Behav Res (2018) 50:989–1010
DOI 10.3758/s13428-017-0921-7

Improved information pooling for hierarchical cognitive models through multiple and covaried regression

R. Anders[1] · **Z. Oravecz**[2] · **F.-X. Alario**[1]

Published online: 11 July 2017
© Psychonomic Society, Inc. 2017

Abstract Cognitive process models are fit to observed data to infer how experimental manipulations modify the assumed underlying cognitive process. They are alternatives to descriptive models, which only capture differences on the observed data level, and do not make assumptions about the underlying cognitive process. Process models may require more observations than descriptive models however, and as a consequence, usually fewer conditions can be simultaneously modeled with them. Unfortunately, it is known that the predictive validity of a model may be compromised when fewer experimental conditions are jointly accounted for (e.g., overestimation of predictor effects, or their incorrect assignment). We develop a hierarchical and covaried multiple regression approach to address this problem. Specifically, we show how to map the recurrences of all conditions, participants, items, and/or traits across experimental design cells to the process model parameters. This systematic pooling of information can facilitate parameter estimation. The proposed approach is particularly relevant for multi-factor experimental designs, and for mixture models that parameterize per cell to assess predictor effects. This hierarchical framework provides the capacity to model more conditions jointly to improve parameter recovery at low observation numbers (e.g., using only 1/6 of trials, recovering as well as standard hierarchical Bayesian methods), and to directly model predictor and covariate effects on the process parameters, without the need for post hoc analyses (e.g., ANOVA). An example application to real data is also provided.

Keywords Hierarchical cognitive modeling · Multiple regression · Psychometrics · Data analysis · Parameter recovery · Bayesian inference

Introduction

Descriptive models for data analysis (e.g., stochastic distributions, regression, principal components analysis) are efficient for measurement purposes. They have relatively few parameters, and can model predictor effects using low numbers of observations (Baayen, Davidson, & Bates, 2008; Lazarsfeld, 1959; Cohen, 1968; Howell, 2012; Jolliffe, 2002; Wilcox, 2012). For example, traditional regression models account for predictor effects with single parameters (e.g., the β coefficients) that are based on the information pooled across an experiment's data cells. This approach is efficient for investigating many predictors in a joint context (simultaneously). Jointly modeling the effects of predictors is preferred because it usually improves the predictive validity of a model, as compared to approaches with independent effects (see Baayen, 2004; Baayen et al., 2008; Barr, Levy, Scheepers, & Tily, 2013). Through modeling several predictors simultaneously, descriptive models can be used to determine *which* predictors are informative. Predictors can be considered informative to the extent that they account for variance in the response data (e.g., smaller/larger magnitudes of choice proportions, accuracy rates, response times).

✉ R. Anders
royce.anders@univ-amu.fr

Z. Oravecz
zita@psu.edu

F.-X. Alario
francois-xavier.alario@univ-amu.fr

[1] Aix Marseille Univ, CNRS, LPC, Marseille, France

[2] Pennsylvania State University, State College, PA, USA

However, these modeling approaches also have their limitations. Particularly, they serve as descriptive analytical tools rather than as explanatory process models. That is, they do not provide any model of *how* the predictors may affect the underlying cognitive process(es) involved in the generation of the observed behaviors (Busemeyer & Diederich, 2010).

In contrast, *process models* (e.g., Anderson, 1996; Busemeyer & Townsend, 1993; Pike, 1973; Van der Linden & Hambleton, 1997) aim to model cognitive dynamics by focusing on one or more cognitive mechanisms that can account for the observed performance differences in an experiment. Particularly, data-driven process models have parameters that map onto cognitive processes. These parameters can be estimated from the observed data and derived for each experimental condition of interest. This allows us to talk about predictors in terms of *how* they affect an underlying cognitive process. Fitting complex process models requires more data than fitting their descriptive counterparts however. Moreover, these models tend to pool information less efficiently, meaning that they are usually fit by experimental cell (e.g., several parameters are added for each cell). As a consequence, such cognitive models can include fewer predictors jointly, which can reduce the validity of their predictions. As noted previously, separate analyses of predictors (e.g., conditions, participants, or items), can cause misattribution errors such as overestimation of effects, type I errors, etc. For example, when certain predictors that significantly account for performance differences are not simultaneously modeled, a model may mistakenly attribute these performance differences to other predictors (Baayen, 2004; Baayen et al., 2008; Barr et al., 2013).

