible: For example, it could very well be that the delayed option is chosen in a trial in which this option offered a very high amount x_j at a very low delay d_j even though the actual discount factor in

that trial (i.e., κ_t) was higher than on average. To overcome this limitation, we suggest to extend the model so that it allows taking RT data into account. Going back to the example, the intuition is that even if the highly attractive delayed option is chosen, a high discount factor κ_t will make this decision more difficult which will result in a comparatively long RT. Applying LOTO to the extended model then allows capturing this RT-specific information.

We extended the HD model by adding the LBA model to it (in a similar but not equivalent way as Rodriguez et al., 2014, 2015). Specifically, we assume that the difference in utility between the immediate and delayed options is mapped onto their LBA's expected drift rates as follows:

$$\nu_i = \frac{1}{2} + \lambda * \left(u_i - u_j\right) , \tag{9a}$$

$$\nu_j = 1 - \nu_i = \frac{1}{2} + \lambda * \left(u_j - u_i\right) . \tag{9b}$$

Where λ is a free parameter that scales the relationship between utility and expected drift rate. Note that these definitions ensure that the two expected drift rates sum up to 1, following the original specification of the LBA (Brown and Heathcote, 2008). Furthermore, these definitions ensure that $\nu_i > \nu_j$ whenever $u_i > u_j$, $\nu_i < \nu_j$ whenever $u_i < u_j$, and $\nu_i = \nu_j = ½$ if $u_i = u_j$ (but note that actual drift rates are assumed to be drawn from independent normal distributions with means ν_i, ν_j and standard deviation s; see Appendix A). Choices and RT were simulated for 160 trials of 30 participants and LOTO was applied to the HD and the HD-LBA models. Single-trial fitting was also applied to the HD-LBA model. Figure 4A depicts correlations of an example participant between the true κ_t and LOTO's estimate for HD (left panel) and for HD-LBA (right panel). For HD-LBA, the correlation is much improved, and LOTO's estimates do not strictly separate choices of the immediate and delayed options anymore. The improved performance of LOTO is confirmed when comparing the correlations of all 30 simulated participants (Figure 4C; immediate chosen: $t(29) = 10.79$, $p < .001$, Cohen's $d = 1.97$; delayed chosen: $t(29) = 8.79$, $p < .001$, $d = 1.61$).

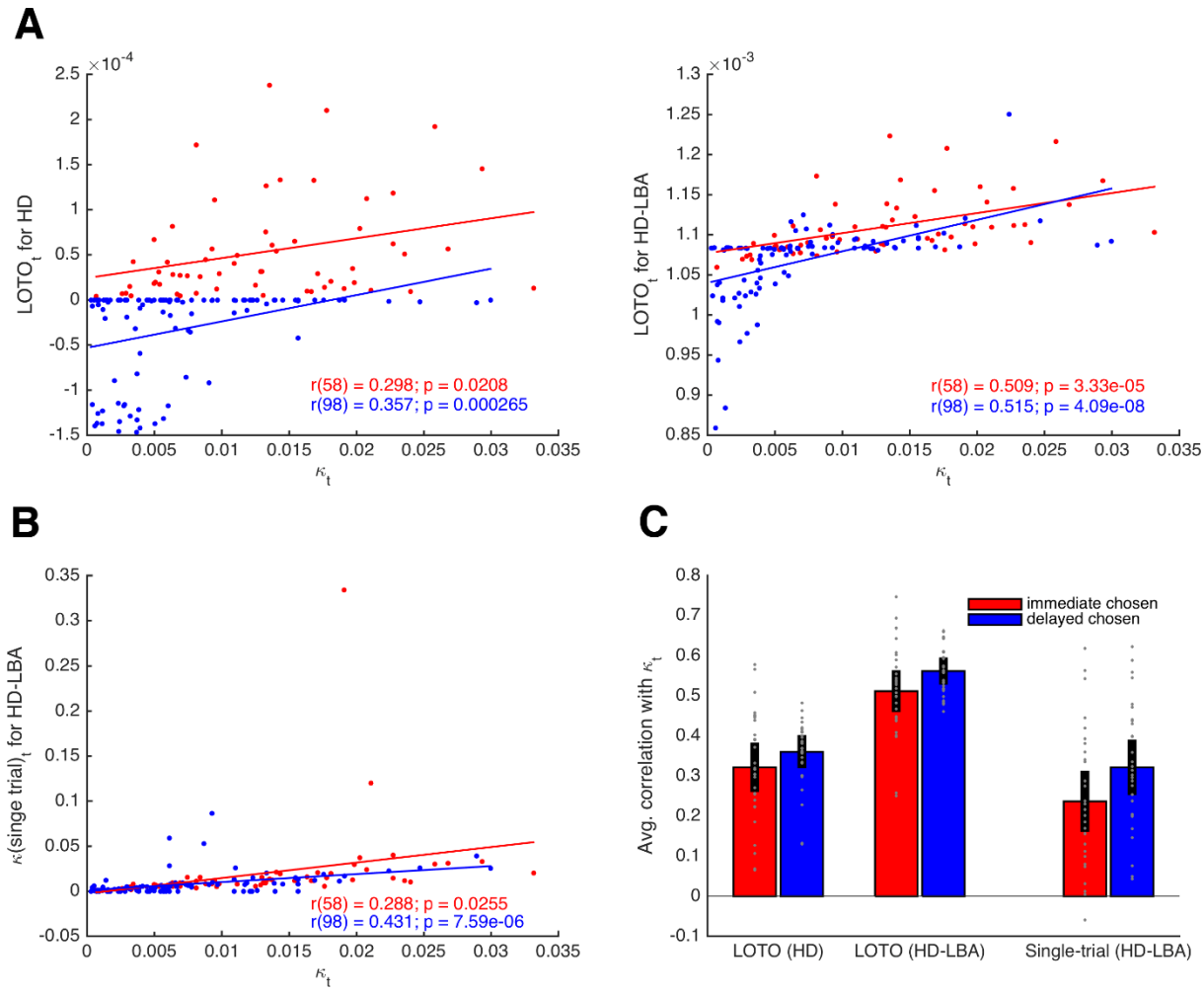

Figure 4. Comparison of LOTO for the HD and the HD-LBA model. **A**) Correlation of the trial-specific parameter κ_t with LOTO's estimates for an example participant. The left panel shows the results when using the HD model, the right panel when using the HD-LBA model. **B**) Same as in **A** but for the single-trial fitting method when using the HD-LBA model. **C**) Average correlation for all 30 simulated participants. Individual values (gray dots) are shown together with 95% confidence intervals of the mean (black error bars).

Notably, the single-trial fitting method applied to the HD-LBA model is also able to capture some variability over and above what is already contained in the decisions (Figure 3B). This is because RT (and the interplay between RT and decisions) add further information about κ_t that can be picked up by single-trial fitting. Nonetheless, applying LOTO to HD-LBA yields significantly better recovery of the trial-wise parameter values (Figure 4C; immediate chosen: $t(29) = 7.24$, $p < .001$, $d = 1.32$; delayed chosen: $t(29) = 7.71$, $p < .001$, $d = 1.41$). Taken together, recovering trial-by-trial changes in a parameter of a cognitive model with LOTO (but also with other techniques) is much more efficient when the model allows to

predict a rich set of single-trial observations, such as RT (and possibly even continuous ratings of choice confidence as in Pleskac and Busemeyer, 2010) in addition to choices or accuracy rates.

VI. A non-parametric test for the presence of systematic trial-by-trial variability

So far, our derivations and simulations rested on the assumption that systematic variability in the parameter of interest exists. As we have argued before, however, this assumption does not have to be correct– not even when the overt behavior appears to be stochastic (Gluth and Rieskamp, 2017). It could be that stochasticity is induced by random fluctuations that are not accounted for by the cognitive model (e.g., randomness in perception or motor execution). When assuming variability in a specific parameter, the risk is that such unsystematic variability is taken up by LOTO even in the absence of any systematic variability in the parameter. Hence, the question is whether one can separate situations in which LOTO captures systematic vs. unsystematic trial-by-trial variability.

To test for the presence of systematic trial-by-trial variability, we propose a non-parametric and simulation-based approach that exploits the improvement in model fit by LOTO: If a model is fitted to n-1 trials, the fit statistic (e.g., the deviance) will improve for two reasons: i.) the statistic for the left-out trial is excluded; ii.) the parameter estimate is adjusted so that it optimally fits the n-1 dataset. The rationale of our proposed approach is that the improvement in fit due to the second reason (i.e., the adjustment) will be lower when there is no systematic trial-by-trial variability in the parameter of interest compared to when there is systematic variability. Practically, we suggest conducting this test in five steps:

1. For each participant, estimate parameters of the cognitive model for all trials and compute the goodness-of-fit per trial.

2. For each participant, generate many (e.g., 1’000) simulations of the model for all trials and all participants using the participants’ estimated parameters and assuming no variability in the parameter of interest (i.e., the null hypothesis).
3. Estimate parameters for these simulated data again, apply LOTO, and compute the LOTO-based improvement in goodness-of-fit of the *n*-1 trials.
4. Do the same as in step 3, but for the empirical data.
5. Test whether the LOTO-based improvement in fit for the empirical data is higher than 95% of the improvements in the simulations that were generated under the null hypothesis.

Note that one advantage of this non-parametric approach is that it does not rely on any assumptions of how the improvement of model fit is distributed under the null hypothesis.

Figure 5A shows the results of this test for the HD model introduced in Section IV. We generated 1’000 simulations per participant under the null hypothesis and 10 additional simulations with increasing levels of variability in parameter κ (for details, see Appendix A). Importantly, the figure shows the average difference in deviance between the “all-trial” fit and the LOTO fits only for the *n*-1 trials that were included in each LOTO step. This means that the trivial improvement in fit due to the fact that LOTO has to explain one trial less is taken out. Thus, the histogram in Figure 5A, which shows the results for the 1’000 simulations under the null hypothesis, confirms our suspicion that LOTO will improve the fit even in the absence of true trial-by-trial variability in parameter κ. However, the continuous vertical lines, which show the results for the simulations with true variability of increasing amount (from blue to red), indicate significantly higher improvements in fit when variability is actually present–except for the lowest level of variability (the dashed black line indicates the 95% threshold).

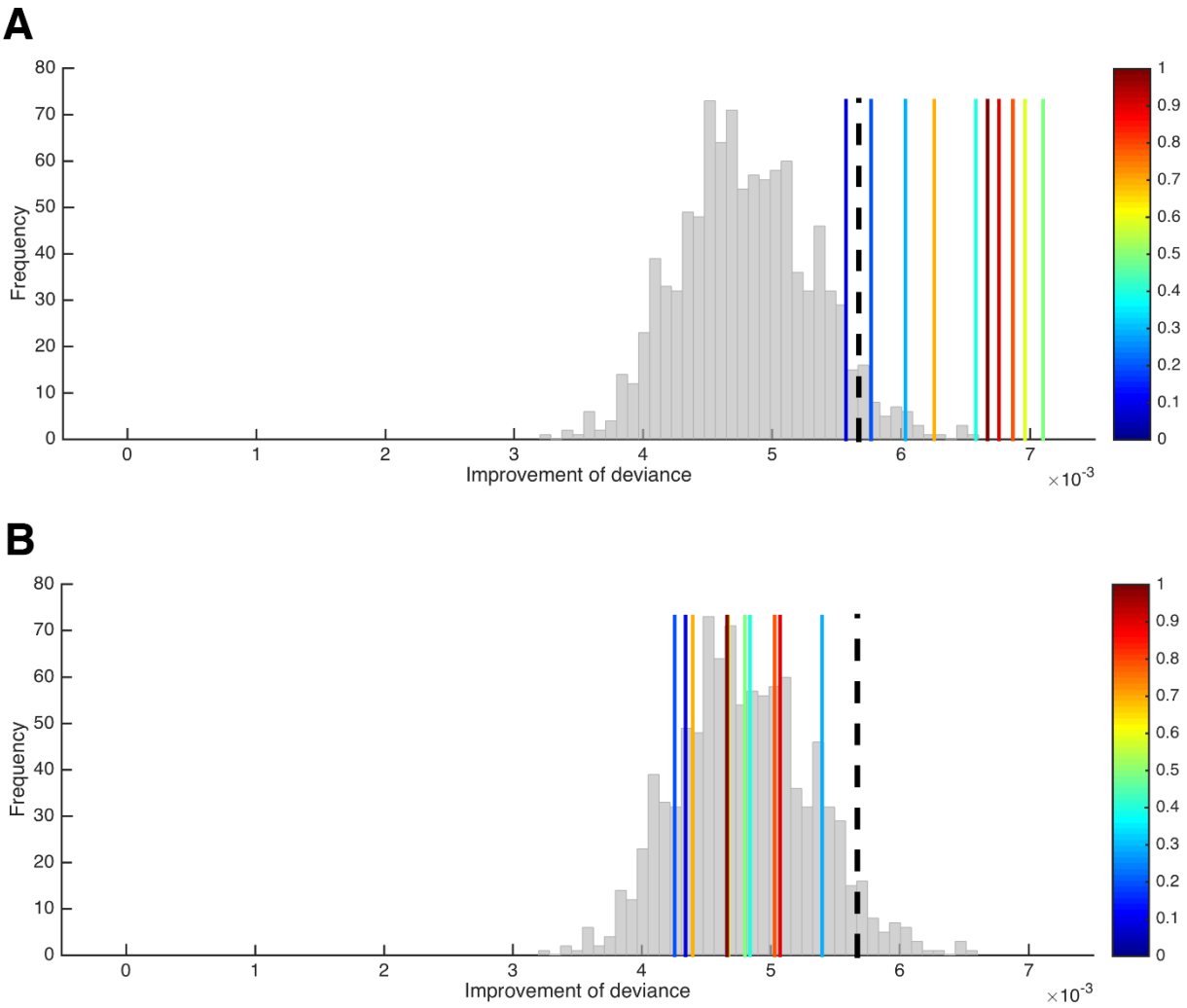

Figure 5. A test for the presence of parameter variability. **A**) Improvement of the HD model's deviance (in the *n*-1 trials included per LOTO step) from the "all-trials" fit to the LOTO fit when variability in κ is absent (histogram) and when variability in κ of increasing amount is present (blue to red vertical lines). The black dashed line indicates the 5% of simulations under the null hypothesis with the highest improvement and serves as the statistical threshold. **B**) Same as in **A**) but with variability in β instead of κ.

What happens if there is no systematic variability in the parameter of interest (here: κ), but in another parameter (here: β)? We can repeat the previous analysis but now simulate variability (of increasing amount) in β instead of κ (note that variability in β would mean that the utility difference between choice options is mapped onto the decision in a more/less deterministic way depending on whether β_t is high/low in a trial; for instance, choosing the immediate option in a trial in which the delayed option is very attractive could be attributed to a low β_t value instead of a high κ_t value). As can be seen in Figure 5B, the improvement in

goodness-of-fit when trying to explain variability in β by modeling variability in κ via LOTO does not exceed the improvement under the null hypothesis, and there is also no systematic relationship with the amount of variability in β. Thus, we can conclude that finding a significantly higher improvement in fit in our data compared to simulated data under the null hypothesis cannot be due to a misattribution of variability in β to variability in κ (of course, this conclusion does not have to hold for other models, parameter sets, and tasks; the test has to be conducted for every different situation anew).

VII. Inferring trial-by-trial variability in more than one parameter

An advantageous feature of LOTO is that the duration of applying the method is almost independent from the number of parameters for which trial-by-trial variability is inferred. The full model is still estimated once for all *n* trials, and LOTO is still applied *n* times (but the estimation might take a bit longer for multiple parameters). In this section, we illustrate the application of LOTO to inferring trial-by-trial variability in two parameters of the HD-LBA model: the discount factor κ and the decision threshold *b* (for details, see Appendix A).

First of all, we used the non-parametric test from the previous section to check for the presence of systematic variability. Thereto, we simulated 1’000 sets of 30 participants under the null hypothesis and tested whether the simulation with true variability in κ and *b* provided a significantly higher improvement of the deviance than those simulations. This could be confirmed, as only 19 of the 1’000 simulations under the null hypothesis provided a better fit. It should be noted that the explanatory power of this test is somewhat limited, because we can only conclude that there is systematic variability in at least one of the two parameters but not necessarily in both of them.

Next, we looked at the correlations between the true κ_t and b_t parameter values and their LOTO estimates for an example participant. As shown in the upper left panel of Figure

6A, LOTO is still capable of recovering the variability in parameter κ from trial to trial for both types of decisions (i.e., immediate, delayed). Similarly, LOTO captures the variability in parameter *b* (lower right panel). The lower left / upper right panels in Figure 6A show the correlations between the true κ_t / b_t with the "wrong" LOTO estimates of b_t / κ_t. Ideally, one would desire correlations close to zero here, that is, LOTO should not attribute variability in one parameter to variability in another parameter. However, we see that there are modest correlations whose sign depend on the choice. When the immediate option was chosen, the correlations are negative; when the delayed option was chosen, the correlations are positive. These results are confirmed when looking at the average correlations across all 30 simulated participants (Figure 6B): LOTO attributes variability in the "correct" parameters independent of choice and variability in the "wrong" parameters depending on which option was chosen (intuitively, a high discount factor κ_t competes with the decision threshold b_t as an alternative explanation for fast choices of the immediate / slow choices of the delayed option; similarly, a high decision threshold b_t competes with the discount factor κ_t as an alternative explanation for choices of the option with highest utility).

Importantly, these results indicate that the observations (i.e., choices and RT) could explain away the undesired correlations while not affecting the desired correlations. To test this, we conducted random-effects linear regression analyses with the true parameter (κ_t, b_t) as dependent variable, and the LOTO estimates of these parameters together with choices and RT as independent variables (Figure 6C). Indeed, the standardized regression coefficient (averaged across simulated participants) for LOTO's estimate of κ_t when explaining variance in true κ_t was significantly positive with a very high effect size ($t(29) = 28.48$, $p < .001$, $d = 5.20$). The coefficient for the "wrong" estimate of b_t when explaining variance in κ_t also reached significance, but the effect was much weaker ($t(29) = 3.01$, $p = .005$, $d = 0.55$). With respect to explaining variance in true b_t, only LOTO's estimate of the "correct" estimate of b_t

was significant and strong ($t(29) = 30.40$, $p < .001$, $d = 5.55$), while the effect for the "wrong" estimate of κ_t did not reach significance ($t(29) = -1.93$, $p = .063$, $d = -0.45$).

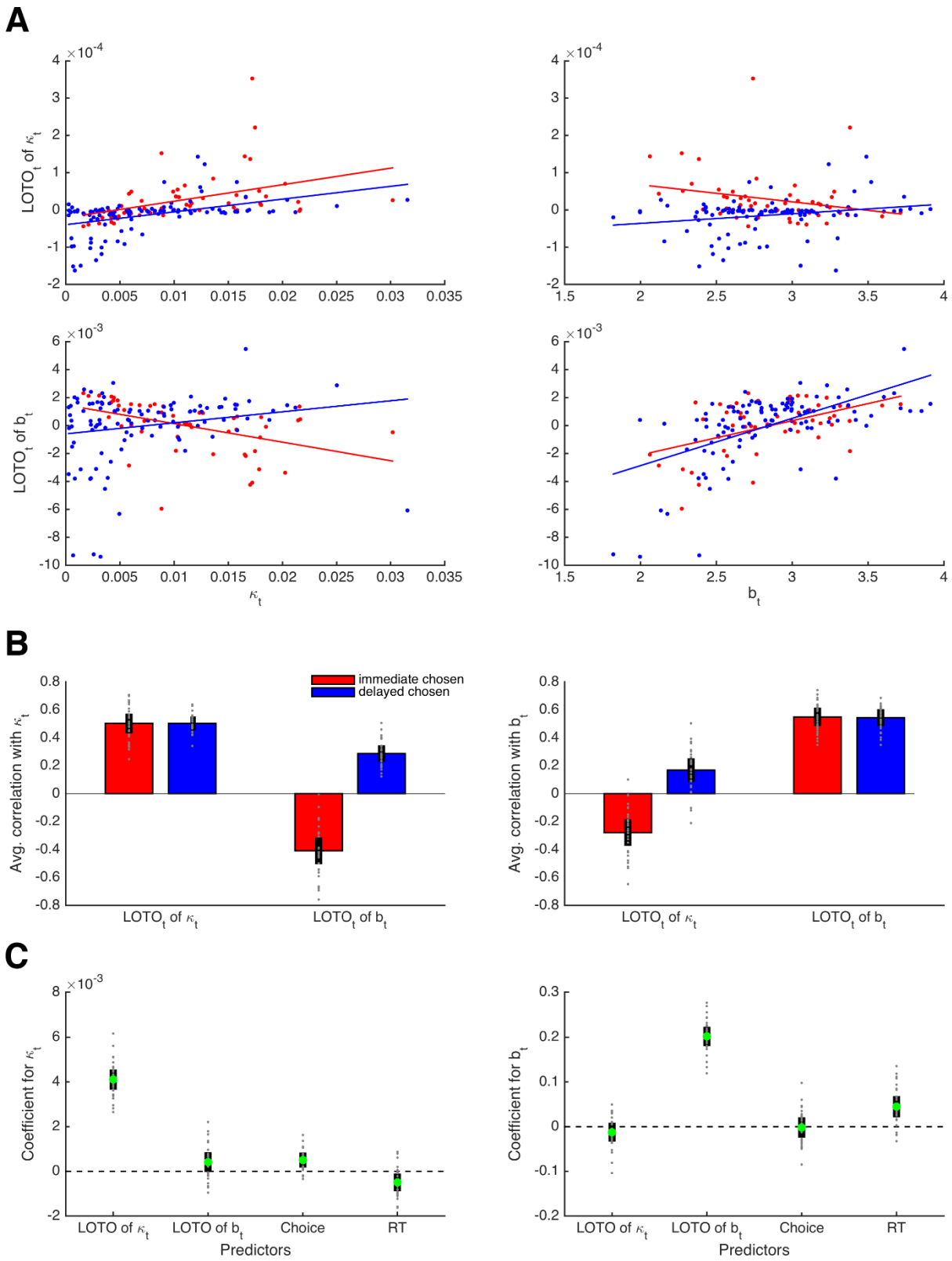

Figure 6. Capturing trial-by-trial variability in two parameters. **A**) Correlations between true parameter values κ_t and b_t with their LOTO estimates for an example participant; the on-diagonal elements show the "correct" inferences, the off-diagonal elements show the "wrong" inferences. **B**) Average correlations for all simulated participants (left panel: correlation with κ_t; right panel: correlation with b_t). **C**) Average standardized regression coefficients of LOTO estimates, choices, and RT when explaining variance in κ_t (left panel) and b_t (right panel).

Taken together, LOTO allows to infer trial-by-trial variability in more than one parameter within the same model over and above what the observations already tell us. At the same time, LOTO's misattributions of variability to other parameters are largely explained away by the observations (as before, these conclusions are restricted to the current setting of task, model, and parameters, and have to be tested for different settings anew). Notably, the random-effects linear regression analysis conducted above is very similar to the conventional approach of analyzing fMRI data on the basis of the General Linear Model (Friston et al., 1994). The ultimate question is not whether LOTO induces some significant misattribution or not, but whether this misattribution can in turn be expected to lead to wrong inferences on the relationship between brain activity and variability in cognitive parameters. We further elaborate on this issue in Section X.

VIII. The performance of LOTO as a function of the number of trials

Trial numbers for neuroimaging studies are often low. fMRI studies are particularly limited, because scanning hours are expensive, and because the low temporal resolution of fMRI requires long inter-trial intervals in the case of event-related fMRI (which is most suitable for trial-by-trial analyses). Therefore, it is important to understand to what extent LOTO's accuracy depends on the number of trials. To test this, we repeated the simulations of the HD and the HD-LBA models for 20, 40, 80, 160, 320, and 640 trials. As shown in Figure 7A, the average height of the correlations between the true parameter values and the respective estimates of LOTO are largely unaffected by the number of trials. Only in the case of 20 trials, the recovery performance is slightly reduced for the threshold parameter *b* of the HD-

LBA model. Note, however, that the variance of the correlations decreases as the number of trials increases. Therefore, LOTO still benefits from increasing the number of trials in terms of achieving higher statistical power, as we show in Section X.

In Section III, we mentioned that the impact of a single trial on the overall model fit as well as on the LOTO estimate decreases as the number of trials increases. If the number of trials is very high, then the improvement in model fit by a change in the parameter value might become smaller than the default tolerance criterion set by the minimization algorithm. In other words, the algorithm assumes convergence because the change in model fit due to taking one trial out is insignificantly small. Figure 7B illustrates this phenomenon. The upper panels show the results for parameter κ of the HD model for 160 trials, the lower panels for 1'600 trials. In the left panels, a strict tolerance criterion was used, in the right panels, the default tolerance criterion in MATLAB was used (for details, see Appendix A). In the case of 1'600 trials, the tolerance criterion matters: If the default criterion is used, the LOTO estimates exhibit discrete jumps, that is, the same parameter value is estimated for trials with relatively similar observations, because a more fine-graded estimation does not improve the fit beyond the default tolerance criterion. This is not the case for the strict tolerance criterion.

According to the example in Figure 7B, the default tolerance criterion appears to lower the correlation between the true values and the LOTO estimates in the case of 1600 trials, but only to a minor extent. This observation is confirmed when comparing the correlations across all 30 simulated participants. In the case of 160 trials, the correlations are not significantly different between the strict and the default tolerance criterion ($t(29) = 1.27$, $p = .213$, $d = 0.23$). In the case of 1600 trials, there is a significant difference ($t(29) = 4.59$, $p < .001$, $d = 0.84$), but the difference is small in absolute terms (i.e., on average the correlation is reduced only by .003). In general, we recommend checking the potential impact of the tolerance criterion on LOTO's performance by running simulations, in particular when LOTO is applied to a task with many trials (i.e., ≥ 1000).