We offer a solution to this problem based on maximizing the information pooled to the process model. With this framework, parameter recovery is improved for the process model at lower numbers of observations; consequently, more predictors can be simultaneously modeled. This is achieved through a hierarchical and covaried multiple regression approach in which information is mapped across experimental cells from the recurrences of all conditions, participants, and items (i.e., cases of repeated measures) to all of the process model's cognitive parameters. In addition, the covariation between the cognitive parameters themselves is modeled. The framework is particularly relevant for multi-factor experiments and for mixture models that parameterize per cell to model predictor effects. It can be an effective approach for advancing empirical modeling methods used for measurement and cognitive-behavioral inferences, known in some domains as *cognitive psychometric* models (see Batchelder, 1998; Batchelder & Riefer, 1999; Riefer, Knapp, Batchelder, Bamber, & Manifold, 2002).

The proposed approach builds upon previous developments in a growing movement known as *hierarchical* cognitive modeling, which has been shown to improve the analytical potential of such process models. Hierarchical cognitive modeling (Lee, 2011; Kruschke, 2011; Rouder, Morey, & Pratte, 2013; Scheibehenne & Pachur, 2015) typically consists of embedding (or nesting) a statistical model at a layer above the process or cognitive model. Major hierarchical approaches have included the following implementations: stochastic population distributions over observed participants/items to constrain estimation error (Rouder & Lu, 2005; Rouder, Lu, Speckman, Sun, & Jiang, 2005; Rouder et al., 2007; a regression model on a parameter to analyze cofactors or trial-by-trial covariates of its value (Cavanagh et al., 2011; Frank et al., 2015; Oravecz, Anders, & Batchelder, 2015; Vandekerckhove et al., 2011); and *latent* predictor modeling as a method for clustering (e.g., of participants/latent signals Anders & Batchelder, 2012, or relevant cognitive abilities across tasks Vandekerckhove, 2014).

The current framework builds most closely upon the previously cited works that emphasize hierarchy through embedded regression models. Specifically, these works demonstrate the advantages of nesting a simple regression on one or two cognitive parameters in order to jointly model a predictor or a trial-by-trial neural activity that covaries with that parameter. The present paper elaborates this framework to a full-fledged regression structure which maps the entire experimental design (conditions and each of their levels, participants, and items). To our knowledge, it is the first work to demonstrate that such a framework can markedly improve parameter recovery at low observation numbers, and hence permit a cognitive model to simultaneously fit more experimental conditions than usual. Hence, the present work focuses on optimal hierarchical methods for experimental designs, and it is complementary research to the previous works. That is, economizing observation numbers for modeling the core experimental design may help to model additional cofactors or trial-by-trial covariates jointly.

The proposed framework involves a hierarchical modeling comprised of (1) a process model, (2) a multiple and covaried regression structure, and (3) by-group (e.g., by predictor, participant, and item intercepts) population distributions. In the following sections, we will demonstrate how this three-tiered approach can markedly improve the information pooled to a process model. Due to the more complex expression of likelihood however, a more advanced estimation approach is typically needed to implement this modeling framework. We therefore utilize the Bayesian estimation approach (Gelman, Carlin, Stern, & Rubin, 2004). A number of advantages have been identified with the Bayesian approach (Lee, 2011; Kruschke, 2011; Rouder et al., 2013; Scheibehenne & Pachur, 2015), including the simultaneous, rather than sequential estimation of model parameters, and the ability to constrain error in estimation, which can improve parameter recovery performance or a model's capacity to make predictions from data.