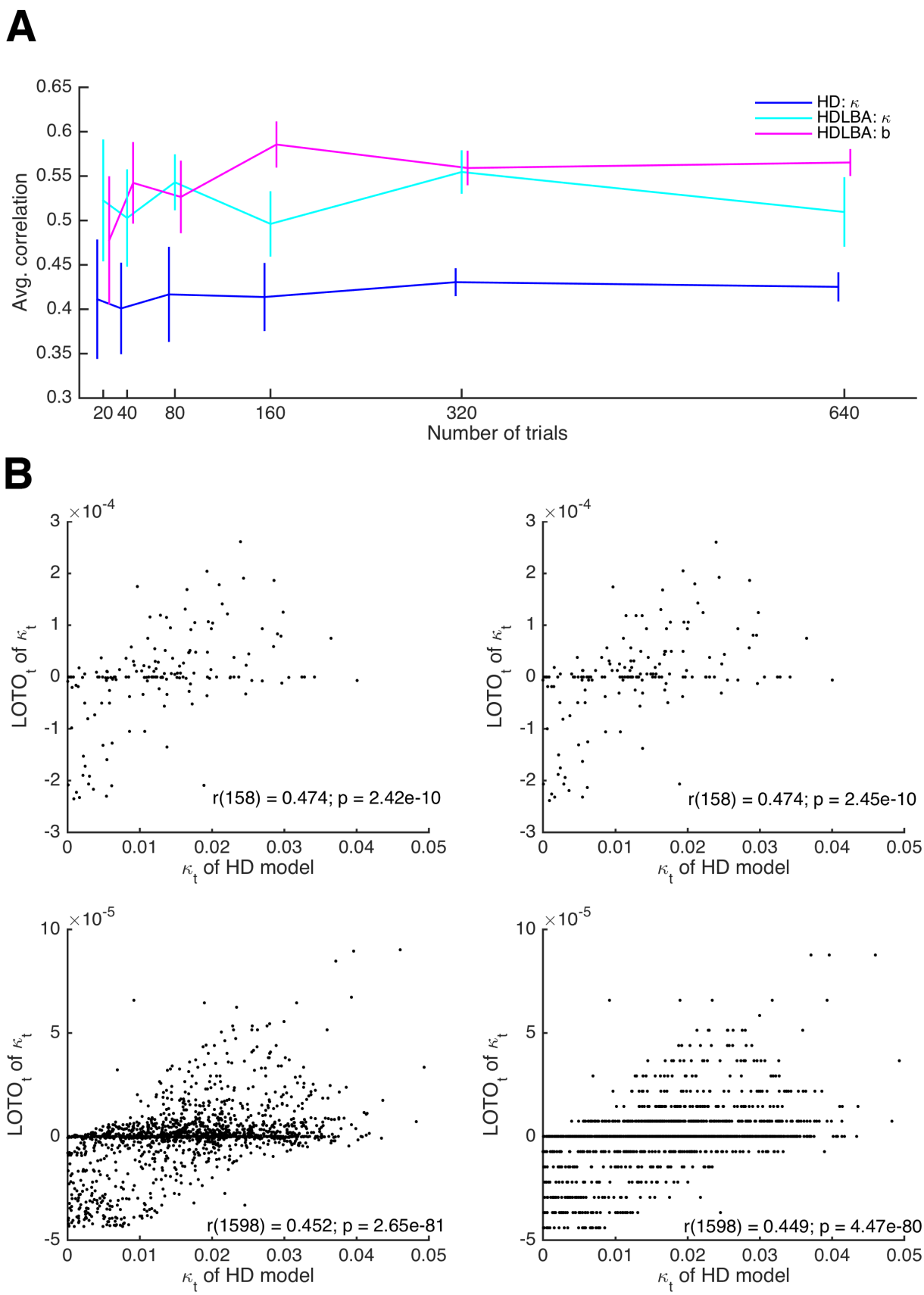

Figure 7. Relationship between the performance of LOTO and the number of trials. **A**) Average correlations between true parameter values and their LOTO estimates for parameters of the HD and the HD-LBA model as a function of the number of trials per simulated participant. **B**) Example correlations between true values and LOTO estimates for 160 trials (upper panels)

and 1'600 trials (lower panels), separately for a strict tolerance criterion for convergence of the minimization algorithm (left panels) and a more lenient, default tolerance criterion (right panels).

IX. Comparison of LOTO with related approaches

In this section, we compare LOTO's ability to recover the underlying trial-by-trial variability of cognitive model parameters with alternative techniques. More specifically, we generate synthetic data (i.e., choices and RT) from the LBA model (Brown and Heathcote, 2008), which assumes that the trial-wise drift rates ν_t and start points A_t are drawn from normal and uniform distributions, respectively (for details, see Appendix A). We compare LOTO with three other methods: i.) with our previously proposed technique (Gluth and Rieskamp, 2017; henceforth "GR17"; see the introduction, for a short description of GR17), ii.) with "single-trial fitting" (see Section IV), and iii.) with the "Single-Trial LBA" approach (van Maanen et al., 2011; henceforth "STLBA") which was specifically developed for retrieving maximum-likelihood estimates of trial-wise drift rates and start points of the LBA.

Figure 8 shows correlations between the true drift rates and start points and the recovered values by each method for an example simulation (upper panels) together with average correlations for multiple simulated experiments (lowest panels). From the example correlations, we see that LOTO and GR17 yield very similar results. This is to be expected, because the two approaches employ similar principles: In both cases, parameters are estimated for all trials, and those parameters for which trial-by-trial variability is not assumed are fixed to their all-trial estimates; then, both approaches use the discrepancy between the observation in each trial and the observations over all trials to infer plausible trial-specific values. However, the comparison across multiple simulations reveals that LOTO performs significantly better than GR17 for both parameters (drift rate: $t(19) = 3.76$, $p = .001$, $d = 0.84$; start point: $t(19) = 4.86$, $p < .001$, $d = 1.09$). The reason for this is that the performance of GR17 depends on the step size with which the range of possible parameter values is scanned,

and we were forced to choose an immediate step size to avoid too high computation times (Appendix A). As a consequence, the GR17 estimates exhibit discrete jumps (similar to those mentioned in the previous section when applying LOTO with a too lenient tolerance criterion), which lower the correlation with the true parameters. In other words, the advantage of LOTO over GR17 is its higher processing speed which is particularly relevant when modeling variability in multiple parameters.

The "single-trial fitting" approach yields acceptable results for the drift rate, but is still outperformed by LOTO ($t(19) = 19.72$, $p < .001$, $d = 4.41$). More dramatically, the method fails completely with respect to recovering the start point values. In our view, this is because adjusting the drift rate is sufficient to account for the observations in a single trial (i.e., whether the choice was correct or not and how fast the choice was), so that there is no information left to be captured by the start point. In contrast, LOTO (and GR17) estimates depend on how the likelihood space of the LBA model changes when one trial is taken out, and this likelihood space depends on both parameters.

Finally, the STLBA method performs worse than LOTO with respect to the drift rate ($t(19) = 52.60$, $p < .001$, $d = 11.76$) but better with respect to the start point ($t(19) = -30.08$, $p < .001$, $d = -6.73$). Looking at the example simulations in Figure 8, we see that this is because many single-trial drift rates (but not start points) of the STLBA are equal to the "all-trial estimate". This is a consequence of LBA's assumption that drift rates are drawn from a normal distribution whereas start points are drawn from a uniform distribution, which induces a "ridge" in the joint density distribution on which one drift rate value (namely the mean of the normal distribution) together with many start points are the maximum-likelihood estimates (L. van Maanen; personal communication, 02/2018). Overall, LOTO appears to provide comparatively good and in fact mostly superior results compared to alternative approaches of capturing trial-by-trial variability in cognitive model parameters.

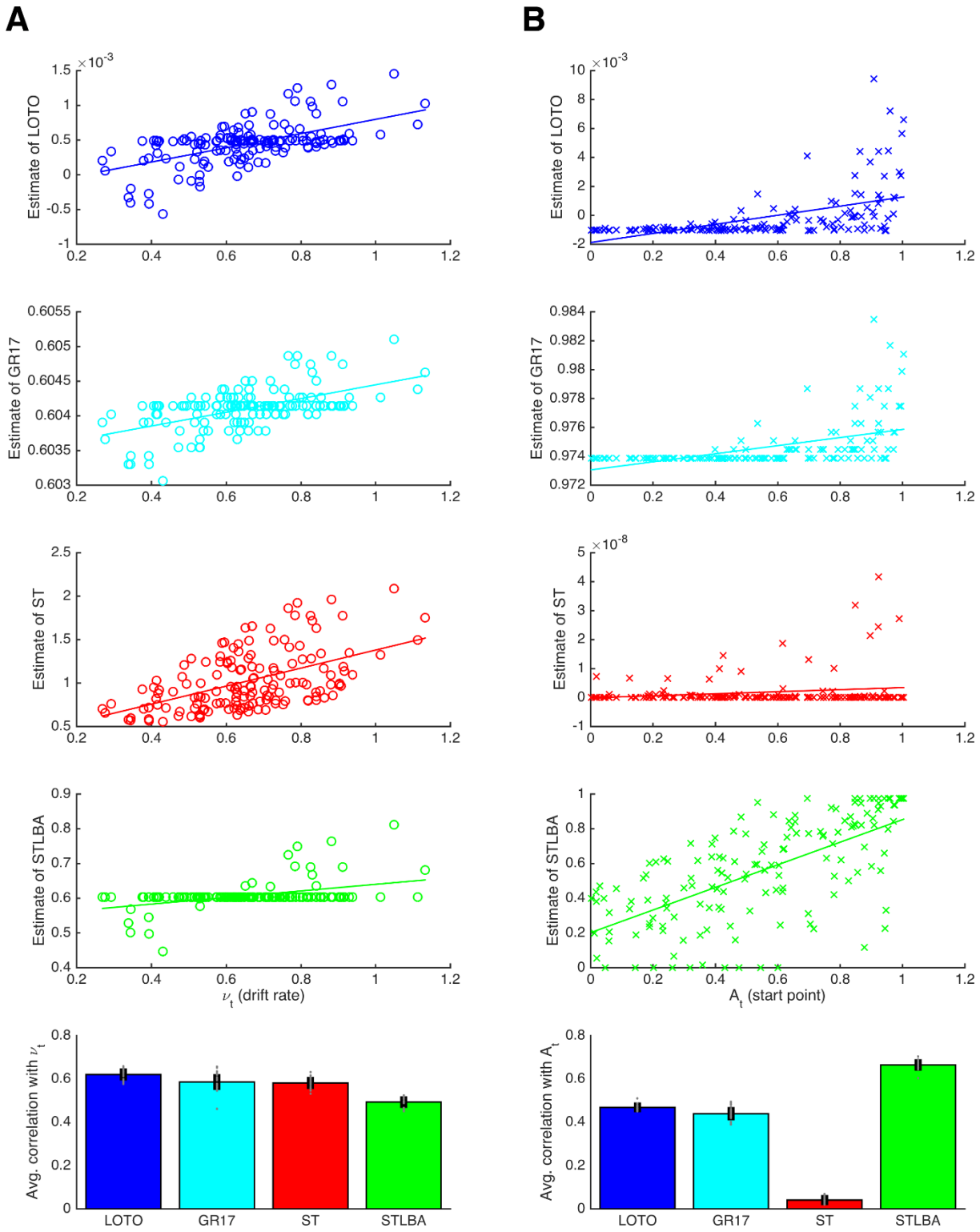

Figure 8. Comparison of methods for capturing trial-wise parameter values. **A**) Results for the LBA drift rate parameter. The first four panels show correlations of an example participant for the methods LOTO (blue), GR17 (cyan), single-trial fitting (ST; red), and STLBA (green). The lowest panel show correlations averaged over all simulations. **B**) Same as in **A**) but for the LBA start point parameter.

X. A power analysis for linking LOTO estimates to neural data

The ultimate goal of LOTO (and related approaches) is to establish a link between the inferred trial-by-trial variability in parameters of cognitive models and the brain activity measured with tools such as fMRI or EEG. Therefore, the question of whether LOTO is an appropriate tool for a given set of experimental design (including task, number of trials, number of participants etc.), cognitive model, and parameter(s) of interest should be answered by testing whether LOTO can be expected to recover the (truly existing) correlation between a neural signal and a parameter. In the case of modeling variability in more than one parameter, it is also important to test whether the potential misattribution of variability by LOTO is not expected to lead to incorrect inferences regarding the brain-parameter relationship (see Section VII). We emphasize here that these requirements are not unique to LOTO. Any approach of investigating the neural correlates of trial-by-trial parameter variability should be accompanied by such analyses of statistical power (i.e., sensitivity) and false-positive rates (i.e., specificity).

To link LOTO estimates to neural data, we reverted to the HD-LBA model for which we have inferred variability in two parameters (i.e., discount factor κ and decision threshold b) in Section VII. When simulating trial-specific parameter values κ_t and b_t as well as choices and RT, we also generated two hypothetical neural signals $F(\kappa)_t$ and $F(b)_t$ that were positively correlated with κ_t and b_t, respectively (for details, see Appendix A). Then, we generated 1’000 datasets of 30 participants with 160 trials each and tested how often the LOTO estimate for a parameter (e.g., κ) correctly explains a significant amount of variability in the associated neural signal (e.g., $F[\kappa]$), and how often it incorrectly explains a significant amount of variability in the neural signal associated with the other parameter (e.g., $F[b]$) over and above the observations. This was done using a random-effects linear regression analysis – similar to the General Linear Model approach in fMRI analysis (Friston et al., 1994). Stated differently,

we tested the sensitivity of LOTO by means of a power analysis, asking how likely it is that LOTO identifies the correct neural signal, and the specificity of LOTO by means of a false-positive analysis, asking how likely it is that LOTO misattributes a neural signal to a different parameter.

Figure 9A depicts the histograms of *p*-values for testing whether LOTO's estimates for κ and *b* are significantly related to the brain signals $F(\kappa)_t$ (left panel) and $F(b)_t$ (right panel), using the setting with 30 participants. According to these results, LOTO achieves a statistical power (or sensitivity) of 67% for parameter *k* and of 76% for parameter *b*, values that are slightly below the threshold of 80% that is usually deemed acceptable. On the other hand, the false-positive rate (or specificity) of LOTO is good with values lying just slightly above the 5% error rate. This means that the small misattribution of parameter variability seen in Section VII does not carry over much to a misattribution of the neural signal. In Figure 9B, we repeated the analysis for a sample size that is increased by 2/3 (i.e., 50 instead of 30 simulated participants). Expectedly, the statistical power of LOTO improves and lies in acceptable ranges. An alternative solution is to increase the number of trials per participant. Figure 9C depicts the results when the number of participants is kept to 30 but the number of trials in increased by 2/3 (i.e., 267 instead of 160 trials). The improvement of statistical power is comparable to that of the increased sample size. Accordingly, one would conclude from these analyses that investigating the neural basis of trial-by-trial variability in parameters κ and *b* of the HD-LBA model might require testing more than 30 participants or testing more than 160 trials per participant, whatever is more suitable in a given context (e.g., adding trials might be difficult for event-related fMRI designs that require long inter-trial intervals but less problematic for EEG designs).

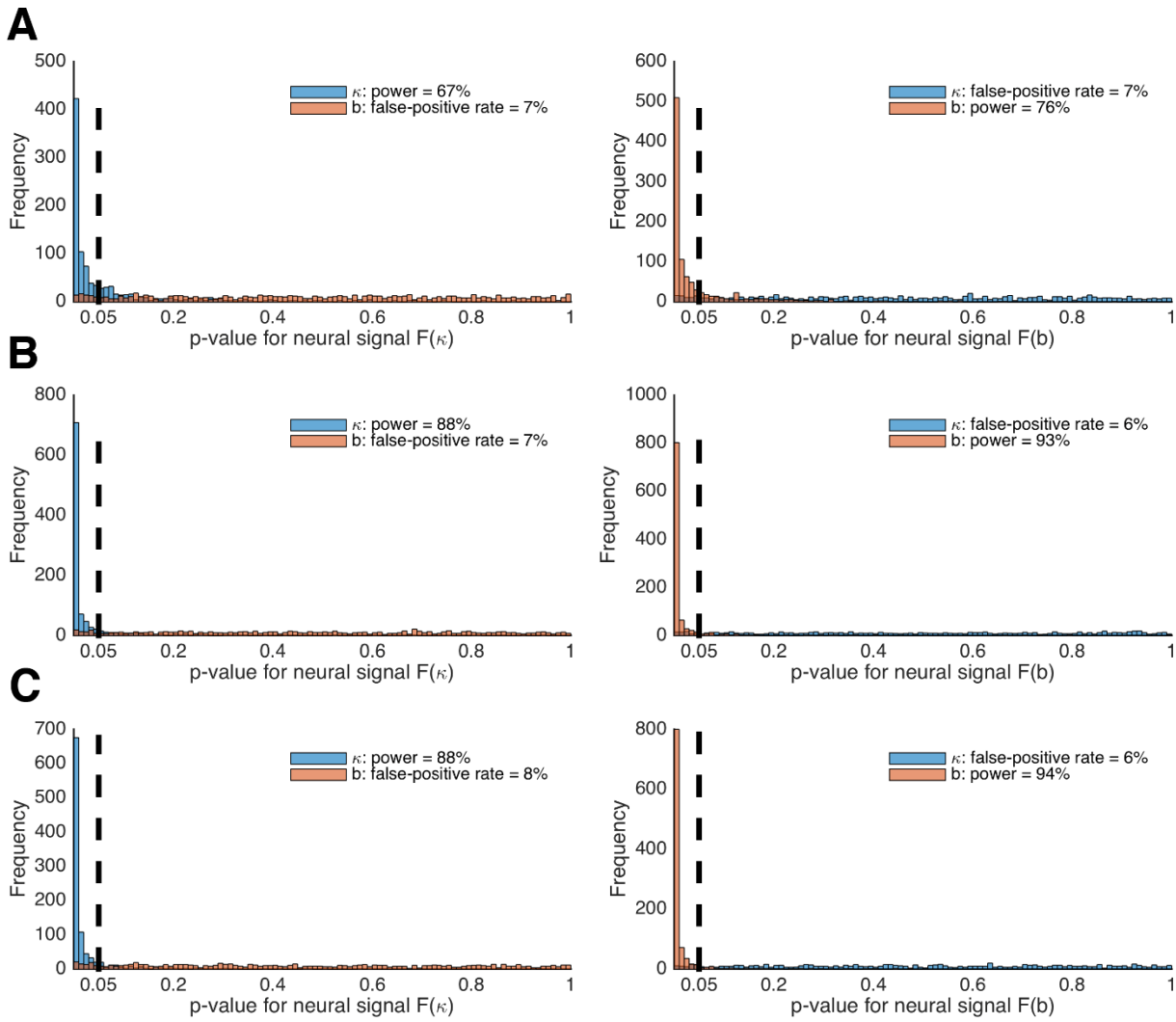

Figure 9. Power analysis of LOTO for detecting brain-parameter correlations. **A**) Histogram of *p*-values for testing significant relationships between κ and *b* and brain signals *F*(κ) (left panel) and *F*(*b*) (right panel). **B**) The same as in **A**) but for 50 instead of 30 participants per simulated experiment. **C**) The same as in **A**) but for 267 instead of 160 trials per participant.

XI. An example of using LOTO for model-based fMRI

In this section, we apply LOTO to a real neuroimaging dataset to further illustrate the method's use and usability in model-based cognitive neuroscience. The dataset is taken from a previously published fMRI study (Gluth et al., 2015), in which participants were asked to make preferential choices from memory. More specifically, participants first encoded associations between food snacks and screen positions and then decided between pairs of food snacks while seeing only the snacks' screen positions (Figure 10A). Participants had either

one or five seconds to retrieve the item-location associations and to deliberate upon the preferred snack option.

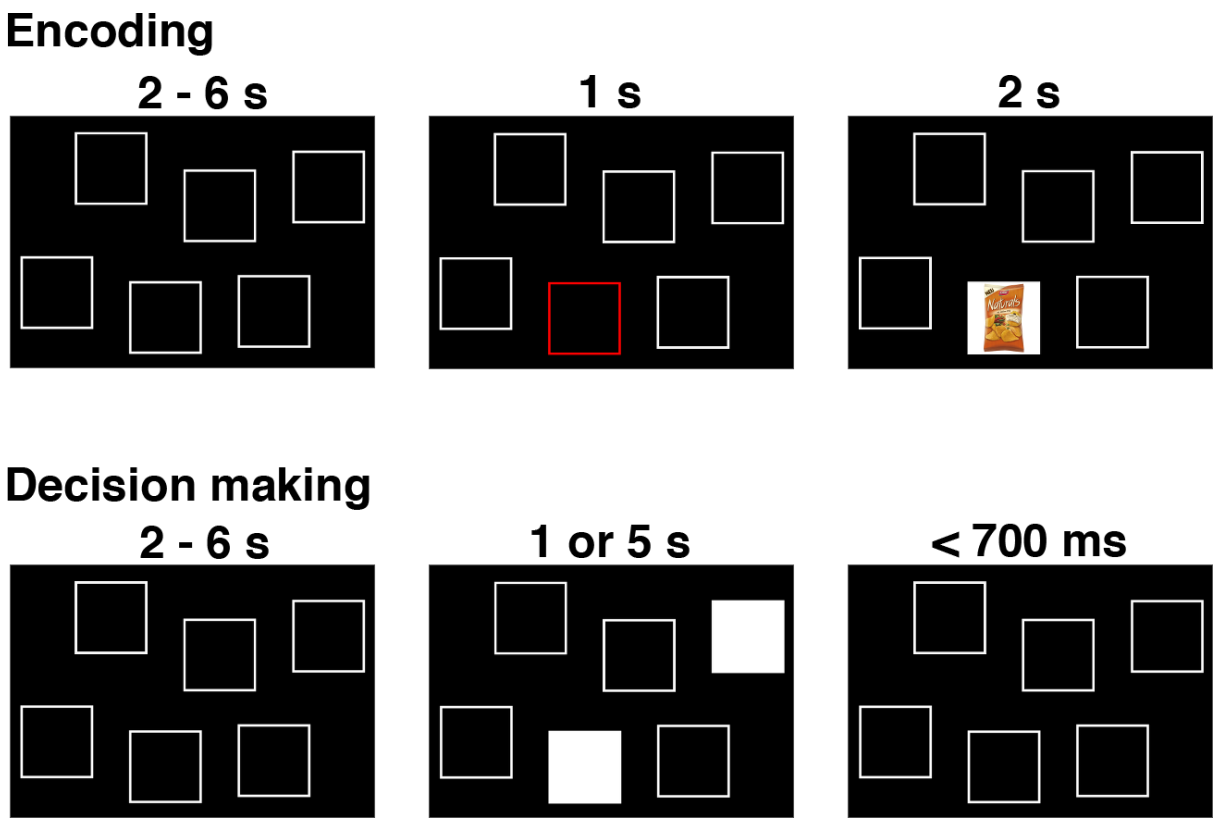

Figure 10. The remember-and-decide task (Gluth et al., 2015). This figure shows one example trial from the encoding and decision-making phases of the task. During an encoding trial, participants learned the association between one out of six different snacks with a specific screen position. During a decision-making trial, two screen positions were highlighted, and participants had to retrieve the two associated snacks from memory to choose their preferred option. Here, we apply LOTO to a decision-making model in order to infer memory-related activity during encoding. For a more extensive illustration of the study design, see Figure 1 in Gluth et al. (2015).

Choice behavior was modeled using an unbounded sequential sampling model which predicted more accurate decisions (i.e., decisions that are more consistent with evaluations of the food snacks in a separate task) with more deliberation time (details on the model and the fMRI analyses are provided in Appendix D; see also Figure 3 in Gluth et al., 2015). Importantly, the model estimates the average memory performance of a participant, that is, the probability of accurately remembering a snack when making decisions. In the original

study (Gluth et al., 2015), we employed our previously proposed GR2017 method for estimating trial-by-trial variability to infer the likelihood of successful memory retrieval for each snack given the choice behavior (intuitively, choosing unattractive snacks is attributed to poor memory performance). Here, we repeat this analysis but use LOTO to make the inference. Notably, LOTO takes advantage of the fact that each snack is a choice option in multiple (i.e., three) choice trials (see Section III). Ultimately, LOTO's item-specific estimate of memory performance is applied to the fMRI analysis of the encoding phase. In other words, a model-based analysis of the well-established *difference in subsequent memory effect* (Paller and Wagner, 2002) is conducted with the prediction that variations in memory performance should be positively correlated with activation of the hippocampus (as has been found in our original publication).

Before using LOTO to inform the fMRI analysis, we checked whether the variability estimates could be linked to other behavioral indicators of memory performance and whether the assumption of systematic variability in memory could be justified. To address the first question, we conducted a logistic regression testing whether the LOTO-based estimate predicted correct responses in the cued recall phase (i.e., a phase following the decision-making phase in which participants were asked to recall each snack; see Appendix D and Gluth et al., 2015). As additional independent variables, this regression analysis also included whether the snack was encoded once or twice, how often a snack was chosen, and the subjective value of the snack. As can be seen in Figure 11A, the LOTO-based estimate of memory explained variance in cued recall performance over and above the other predictor variables ($t(29) = 7.44$, $p < .001$, $d = 1.36$). To address the question of systematic variability, we followed the procedures described in Section VI and generated 1'000 datasets under the null hypothesis of no variability in memory performance and compared the LOTO-based improvement in model fit in these datasets with the improvement in the real data. Indeed, the

improvement in the real data was much higher than in any of the datasets generated under the null hypothesis (Figure 11B).