The paper is organized as follows. The first section, "Process parameters as a function of a hierarchical multiple regression structure" develops the framework. Next, "Data-driven process models for performance data: Sequential sampling" introduces a popular genre of cognitive models that is frequently used for performance data (e.g., response times and accuracy), known as sequential sampling (Busemeyer & Townsend, 1992; Townsend & Ashby, 1983). Consequently, we will demonstrate the framework using a standard sequential sampling model, which can be generalized to a number of experiments involving response time analyses. Then using hierarchical Bayesian methods, we develop an estimation approach in "Bayesian estimation". We will also remark as to how the framework can be easily adapted to a variety of other process models. Next, "Fitting approach" discusses important fitting topics for the proposed approach, and "Bayesian sampler settings" includes the recommended specifications. "Application to simulated data" demonstrates the aforementioned advantages of the approach through several large simulation studies. "Application to experimental data" provides an empirical application, and finally "Discussion" includes the general discussion.

Process parameters as a function of a hierarchical multiple regression structure

Consider an experiment that is designed with F factors, each having L_f levels (for example a 2×2 design or a $3 \times 3 \times 2$ design), which will be tested with multiple participants, P, and/or items, I, of interest. This kind of experimental design gives rise to a number of unique experimental design cells, C, each having a unique combination of factor levels per participant and/or item. The experimenter(s) will collect N observations in each cell, as y_{j_c}, in which observation $j \in 1 \ldots, N$ and design cell $c \in 1, \ldots, C$. This set of N observations along C experimental cells is defined as the response data.

To learn more about the underlying cognitive process(es), and how the predictors may affect them, a researcher selects a data-driven process model that possesses an expression of likelihood, or a probability density function $f(\cdot)$, that can be used to fit the data (y_{j_c}) as a function of a set of cognitive parameters, Ω. A maximally data-driven model will estimate these parameters per every design cell c, as Ω_c, and is known as a finite mixture-model implementation (Everitt, 1981).[1] The model parameters for a given cell are typically used to model the distribution of observed data in that cell (e.g., central trends and variance across trials). Then a general stochastic expression of the data *per design cell*, y_{j_c}, as related to the mixture application of the cognitive process model, can be expressed as

$$y_{j_c} \sim f(\Omega_c = \{\omega_{1c}, \omega_{2c}, \ldots, \omega_{Kc}\}), \tag{1}$$

where Ω_c contains the K cognitive parameters ω_k that model design cell c. From this modeling, predictors (or how experimental conditions, participants, and items modify the cognitive parameters) can then be retrieved by a posterior analysis of the parameters across cells, for example by an analysis of variance (ANOVA, Iversen, & Norpoth, 1987; Cohen & Cohen, 1988).

However, in this by-cell modeling approach, extra parameters are specified per cell, and there is much information shared between cells that is lost. As a result, more data are required per cell, and less predictors can be modeled. Hierarchical modeling has made developments on this issue, particularly by addressing the recurrence of the same participants or items in different cells, which are cases of repeated measures. A current standard in the field is to nest a population distribution at a layer above the participants or items in the process model. This approach can utilize the between-subject variance to improve the within-subject estimates, and according to recurring subjects across conditions, can improve the within-subject parameter estimates. Furthermore, one can also use the subjects' group parameters to make generalizations about the population itself (Rouder & Lu, 2005; Rouder, Lu, Speckman, et al., 2005; Rouder, Lu, Sun, et al., 2007). As opposed to a simple distribution nesting, this approach can also be implemented through nesting a regression structure at a layer above the model (e.g., population intercept and error term, see Vandekerckhove et al., 2011). Nesting a regression structure also has the advantage to allow for jointly modeling a covariate or between-trial effects (Cavanagh et al., 2011; Frank et al., 2015). However for a given process model, it is good practice to first verify through simulation analyses that there are enough observations to appropriately fit the data at such granularity.

While the benefits from sharing information across cells of repeated participants and/or items has been largely recognized, it has not yet been quantified to what extent one benefits from sharing information across cells of repeated condition levels (in tandem with the participants and items). Furthermore, it has not been studied how this information can be effectively mapped to *all* of the process model parameters. In this work, we pursue such a study and develop an approach to implement the methodology. This is accomplished through a multiple and covaried regression structure that is embedded hierarchically, which informs all of the cognitive parameters and also models their correlation. An illustration of how such an approach can improve the information pooled into the process model is provided in the

[1] A less data-driven model may economize observation numbers by estimating select parameters from data *pooled* across cells (fewer parameters estimated), rather than having the opportunity to observe if the data instead suggest the parameters vary across individual cells.

following paragraph. Then through several implementations and large recovery analyses, the Section "Application to simulated data" demonstrates the advantages of the approach.

In experimental designs, frequently many of the same condition levels, L_f, along a number of factors, F, participants, P, and items, I, are found in other cells. Information about their effects across unique cells (e.g., in the context of other predictor/participant/item combinations) can be pooled by a hierarchical multiple regression structure. Note that regression maps information through indicator values, x, that pool information from all recurrences of a condition, regardless of its cell membership. Consider a cognitive model in which there are $K = 3$ process model parameters $(\omega_1, \omega_2, \omega_3)$, for which at least $N = 60$ observations in a cell c are needed to appropriately fit the parameters. Secondly, suppose that the experimental data is derived from a 2×2 design, which provides for $L = 2$ levels, for each of $F = 2$ factors, and involves $P = 10$ participants and $I = 6$ items. Let lowercase script be the index for each condition (factor, participant, item), then the parameters in Ω_c are quadruply indexed as $\{\omega_{1pil_{f_1}l_{f_2}}, \omega_{2pil_{f_1}l_{f_2}}, \omega_{3pil_{f_1}l_{f_2}}\}$, which expands the total number of experimental cells to 240 (e.g., $L_{f_1} \times L_{f_2} \times P \times I = 2 \times 2 \times 10 \times 6 = 240$). Hence, each participant would need to complete $24 \times 60 = 1440$ trials to reliably estimate these predictors in a joint context. While these numbers are realistic for this simple experimental design, they are still largely inconvenient to obtain. However, as we will later show in Table 1, by implementing the proposed hierarchical multiple regression approach, one could have performed the same process modeling with only 1/6 of the observations (240 per subject instead of 1440). Therefore, such an approach allows for smaller experiments to be modeled, more conditions to be jointly modeled (e.g., covariates, between-trial effects, or experiments with additional predictors), and improved parameter recovery at low observation numbers. These developments can benefit a researcher's capacity to make inferences from data with cognitive process models.

The hierarchical multiple regression approach is formally specified as follows: suppose each cognitive parameter in the set $\Omega_c = \{\omega_{1c}, \omega_{2c}, \ldots, \omega_{Kc}\}$ contains a full regression structure of coefficients, b_w, with $w \in 1, \ldots, W$ for the levels among F specified predictors, as well as participant b_p and item b_i intercepts. Then the model is specified by:

$$\begin{aligned}
\omega_{1c} &= b_{1_{\omega_1}} x_{c_1} + \ldots + b_{W_{\omega_1}} x_{cW} + b_{p_{\omega_1}} + b_{i_{\omega_1}} + \varepsilon_1 \\
\omega_{2c} &= b_{1_{\omega_2}} x_{c_1} + \ldots + b_{W_{\omega_2}} x_{cW} + b_{p_{\omega_2}} + b_{i_{\omega_2}} + \varepsilon_2 \\
&\vdots \qquad\qquad \vdots \qquad\qquad \vdots \\
\omega_{Kc} &= b_{1_{\omega_K}} x_{c_1} + \ldots + b_{W_{\omega_K}} x_{cW} + b_{p_{\omega_K}} + b_{i_{\omega_K}} + \varepsilon_K \\
\varepsilon_{1:K} &\sim \text{Multivariate Normal}(0, \mathbf{\Sigma}_{K \times K}) \,,
\end{aligned} \tag{2}$$

and the error ε_k of parameter ω_{kc} from the regression, as influenced by deviation from the model and the covariance of the other ω_{kc} parameters, is modeled by the multivariate normal with mean 0 and the $K \times K$ covariance matrix $\mathbf{\Sigma}$. The x_c are the indicators that link the W regressed conditions (and covariate effects if desired) to the corresponding experimental design cell c, which also has intercepts, I, according to participant p and item i.