Having established that LOTO has the potential to identify meaningful neural correlates of item-specific memory performance, we turned to the fMRI analysis. Here, we employed a frequent approach to model-based fMRI (O'Doherty et al., 2007), which is to add the predictions of a cognitive model as a so-called parametric modulator (PM) of the hemodynamic response at a time window of interest. In our case, this time window is the presentation of snack food items during encoding with the LOTO-based variability estimates serving as the PM (in order to test for a model-based subsequent memory effect; see above). As can be seen in Figure 11C, this analysis revealed fMRI signals in bilateral hippocampus (MNI coordinates of peak voxel in left hippocampus: x = -18, y = -10, z = -12; right hippocampus: 38, -12, -26) which surpassed the statistical threshold corrected for multiple comparisons (left: $Z = 3.73$; $p = .033$; right: $Z = 4.46$, $p = .002$). Remarkably, these effects were even stronger than those reported in the original publication, which were based on the GR2017 method (note that as in the original publication, we ruled out alternative explanations for these results by including cued recall performance and subject values of snacks as additional PMs in the analysis; see Appendix D).

To illustrate a second analysis approach, we also conducted a region-of-interest (ROI) analysis by extracting the average fMRI signal from anatomically-defined masks of the left and right hippocampi (Figure 11D). The average signal in these ROIs were significantly positive (left: $t(29) = 2.93$, $p = .007$, $d = 0.53$; right: $t(29) = 4.36$, $p < .001$, $d = 0.80$). Similar to the previous analysis, the effects of this ROI analysis were stronger than those reported in the original publication, in which only the right hippocampus ROI reached significance. Taken together, we could reproduce the analysis of a model-based subsequent memory effect reported in Gluth et al. (2015) and obtained even superior results with LOTO.

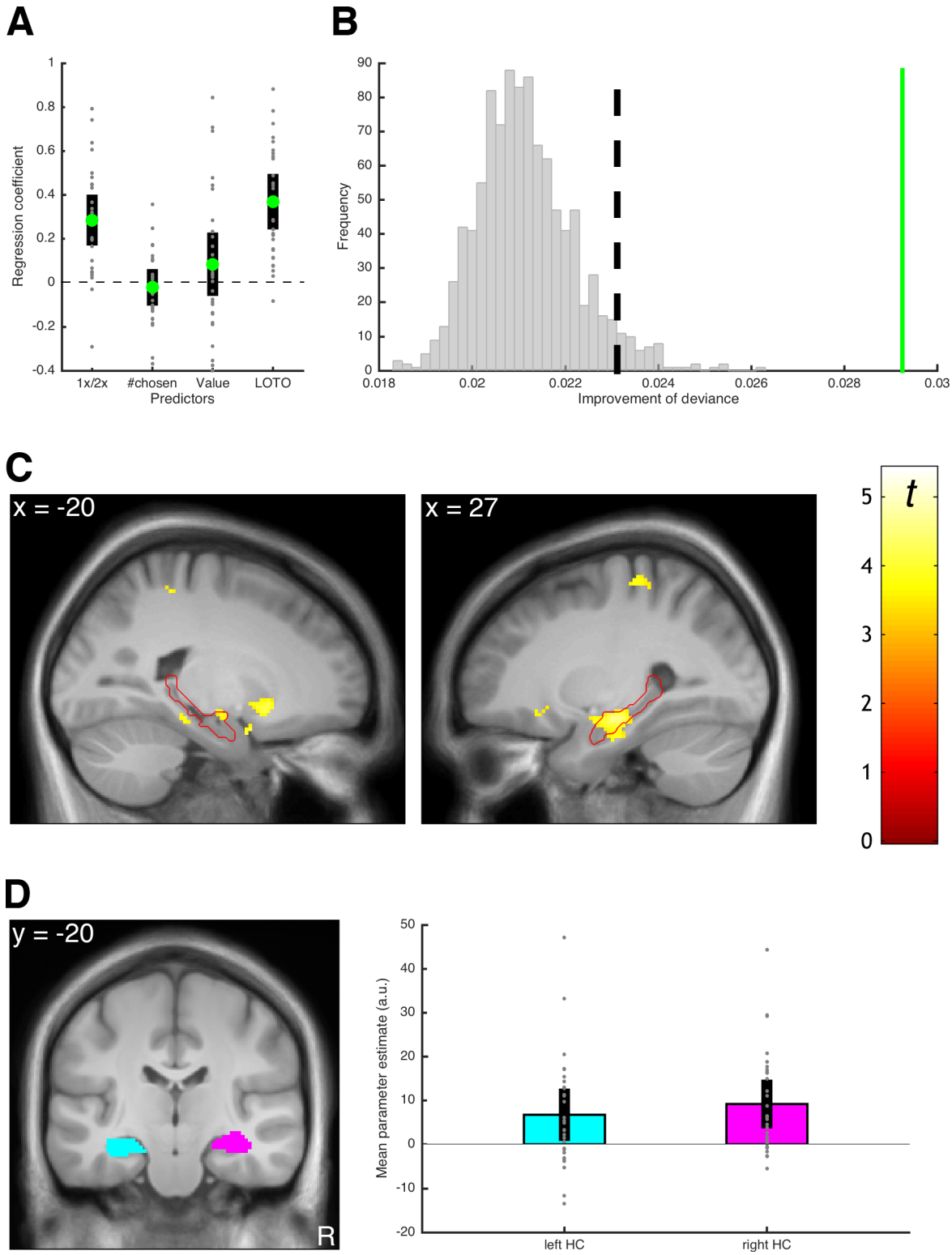

Figure 11. Results of the LOTO example application. **A**) Average standardized regression coefficients for predicting memory performance in cued recall (1x/2x refers to how often an item was encoded). **B**) Testing for systematic variability in memory; the green line shows the average trial-wise improvement of fit for the real dataset. The vertical dashed line represents the significance threshold based on the simulated data. **C**) fMRI signals linked to the LOTO-based estimate of item-specific memory; the red outline depicts the anatomical mask of the hippocampus. **D**) ROI analysis with anatomical masks of left and right hippocampus (HC) and average fMRI signals in these ROIs (the analyses in **C** and **D** are analogous to those in Figure 5 in Gluth et al., 2015).

Discussion

In this work, we proposed LOTO, a simple yet efficient technique to capture trial-by-trial measures of parameters of cognitive models for relating those measures to neural or physiological data. The core principle is to infer single-trial values from the difference in the parameter estimates when fitting a model with all trials vs. with one trial being left out. The application of LOTO is simple: First, the full set of parameters of a model is estimated for all n trials; then, the parameters of interest are estimated again n times with each trial being left out once, while the remaining parameters are fixed to their all-trial estimates. In addition to its simplicity, LOTO is also comparatively fast: If fitting a model once to all trials takes u time units, then LOTO's runtime will be $u \times n$ at most (in fact, it can be expected to be even faster because not all parameters are estimated again). And in contrast to the GR2017 approach, this runtime will not increase (much) if variability in more than one parameter is estimated. Compared to related methods, LOTO appears to provide either comparable or superior results. Finally, it is a very general approach to capturing variability in parameters. In principle, it can be applied to any cognitive model, although we stress the importance of testing the feasibility of LOTO for each model and study design anew.

The LOTO recipe

In the following, we provide a "LOTO recipe" that summarizes the critical steps, tests and checks that one should follow when seeking to apply the method. We link each point to the relevant section(s) in the paper. Thus, despite our recommendation that readers with limited statistical and mathematical knowledge should first consult this recipe, it is indispensable for interested cognitive neuroscientists to read the relevant sections of the article before trying to apply LOTO to their own data. Notably, Steps 1 and 2 of this recipe are to be taken prior to conducting the experiment. Therefore, LOTO is not suitable for post-hoc analyses of data that have already been acquired with different study goals in mind. Also note that the computer

codes associated with this article and LOTO are freely available on the Open Science Framework (https://osf.io/du85f/).

1. Specify the cognitive model and the parameters for which trial-by-trial variability should be estimated; similarly, specify the design of the study (number of trials and participants, experimental manipulations, neuroscientific method).
    a. What type of observations does the model explain? If it explains only choices, try to extend the model to allow predicting RT (Section V).
    b. How are the experimental manipulations or conditions related to the model (i.e., which manipulation is predicted to selectively influence a given parameter)? If the model has to be fit to each condition separately, try to adjust the model so that its predictions change directly as a function of the condition (Section IV).
    c. Should variability be captured in each trial or across multiple trials (e.g., small blocks of trials)? If simulations of LOTO (see Step 2 below) do not yield acceptable results, consider to chunk multiple trials into small blocks and try to capture parameter variability from block to block (Section III).
    d. Does the neuroscientific method allow measuring signals that can be related to trial-by-trial variability? Block-designs for fMRI and ERPs for EEG are less suitable than event-related designs for fMRI and time-frequency analyses for EEG.
2. Test the feasibility of the LOTO approach for the chosen model and study design.
    a. Test whether LOTO can (in principle) provide sensible estimates of the true trial-specific values of the targeted model parameter(s). Specifying the first derivative of the log-likelihood and the Fisher information (as done in Sections III and IV) is desirable but often cumbersome or even impossible. Instead, we suggest running simulations (Sections III – V).

b. Check whether LOTO might misattribute random noise by conducting the non-parametric test for presence of systematic variability (Section VI). Similarly, check whether LOTO might misattribute variability in other parameters to the parameters of interest by conducting the regression analyses outlined in Section VII.

c. Estimate the statistical power and if applicable the specificity of LOTO given the model and study design (Section X). This might require running a pilot study to obtain a rough estimate of the amount of trial-by-trial variability in model parameters that can be expected (alternatively, existing data sets could be inspected for this purpose).

d. For experiments with many trials per participant (i.e., ≥ 1’000), check whether the default tolerance criterion of the parameter estimation method is sufficient or whether there are discrete “jumps” in the LOTO estimates. If necessary, choose a stricter criterion (section VIII).

e. Always use the actually planned number of trials, number of participants and manipulations when doing simulations. If the simulations (see Steps 2a,b) do not yield acceptable results, consider increasing the number of participants or the number of trials per participant (Section X), or consider adapting the model or experimental paradigm (Sections III – V).

3. After conducting the study, apply LOTO (Section II), conduct the non-parametric test for the presence of systematic variability (Section VI), and use LOTO’s estimate for analyzing the recorded neural data (e.g., Section XI).

Of note, the test for the presence of variability (Section VI; Steps 2b and 3) and the power analyses (Section X; Step 2c) are very time-consuming, because they require simulating a large number (e.g., 1’000) of entire experiments and applying LOTO to it. For example, for the first power analysis reported in Section X, we ran 1’000 simulations of 30 participants with 160 trials each, meaning that the HD-LBA model had to be estimated ~5 million times.

Therefore, it may not be possible to conduct an extensive search for optimal sample size and trial number to meet a desired statistical power such as 80%. Instead, trying out a limited set of realistic scenarios appears most appropriate. Note, however, that other approaches such as GR17 or hierarchical Bayesian modeling can be much more time-consuming than LOTO, so that even the simplest simulation-based power analyses would become infeasible.

The principle of leaving one out

The idea of leaving one data point out, also termed the "jackknife" principle, is not new. In fact, the approach was proposed by statisticians in the 1950s (Quenouille, 1956; Tukey, 1958) as a way to estimate the variance and bias of estimators when dealing with small samples. From a technical point of view, LOTO bears obvious similarities to statistical methods such as leave-one-out cross-validation, which is often employed in classification-based analyses (Bishop, 2006), or the leave-one-subject-out approach, which is used to circumvent the non-independence error in statistical analyses of fMRI data (Esterman et al., 2010). Notably, the jackknife principle has also been adopted for inferring trial-specific onset latencies of ERPs (Miller et al., 1998; Stahl and Gibbons, 2004) as well as for quantifying trial-specific functional connectivity between different brain regions (Richter et al., 2015). Thus, a combination of LOTO for inferring variability in cognitive model parameters with the same principle used for identifying trial-specific brain activation patterns could be a promising approach for future research in model-based cognitive neuroscience.

LOTO as a two-stage approach to linking behavioral and neural data

As mentioned in the introduction, LOTO is one out of several approaches that have been put forward to link computational models of cognition and behavior with neural data. Turner and colleagues (Turner et al., 2017) proposed a taxonomy of these approaches based on whether a method i.) uses neural data to constrain the (analysis of the) cognitive model, ii.) uses the

cognitive model to constrain the (analysis of the) neural data, or iii.) whether the two types of data constrain each other in a reciprocal manner. Within this taxonomy, LOTO is best classified as belonging to the second class, that is, it infers trial-wise values of the cognitive model to then inform the analysis of neural data. In other words, LOTO is a "two-stage approach" to model-based cognitive neuroscience, in which modeling of behavior precedes the analysis of brain data (as exemplified in Section XI). This approach misses out on the advantage of reciprocity exhibited by the joint modeling approach (Turner et al., 2015). However, LOTO's strength lies in its ease, speed and generality of application. Moreover, it retains the flexibility of the two-stage approach with respect to identifying the neural correlates of trial-by-trial parameters: One does not need to have an a priori hypothesis of which brain region or EEG component might be related to variability in a model parameter. Instead, the variability estimate captured by LOTO can simply be plugged into, for instance, a GLM-based fMRI whole-brain analysis (Gluth and Rieskamp, 2017; O'Doherty et al., 2007; see also Figure 11C).

Concluding remarks

Today, neuroimaging techniques such as EEG or fMRI and psychophysiological methods such as eye-tracking allow us to shed light on the process of single actions. However, the development of cognitive models has only partially kept up with these advances, because the central tenet of cognitive modeling remains to be the identification of invariant features of behavior by averaging over repeated observations. We envision that LOTO has the potential to fill this gap, in particular because it will neither be restricted to a handful of specific models nor to a handful of experts in the field who are highly proficient in computational modeling.

References

Bishop, C.M. (2006). Pattern Recognition and Machine Learning (New York: Springer).

Brown, S.D., and Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. Cognit. Psychol. *57*, 153–178.

Esterman, M., Tamber-Rosenau, B.J., Chiu, Y.-C., and Yantis, S. (2010). Avoiding non-independence in fMRI data analysis: leave one subject out. NeuroImage *50*, 572–576.

Farrell, S., and Lewandowsky, S. (2018). Computational Modeling of Cognition and Behavior (New York, NY: Cambridge University Press).

Forstmann, B.U., Wagenmakers, E.-J., Eichele, T., Brown, S., and Serences, J.T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? Trends Cogn. Sci. *15*, 272–279.

Frederick, S., Loewenstein, G., and O'Donoghue, T. (2002). Time discounting and time preference: a critical review. J. Econ. Lit. *40*, 351–401.

Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., and Frackowiak, R.S. (1994). Statistical parametric maps in functional imaging: A general linear approach. Hum. Brain Mapp. *2*, 189–210.

Gluth, S., and Rieskamp, J. (2017). Variability in behavior that cognitive models do not explain can be linked to neuroimaging data. J. Math. Psychol. *76*, 104–116.

Gluth, S., Rieskamp, J., and Büchel, C. (2012). Deciding when to decide: time-variant sequential sampling models explain the emergence of value-based decisions in the human brain. J. Neurosci. *32*, 10686–10698.

Gluth, S., Sommer, T., Rieskamp, J., and Büchel, C. (2015). Effective connectivity between hippocampus and ventromedial prefrontal cortex controls preferential choices from memory. Neuron *86*, 1078–1090.

Gluth, S., Hotaling, J.M., and Rieskamp, J. (2017). The attraction effect modulates reward prediction errors and intertemporal choices. J. Neurosci. *37*, 371–382.

Hawkins, G.E., Mittner, M., Forstmann, B.U., and Heathcote, A. (2017). On the efficiency of neurally-informed cognitive models to identify latent cognitive states. J. Math. Psychol. *76*, 142–155.

Ly, A., Marsman, M., Verhagen, J., Grasman, R., and Wagenmakers, E.-J. (2017). A tutorial on Fisher information. ArXiv 1705.01064.

van Maanen, L., Brown, S.D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., and Forstmann, B.U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. J. Neurosci. Off. J. Soc. Neurosci. *31*, 17488–17495.

Mazur, J.E. (1987). An adjusting procedure for studying delayed reinforcement. In Quantitative Analyses of Behavior: Vol. 5. The Effect of Delay and of Intervening Events on

Reinforcement Value, (Hillsdale, NJ: Erlbaum), pp. 55–73.

Miller, J., Patterson, T., and Ulrich, R. (1998). Jackknife-based method for measuring LRP onset latency differences. Psychophysiology *35*, 99–115.

Nunez, M.D., Vandekerckhove, J., and Srinivasan, R. (2017). How attention influences perceptual decision making: Single-trial EEG correlates of drift-diffusion model parameters. J. Math. Psychol. *76*, 117–130.

O'Doherty, J.P., Hampton, A., and Kim, H. (2007). Model-based fMRI and its application to reward learning and decision making. Ann. N. Y. Acad. Sci. *1104*, 35–53.

Palestro, J.J., Bahg, G., Sederberg, P.B., Lu, Z.-L., Steyvers, M., and Turner, B.M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. J. Math. Psychol. *84*, 20–48.

Paller, K.A., and Wagner, A.D. (2002). Observing the transformation of experience into memory. Trends Cogn. Sci. *6*, 93–102.

Peters, J., Miedl, S.F., and Büchel, C. (2012). Formal comparison of dual-parameter temporal discounting models in controls and pathological gamblers. PloS One *7*, e47225.

Pleskac, T.J., and Busemeyer, J.R. (2010). Two-stage dynamic signal detection: a theory of choice, decision time, and confidence. Psychol. Rev. *117*, 864–901.

Quenouille, M.H. (1956). Notes on bias in estimation. Biometrika *43*, 353–360.

Ratcliff, R. (1978). A theory of memory retrieval. Psychol. Rev. *85*, 59–108.

van Ravenzwaaij, D., Provost, A., and Brown, S.D. (2017). A confirmatory approach for integrating neural and behavioral data into a single model. J. Math. Psychol. *76*, 131–141.

Richter, C.G., Thompson, W.H., Bosman, C.A., and Fries, P. (2015). A jackknife approach to quantifying single-trial correlation between covariance-based metrics undefined on a single-trial basis. NeuroImage *114*, 57–70.

Rodriguez, C.A., Turner, B.M., and McClure, S.M. (2014). Intertemporal choice as discounted value accumulation. PloS One *9*, e90138.

Rodriguez, C.A., Turner, B.M., Van Zandt, T., and McClure, S.M. (2015). The neural basis of value accumulation in intertemporal choice. Eur. J. Neurosci. *42*, 2179–2189.

Stahl, J., and Gibbons, H. (2004). The application of jackknife-based onset detection of lateralized readiness potential in correlative approaches. Psychophysiology *41*, 845–860.

Tukey, J.W. (1958). Bias and confidence in not quite large samples. Ann. Math. Stat. *29*, 614–623.

Turner, B.M., Forstmann, B.U., Wagenmakers, E.-J., Brown, S.D., Sederberg, P.B., and Steyvers, M. (2013). A Bayesian framework for simultaneously modeling neural and behavioral data. NeuroImage *72*, 193–206.

Turner, B.M., van Maanen, L., and Forstmann, B.U. (2015). Informing cognitive abstractions through neuroimaging: the neural drift diffusion model. Psychol. Rev. *122*, 312–336.

Turner, B.M., Forstmann, B.U., Love, B.C., Palmeri, T.J., and Van Maanen, L. (2017). Approaches to analysis in model-based cognitive neuroscience. J. Math. Psychol. *76*, 65–79.

Wiecki, T.V., Sofer, I., and Frank, M.J. (2013). HDDM: hierarchical Bayesian estimation of the drift-diffusion model in Python. Front. Neuroinformatics *7*.

Appendix A: Detailed specifications of the simulations

A1. Bernoulli and binomial distributions

We drew n = 300 values of θ_t from a uniform distribution over a range of .2 (e.g., [.4, .6]), and then generated one Bernoulli event per trial t using θ_t as the underlying probability. LOTO was then used to recover θ_t by estimating and subtracting $\theta_t - \theta_{\neg t}$, for every t according to Equation 2. This was done for every range from [0, .2] to [.8, 1] in steps of .01 (i.e., 80 times) and repeated 1'000 times. Pearson product-moment correlations between θ_t and $LOTO_t$ were calculated and averaged over the 1'000 repetitions.

In a second step, we generalized to the binomial distribution by assuming that θ remains stable over a limited number of trials m ($m < n$). The goal was then to recover θ_m with LOTO. We chose four different levels of m (i.e., $L(m) = \{1, 5, 10, 20\}$), but kept n constant (i.e., n = 300). Thus, the number of different values of θ_m decreased by n/m as m increased. Again, simulations were run for every range from [0, .2] to [.8, 1] in steps of .01 and repeated 1'000 times, LOTO was applied (Equation 5a), and correlations between θ_m and $LOTO_m$ were calculated and averaged over the 1'000 repetitions (for each level of m separately).

A2. Intertemporal choice sets and HD model

For the sake of plausibility, choice sets for the intertemporal choice task were created in rough correspondence to typical settings in fMRI studies (Peters et al., 2012). We chose to simulate 160 trials per "participant", which would result in a realistic total duration of 40 minutes of an fMRI experiment, in which one trial lasts ~15 seconds (e.g., Peters and Büchel, 2009). As often done in fMRI studies on intertemporal choice, the immediate choice option was identical in all trials (i.e., $d_i = 0$; $x_i = 20$), but the delay and amount of the delayed choice

options varied from trial to trial. In a first step, delays d_j were drawn uniformly from the set $S(d_j) = \{1, 7, 14, 30, 90, 180, 270, 360\}$, and amounts x_j were drawn from a uniform distribution with limits [20.5; 80], discretized to steps of 0.5. In a second step, this procedure was repeated until a choice set was found for which the utility of the immediate option u_i was higher than the utility of the delayed option u_j in ~50% of the trials, given the simulated participant's baseline parameter κ (see below). Specifically, a choice set was accepted when $u_i - u_j > 0$ in more than 45% but less than 55% of trials. Otherwise, a new choice set was created. This was repeated up to 1'000 times. If no choice set could be found, the criterion was relaxed by 10% (i.e., a choice set was now accepted when $u_i - u_j > 0$ in more than 40% but less than 60% of trials). This procedure was repeated until an appropriate choice set was found. Note that using such adaptive designs is very common and recommended for intertemporal choice tasks to avoid having participants who choose only immediate or only delayed options (Gluth et al., 2017; Koffarnus et al., 2017; Peters et al., 2012). In practice, adaptive designs require a pre-test to infer the individual discount rate κ.

Simulations of the HD model were also based on empirical data (Peters et al., 2012). For each simulated participant, we first determined "baseline" parameters κ and β (for the sake of simplicity, we omit the indicator variable for participants here and for all following notations) by drawing from log-normal distributions truncated to ranges [.001, .07] and [0.01, 10], respectively. Means μ and standard deviations σ of these log-normal distributions were $\{\mu_\kappa = -4.8, \sigma_\kappa = 1\}$ and $\{\mu_\kappa = -0.77, \sigma_\kappa = 0.71\}$, respectively, and were chosen so that the medians and interquartile ranges of generated values would match those reported in Peters et al. (2012). Log-normal distributions were used to account for the fact that distributions of empirical discount rates and sensitivity parameters are often right-skewed. If trial-by-trial variability in κ and β was assumed, then this was realized by drawing trial-specific values of

κ_t and β_t from a normal distribution (truncated to values ≥ 0) with means κ and β, and standard deviations 0.01 and 1, respectively. Otherwise, we assumed that $\kappa_t = \kappa$, $\beta_t = \beta$.

After specifying choice sets and trial-specific parameter values, choices were simulated and the parameters of the HD model were estimated for all trials. This was done in a two-step procedure starting with a grid search (with 25 different values per parameter) to find suitable starting values for the parameters that were then used in a constrained simplex minimization. Then, LOTO was applied by leaving each trial out once and estimating the parameter(s) of interest (i.e., either κ_t or β_t or both) again, using the "all-trials" parameter estimates as starting values. Similarly, single-trial fitting was conducted whenever necessary by fitting each trial individually, also using the "all-trials" estimates as starting values.

Unless stated otherwise, this procedure was conducted for 30 participants and repeated multiple times whenever necessary (e.g., when testing for the presence of systematic trial-by-trial variability in Section VI). The example participant shown in Figure 3 was taken from such a set of simulations. This participant had baseline parameters $\kappa = .0076$ and $\beta = 2.11$, but only κ_t varied from trial to trial. When testing for the presence of systematic trial-by-trial variability in Section VI, we generated 1'000 simulations of 30 participants under the null hypothesis and 10 simulations of 30 participants assuming 10, 20, 30, 40, 50, 60, 70, 80, 90, 100% variability in κ compared to the amount of variability used in our initial simulations (see above).

A3. HD-LBA model

Procedures for simulating the HD-LBA model were similar to those for the HD model. As the HD model, the HD-LBA model includes parameter κ, and values for κ and κ_t were generated in the same way as before. However, HD-LBA does not use the soft-max choice function (see Equation 7) to specify choice probabilities but the LBA model (Brown and

Heathcote, 2008) to specify joint choice and response time (RT) distributions. Thus, instead of the soft-max choice sensitivity parameter β, it includes parameter λ that scales the relationship between utility and drift rate (see Equations 9a and 9b), and four additional parameters from the LBA model: parameter s that models (unsystematic) trial-by-trial variability in the drift rates (i.e., drift rates for options i, j are drawn from independent normal distributions with means ν_i, ν_j and standard deviation s), parameter b that represents the decision threshold, parameter A that represents the height of the uninform distribution of the start point, and parameter t_0 that accounts for the non-decision time of the RT. Most of these parameters were kept identical for all simulated participants (i.e., $\lambda = 0.1$, $s = 0.3$, $A = 1$, $t_0 = 0.5$). Besides κ, only the decision threshold b varied across participants (because it was used to test how well LOTO captures trial-by-trial variability in two parameters simultaneously) by drawing individual values from a uniform distribution with limits [2, 3]. When testing LOTO on two parameters simultaneously, trial-specific values of κ_t were drawn as stated above, and trial-specific values of b_t were drawn from a normal distribution with mean b and standard deviation 0.04 (with the constraint that $b_t \geq$ A).

Parameter estimation again started with an initial grid search. To reduce computation time, grid search was extensive (i.e., 15 grid search values) for parameters κ and b but limited (i.e., 3 grid search values) for the remaining parameters that were identical across simulated participants. As for the HD model, best parameter estimates from grid search were used as starting values for the constrained simplex minimization, followed by applying LOTO and single-trial fitting.

A4. LBA model

We compared four different methods for capturing trial-by-trial variability with respect to how well these methods recover trial-wise drift rate and start point values of the

LBA model (Brown and Heathcote, 2008). We chose these two LBA parameters because one of the tested methods, the single-trial LBA (STLBA) approach (van Maanen et al., 2011), has been developed specifically for this set of parameters. The other three methods were LOTO, the "single-trial fitting" method (see Section IV), and an approach that we proposed in a previous publication (Gluth and Rieskamp, 2017; henceforth "GR17").

In total, we generated synthetic data of 20 experiments with 20 simulated participants each, and with 200 trials per participant. The decision threshold and non-decision time parameters were fixed across participants (i.e., $b = 1.5$; $t_0 = 0.2$). The remaining three parameters of the LBA model were sampled individually for each participant: i.) a value representing the average drift rate parameter ν was drawn from a normal distribution with mean $\mu_\nu = 0.6$ and standard deviation $\sigma_\nu = 0.02$, ii.) a value representing the standard deviation parameter s (for trial-by-trial variability of the drift rate) was drawn from a normal distribution with $\mu_s = 0.2$ and $\sigma_s = 0.02$, and iii.) a value representing the height of the start point distribution A was drawn from a normal distribution with $\mu_A = 1$ and $\sigma_A = 0.1$ (with the constraint that $A \leq b$). Trial-specific drift rates ν_t were drawn from a normal distribution with $\mu(\nu_t) = \nu$ and $\sigma(\nu_t) = s$. Trial-specific start points A_t were drawn from a uniform distribution with limits $[0, A]$. Note that these values refer to the (trial-specific) drift rates and start points for correct responses (as, for instance, in a two-alternative forced-choice perceptual or inferential choice task). Similarly, trial-specific drift rates and start points for incorrect responses were drawn from distributions with the same specifications, except that the mean of the drift rate was $1-\nu$. We chose the above-mentioned values so that accuracy rates and RTs for simulated participants were kept in realistic ranges (i.e., mean accuracy rates and RTs per simulated participant ranged from 53 to 89 percent and from 1.5 to 2.1 seconds, respectively).

Parameter estimation procedures were very similar to those of the HD-LBA model. An initial grid search with extensive search (i.e., 15 grid search values) for parameters ν and A

and limited search (i.e., 3 grid search values) for the remaining parameters was used to find good starting values for the constrained simplex minimization, followed by applying the four methods for capturing trial-by-trial variability. The STLBA method was applied according to the equations provided in van Maanen et al. (2011). The GR17 method required specifying a range of possible parameter values for ν and A as well as a step size with which this range of values was scanned (i.e., prior, likelihood and posterior values were calculated for each step within the range of possible values; for more details, see Gluth & Rieskamp, 2017). Possible values for both parameters were restricted to be within 3 standard deviations below and above the parameters' group means, with 1'000 steps per parameter. This implied 1 million calculations for both prior and likelihood distributions, because the joint distributions over both parameters needed to be specified (with these specifications, the GR17 method required about three times the computation time of the three other methods combined). To test the methods' abilities to recover trial-specific drift rates and start points, we restricted the analysis to correct trials only (as often done in perceptual or inferential choice tasks).

A5. Simulation of neural signals

To test whether LOTO can capture neural signals that systematically vary with variability in a cognitive model parameter, we simulated such signals in a very basic form. We assumed that the neural signal $F(\pi)$ in trial t (which could, for instance, represent the amplitude of the fMRI blood-oxygen-level-dependent signal in a circumscribed region in the brain or the average EEG spectral power in a specific frequency band during a circumscribed time period of the trial) that is systematically linked to variability in a parameter π is drawn from a normal distribution with mean π_t and standard deviation $\sigma_{F(\pi)}$ (see Palestro et al., 2018, for a similar assumption of normally distributed neural data in related work). The value for the standard deviation $\sigma_{F(\pi)}$ was specified as follows: We simulated $F(\pi)_t$ together with π_t and the observations (choices and RT) for the desired number of trials (160) participants (30), and

then conducted linear regressions with $F(\pi)_t$ as dependent and π_t and the observations as independent variables in each simulated participant (independent variables were standardized). Then, we estimated the effect size of the regression coefficient of π_t at the group level. The goal was to obtain an effect size of $d = 1$ by trying out different values for $\sigma_{F(\pi)}$ (the choice of the effect size is justified in the next paragraph). This iterative procedure was stopped as soon as a value of $\sigma_{F(\pi)}$ was found, for which the 95% confidence interval of the effect size overlapped with 1 when generating 1'000 simulations of 30 participants with 160 trials. For parameter κ of the HD-LBA model, we obtained $\sigma_{F(\kappa)} = 0.92$; for parameter b of the HD-LBA model, we obtained $\sigma_{F(b)} = 5$. Accordingly, this approach leads to a simulated neural signal for parameter κ that is much less variable than the signal for parameter b, because the trial-by-trial variability in the former parameter is much lower than in the latter (see above). This would be implausible from a neurophysiological point of view. However, this implausibility does not matter in the context of the current work, because we are only interested in the correlation between $F(\pi)_t$ and π_t (and the LOTO estimate associated with π_t), which is unaffected by a linear transformation of $F(\pi)_t$.

The rationale of this approach is that in an idealized world, in which we would know the true trial-specific values π_t, we could expect a very high effect size (such as $d = 1$) for detecting a relationship between π_t and $F(\pi)_t$. This effect size is slightly higher than, for instance, current estimates of realistic effect sizes for fMRI studies, which are based on large-scale empirical studies (Geuter et al., 2018; Poldrack et al., 2017). But again, the chosen effect size refers to a hypothetical world, in which the true trial-by-trial value of a cognitive function is known and then related to brain activity. For such a fortunate scenario, we can expect to obtain a slightly higher effect size than in reality.

We generated 1'000 datasets of 30 simulated participants with 160 trials for the HD-LBA model. For each of these datasets, we conducted random-effects linear regression

analyses, that is, we ran two linear regression analyses with $F(\kappa)_t$ / $F(b)_t$ as dependent variable and the LOTO estimate for π_t and b_t together with the observations (i.e., choices and RT) as independent variables in each simulated participant and then conducted two-sided one-sample *t*-tests on the regression coefficients for π_t and b_t at the group level. The proportion of significant effects for "correct" / "wrong" pairings of *F* and π determine the statistical power / false-positive rate of LOTO.

A6. General settings

All simulations and analyses were conducted in MATLAB. Function minimization via a (constrained) simplex algorithm was conducted using the function *fminsearchcon*. For determining convergence of parameter estimation and termination of search, a very strict tolerance criterion of 10^{-13} was chosen, because LOTO needs to work with very small differences in deviance when trying to adjust a parameter value from the "all-trial" estimate to the *n*-1 estimate, in particular when *n* is high. To illustrate the effect of the tolerance criterion in section VIII, the performance of LOTO for the HD model was compared between the strict tolerance criterion and the default criterion in MATLAB, which is 10^{-4}.

A7. Code availability

The computer codes associated with this article and the LOTO method are freely available on the Open Science Framework (https://osf.io/du85f/). Although running these codes requires access to the proprietary MATLAB software, we stress that LOTO is a technique rather than a software package and can therefore be realized in various programing languages (including open source software such as R or Python).

References of Appendix A

Brown, S.D., and Heathcote, A. (2008). The simplest complete model of choice response time: linear ballistic accumulation. Cognit. Psychol. *57*, 153–178.

Geuter, S., Qi, G., Welsh, R.C., Wager, T.D., and Lindquist, M.A. (2018). Effect size and power in fMRI group analysis. bioRxiv.

Gluth, S., and Rieskamp, J. (2017). Variability in behavior that cognitive models do not explain can be linked to neuroimaging data. J. Math. Psychol. *76*, 104–116.

Gluth, S., Hotaling, J.M., and Rieskamp, J. (2017). The attraction effect modulates reward prediction errors and intertemporal choices. J. Neurosci. *37*, 371–382.

Koffarnus, M.N., Deshpande, H.U., Lisinski, J.M., Eklund, A., Bickel, W.K., and LaConte, S.M. (2017). An adaptive, individualized fMRI delay discounting procedure to increase flexibility and optimize scanner time. NeuroImage *161*, 56–66.

van Maanen, L., Brown, S.D., Eichele, T., Wagenmakers, E.-J., Ho, T., Serences, J., and Forstmann, B.U. (2011). Neural correlates of trial-to-trial fluctuations in response caution. J. Neurosci. *31*, 17488–17495.

Palestro, J.J., Bahg, G., Sederberg, P.B., Lu, Z.-L., Steyvers, M., and Turner, B.M. (2018). A tutorial on joint models of neural and behavioral measures of cognition. J. Math. Psychol. *84*, 20–48.

Peters, J., and Büchel, C. (2009). Overlapping and distinct neural systems code for subjective value during intertemporal and risky decision making. J. Neurosci. *29*, 15727–15734.

Peters, J., Miedl, S.F., and Büchel, C. (2012). Formal comparison of dual-parameter temporal discounting models in controls and pathological gamblers. PloS One *7*, e47225.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.-B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. *18*, 115–126.

Appendix B: Derivatives of the hyperbolic discounting (HD) model

The utilities u_i and u_j of the immediate and delayed choice options as well as the corresponding choice probabilities p_i and p_j according to the HD model are provided in Equation 6 and 7, respectively.

The log-likelihood for a single choice of the immediate option is given by the natural logarithm of the choice probability p_i and thus:

$$LL_i = \ln(p_i) = \ln\left[\frac{1}{1+\exp(-\beta*[u_i-u_j])}\right] = \ln[1] - \ln\left[1+\exp\left(-\beta*\left[u_i-u_j\right]\right)\right]$$

$$= -\ln\left[1+\exp\left(-\beta*\left[u_i-u_j\right]\right)\right] \quad . \qquad \text{(B1)}$$

The derivative of LL_i with respect to κ (which is hidden in u_j in Equation B1) is therefore:

$$\frac{\partial LL_i}{\partial \kappa} = -\frac{1}{1+\exp(-\beta*[u_i-u_j])} * \frac{\partial[1+\exp(-\beta*[u_i-u_j])]}{\partial \kappa} = -\frac{\exp(-\beta*[u_i-u_j])}{1+\exp(-\beta*[u_i-u_j])} * \frac{\partial(-\beta*[u_i-u_j])}{\partial \kappa} =$$

$$-\left[1-\frac{1}{1+\exp(-\beta*[u_i-u_j])}\right] * \frac{\partial\left(-\beta*\left[x_i-\frac{x_j}{1+\kappa*d_j}\right]\right)}{\partial \kappa} = -(1-p_i) * \frac{\partial\left(\beta*x_j*[1+\kappa*d_j]^{-1}-\beta*x_i\right)}{\partial \kappa} =$$

$$-(1-p_i) * -\left(\beta*x_j*\left[1+\kappa*d_j\right]^{-2}\right) = \frac{\beta*x_j*d_j}{[1+\kappa*d_j]^2} * (1-p_i) \ . \qquad \text{(B2)}$$

Similarly, the log-likelihood for a single choice of the delayed option is:

$$LL_j = \ln(p_j) = \ln\left[\frac{1}{1+\exp(-\beta*[u_j-u_i])}\right] = \ln[1] - \ln\left[1+\exp\left(-\beta*\left[u_j-u_i\right]\right)\right]$$

$$= -\ln\left[1+\exp\left(-\beta*\left[u_j-u_i\right]\right)\right] \quad . \qquad \text{(B3)}$$

And the derivative of LL_j with respect to κ is therefore:

22 $$\frac{\partial LL_j}{\partial \kappa} = -\frac{1}{1+\exp(-\beta*[u_j-u_i])} * \frac{\partial[1+\exp(-\beta*[u_j-u_i])]}{\partial \kappa} = -\frac{\exp(-\beta*[u_j-u_i])}{1+\exp(-\beta*[u_j-u_i])} * \frac{\partial(-\beta*[u_j-u_i])}{\partial \kappa} =$$

23 $$-\left[1-\frac{1}{1+\exp(-\beta*[u_j-u_i])}\right] * \frac{\partial\left(-\beta*\left[\frac{x_j}{1+\kappa*d_j}-x_i\right]\right)}{\partial \kappa} = -\left(1-p_j\right) * \frac{\partial\left(\beta*x_i-\beta*x_j*\left[1+\kappa*d_j\right]^{-1}\right)}{\partial \kappa} =$$

24 $$-p_i * \left(\beta * x_j * \left[1+\kappa * d_j\right]^{-2}\right) = -\frac{\beta*x_j*d_j}{[1+\kappa*d_j]^2} * p_i \ . \qquad \text{(B4)}$$

Appendix C: Example derivatives of LOTO for the hyperbolic discounting (HD) model

Here, we show that LOTO will produce estimates of a parameter that are related to both the choices and task features and that are not extreme, as soon as each option has been chosen at least 2 times (in the case of a two-alternative choice task). We will use the examples of the HD model mentioned in the main text. Assume that there are two trials with identical features (i.e., identical amount x_i of the immediate option, identical amount x_j of the delayed option, identical delay d_j of the delayed option) and that the immediate option is chosen in one trial but the delayed option is chosen in the other trial (we assume identical features only to simplify the derivatives; as stated in the main text, it is important to vary the features across trials). In this case, the derivative of the log-likelihood w.r.t. the discount factor κ is given by Equation 8c. Setting this equation to 0 yields the maximum likelihood estimate (MLE) of the parameter:

$$0 = \frac{\partial LL_{\{i,j\}}}{\partial \kappa} = \frac{\partial LL_i}{\partial \kappa} + \frac{\partial LL_j}{\partial \kappa} = \frac{\beta * x_j * d_j}{[1+\kappa * d_j]^2} * (1 - p_i) - \frac{\beta * x_j * d_j}{[1+\kappa * d_j]^2} * p_i \ . \qquad \text{(C1)}$$

We can easily generalize this to the case of n trials in which the immediate option was chosen n_i times and the delayed option was chosen n_j times (i.e., $n_i + n_j = n$):

$$0 = n_i * \frac{\partial LL_i}{\partial \kappa} + n_j * \frac{\partial LL_j}{\partial \kappa} = n_i * \frac{\beta * x_j * d_j}{[1+\kappa * d_j]^2} * (1 - p_i) - n_j * \frac{\beta * x_j * d_j}{[1+\kappa * d_j]^2} * p_i \ . \qquad \text{(C2a)}$$

This equation has to be solved for κ which requires some steps. By putting the second part of the right-hand side to the left hand side, we get:

$$n_i * \frac{\beta * x_j * d_j}{[1 + \kappa * d_j]^2} * (1 - p_i) = n_j * \frac{\beta * x_j * d_j}{[1 + \kappa * d_j]^2} * p_i$$

This simplifies to:

$$n_i * (1 - p_i) = n_j * p_i$$

We re-arrange this to:

$$\frac{n_j}{n_i} = \frac{(1 - p_i)}{p_i} = \frac{1}{p_i} - 1$$

Note that p_i is specified in Equation 7. Therefore:

$$\frac{n_j}{n_i} = \frac{1}{\frac{1}{1 + \exp(-\beta * [u_i - u_j])}} - 1 = 1 + \exp(-\beta * [u_i - u_j]) - 1 = \exp(-\beta * [u_i - u_j])$$

Taking the logarithm on both sides yields:

$$\ln\left(\frac{n_j}{n_i}\right) = -\beta * (u_i - u_j) = \beta * u_j - \beta * u_i$$

We re-arrange this to:

$$\ln\left(\frac{n_j}{n_i}\right) + \beta * u_i = \beta * u_j$$

Note that u_i and u_j are given by Equation 6. Therefore:

$$\ln\left(\frac{n_j}{n_i}\right) + \beta * x_i = \beta * \frac{x_j}{1 + \kappa * d_j}$$

Finally, we re-arrange this to isolate κ:

$$(1 + \kappa * d_j) * \left(\ln\left[\frac{n_j}{n_i}\right] + \beta * x_i\right) = \beta * x_j$$

$$1 + \kappa * d_j = \frac{\beta * x_j}{\left(\ln\left[\frac{n_j}{n_i}\right] + \beta * x_i\right)}$$

$$\kappa * d_j = \frac{\beta * x_j}{\left(\beta * x_i + \ln\left[\frac{n_j}{n_i}\right]\right)} - 1$$

$$\kappa = \frac{1}{d_j} * \left[\frac{\beta * x_j}{(\beta * x_i + \ln[n_j] - \ln[n_i])} - 1\right]. \qquad \text{(C2b)}$$

For instance, in the case of $n_i = n_j$, this equation simplifies to:

$$\kappa = \frac{x_j - x_i}{d_j * x_i},$$

which yields a κ of 0.075 and 0.182 for the two example trials mentioned in the main text (compare with the right panels of Figure 2).

Importantly, Equation C2b can be expected to yield reasonable results as long as n_i and n_j are ≥ 1. If either n_i or n_j is 0, then the first term in the parentheses on the right-hand side of Equation C2b will be 0, because we have $\ln[0] = \pm\infty$ in the denominator, and the estimation of κ will simply be $-1/d_j$, which is negative and thus lies outside of the range of values allowed for κ. When applying LOTO, we therefore require n_i and n_j to be ≥ 2, because the parameter is estimated once for n and once for n-1 trials, and in the n-1 case LOTO will yield an unreasonable estimate in one trial (i.e., if $n_i = 1$, then the error will be in that one trial in which the immediate option was chosen; if $n_j = 1$, then the error will be in that one trial in which the delayed option was chosen).

Appendix D: Detailed specifications of the example application of LOTO in Section XI

D1. Essential study information

A total of 84 participants took part in the study (Gluth et al., 2015) which consisted of three groups (one group of participants that underwent fMRI and two groups that conducted behavioral tasks only). The final sample of the fMRI group included 30 participants, and the analyses in the current article are restricted to this sample. All participants gave written informed consent and the study was approved by the Ärztekammer Hamburg, a local ethics committee in Hamburg (Germany) where the study was conducted.

Participants were asked to not to eat for 3 hours prior to the experiment. They were familiarized with a set of 48 common food snacks from German supermarkets. To infer the subjective values of food snacks, participants conducted a Becker-DeGroot-Marschak (BDM) auction (Becker et al., 1964), in which they stated their willingness to pay money for each snack. To incentivize accurate bids, one randomly selected bid trial was realized at the end of the experiment. Inside the MR scanner, participants conducted the remember-and-decide task which consisted of multiple runs of encoding, distraction, decision-making and recall phases. During encoding, participants learned associations between six different food snacks and six different screen positions. To manipulate memory performance, half of the snacks were encoded once, and the other half twice. During distraction, participants worked on a 2-back working memory task for 30 seconds. Afterwards, participants made nine decisions between pairs of snacks. Snacks were not shown directly, but the respective screen positions were highlighted and participants had to use their memory to make accurate decisions. To manipulate deliberation time, participants were given either a one- or a five-second time window before they had to make a choice. Finally, the ability to remember snacks correctly was assessed via cued recall (during this phase, MR scanning was paused to enable communication between participant and experimenter). Participants were incentivized to

make accurate choices by realizing one choice from a randomly selected decision-making trial. Further information can be found in Gluth et al. (2015).

D2. Cognitive model

To predict preferential choices and to inform the LOTO-based fMRI analysis, we used a slightly modified version of the cognitive model proposed in the original study (each modification is explained below). This model assumes that decisions are based on a noisy evidence accumulation (diffusion) process which is informed by the subjective values of the available choice options. The model also accounts for the possibility that options are not recalled and for participants' tendency to prefer remembered over forgotten options (the *memory bias*; see also Gluth et al., 2015; Mechera-Ostrovsky and Gluth, 2018). Because participants were forced to wait for a specified amount of time (i.e., either one or five seconds) and then had to indicate their decision promptly, the model does not assume a decision threshold but an unbounded diffusion process that is terminated at the specified amount of time. The probability to choose an option is thus given by the probability that the diffusion process ends with a positive or negative value. More specifically, the probability of choosing the left option l over the right option r is:

$$p(l|\{l,r\}) = \Phi\left(\frac{\Delta_V * \sqrt{t}}{\sigma}\right) \quad , \tag{D1}$$

where $\Phi(x)$ refers to the value of the cumulative distribution function of the standard normal distribution at x, t refers to the deliberation time (i.e., $t = 1$ or $t = 5$), σ accounts for the amount of noise in the diffusion process ($\sigma \geq 0$), and Δ_V refers to the value difference between l and r, which is given by:

$$\Delta_V = M_l * (V_l - g) - M_r * (V_r - g) \quad , \tag{D2}$$

where M_l is the probability to remember option l, V_r is the standardized subjective value of option r, and g represents the memory bias. The memory probability M equals either α or β,

depending on whether an item has been shown once or twice during encoding, and α and β are two free parameters ($0 \leq \alpha, \beta \leq 1$). This definition differs from the model in Gluth et al. (2015) in which β represented the improvement of memory performance for items encoded twice compared to items encoded once. We changed this definition because here we are not interested in testing whether there is a significant improvement of memory but instead in the influence of memory regardless of whether it was improved. The second adaption of the model is that the memory bias *g* is not a free parameter. Instead, we took each participant's intercept coefficient from the logistic regression that tested for the presence of a memory bias (see Equation 1 in Gluth et al., 2015). This adaption was made because contrary to the original study, we did not estimate parameters with hierarchical Bayesian modeling here, and treating the memory bias as a free parameter led to severe misspecifications of the other free parameters in some participants when using maximum likelihood estimation.

The three free parameters of the model (i.e., σ, α, β) were estimated via maximum likelihood estimation using the minimization techniques described in Appendix A. A grid , but the minimization was repeated ten times with randomly generated starting values. After estimating the full model, LOTO was applied to extract trial-by-trial (or more exactly: item-by-item) variability in memory probability *M* by fixing σ to the full-model estimate and re-estimating either α or β with each item taken out once (e.g., for an item *i* that was encoded twice, α was also fixed to the full-model estimate and β was re-estimated by taking out those three decision-making trials that included *i* as a choice option). The difference between the *n* item and the *n*-1 item parameter estimates were then computed and used for the fMRI analysis. To test for the presence of systematic variability in memory probability (see Section VI), we generated 1'000 datasets of each participant using the full-model parameter estimates and assuming no variability in α or β. LOTO was then applied to these datasets and the

improvement in model fit (relative to the all-trial model fit) was compared to that in the real data.

D3. Analysis of fMRI data

Because the fMRI analysis closely followed the procedures in Gluth et al. (2015), we provide only a brief overview here. Details on fMRI data acquisition, pre-processing and statistical analyses can be found in the original publication. Statistical analysis was based on the General Linear Model (GLM) approach as implemented in SPM12. As in the original study, twelve onset vectors were defined including two vectors for the encoding phase, one for items shown once and one for items shown twice. Both of these two onset vectors were accompanied by three parametric modulators (PM) that modeled the predicted change of the hemodynamic response as a function of the following variables of interest: i.) the subjective value of the encoded snack; ii.) whether the snack was correctly remembered or not in the subsequent cued recall phase; iii.) the LOTO-based estimate of item-by-item variability in memory probability (see above). Importantly, the LOTO-based PM was entered into the GLM last, so that correlated fMRI signals could be uniquely attributed to this PM. Thus, we conducted a highly rigorous test of whether LOTO can be used to link trial-by-trial variability of a cognitive model to brain activity. For each participant, one contrast image was generated that combined the effect of the LOTO-based PM for both onset vectors of the encoding phase.

At the group level, individual contrast images were subjected to a one-sample *t*-test. For illustration purposes, we performed two different statistical analyses. In the first analysis, we tested for significant voxels within the bilateral hippocampus, using a family-wise error correction for small volume at $p < .05$. The small volume was defined by a bilateral anatomical mask of the hippocampus taken from the AAL brain atlas (Tzourio-Mazoyer et al., 2002). The results were displayed on the mean anatomical (T1) image of all participants using an uncorrected threshold of $p < .001$ with at least 10 contiguous voxels. The second analysis

was a region-of-interest (ROI) analysis for which we averaged the fMRI signal for left and right hippocampus separately (using the same anatomical mask) and tested whether the average signals in the ROI was significantly positive (at $p < .05$, two-sided).

References of Appendix D

Becker, G.M., DeGroot, M.H., and Marschak, J. (1964). Measuring utility by a single-response sequential method. Behav. Sci. *9*, 226–232.

Gluth, S., Sommer, T., Rieskamp, J., and Büchel, C. (2015). Effective connectivity between hippocampus and ventromedial prefrontal cortex controls preferential choices from memory. Neuron *86*, 1078–1090.

Mechera-Ostrovsky, T., and Gluth, S. (2018). Memory beliefs drive the memory bias on value-based decisions. Sci. Rep. *8*.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., and Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage *15*, 273–289.