Then the notation in Eq. 2 can be more simply compacted into a single multiplicative term as in Eq. 3, by introducing additional x_c terms, which by values of 0 or 1, appropriately index these intercepts. Consequently, the notation for each parameter is simplified as a single vector of weights (which includes intercepts), β_{ω_k}, and a vector of indicator values for the cell X_c, resulting in a generalized case notation where

$$\begin{aligned}
\omega_{1c} &= \beta_{\omega_1} X_c + \varepsilon_1 \\
\omega_{2c} &= \beta_{\omega_2} X_c + \varepsilon_2 \\
\vdots \quad &\quad \vdots \qquad \vdots \\
\omega_{Kc} &= \beta_{\omega_K} X_c + \varepsilon_K \\
\varepsilon_{1:K} &\sim \text{Multivariate Normal}(0, \mathbf{\Sigma}_{K \times K}) \,.
\end{aligned} \tag{3}$$

Furthermore, letting $k \in 1, \ldots, K$ be the index of the appropriate parameter in Ω_c, the notation is further summarized as

$$\begin{aligned}
\omega_{kc} &= \beta_{\omega_k} X_c + \varepsilon_k \\
\varepsilon_{1:K} &\sim \text{Multivariate Normal}(0, \mathbf{\Sigma}_{K \times K}) \,.
\end{aligned} \tag{4}$$

These formulas summarize the nested multiple regression approach. In the following section, we acquaint the reader with a popular class of cognitive process models used for handling performance differences, known as sequential sampling models. We will utilize a canonical example of this class to demonstrate an implementation of the approach.

Data-driven process models for performance data: Sequential sampling

Among cognitive process models for handling performance data (such as reaction times, responses), sequential sampling models are currently very popular in several domains (Busemeyer & Townsend, 1992; Townsend & Ashby, 1983). Sequential sampling can be conceived of as a time-based extension of the predominant framework for modeling response data, known as signal detection theory (SDT, Green & Swets 1966; Pike, 1973). Sequential sampling posits that performance differences, in the context of time, may be modeled by a noisy accumulation of information toward a threshold, whose crossing triggers the response. Furthermore, these models involve a parameter that distinguishes the time elapsed in this decision process from time elapsed in external processes, such as during the motor movement that ensues after the threshold is triggered. This framework has been effective in accounting for performance differences through such a mechanism, and these models can closely fit response time (RT) distributions. The approach has experienced continued support since its beginnings in the 1960s (Stone, 1960; Laming, 1968; Gerstein & Mandelbrot, 1964; Ratcliff, 1978) in both theoretical (e.g., simulation exploration) and real data applications of experimental psychology (Ratcliff, Van Zandt, & McKoon, 1999; Ratcliff, Gomez, & McKoon, 2004; Ratcliff, Thompson, & McKoon, 2015; Ratcliff & McKoon, 2008; Anders, Riès, van Maanen, Alario, 2015) and neuroscience (Dehaene, 2008; Kelly & O'Connell, 2013; O'Connell, Dockree, & Kelly, 2012).

Next, we will specify the key model parameters involved. The three principal components that describe a sequential sampling process are at minimum: an accumulation rate γ of quantity X (the behavioral activation level that accumulates), an absorbing threshold of value α, and an external time θ. Using these three parameters (γ, α, θ), a standard sequential sampling process is illustrated in the left plot of Fig. 1. This process models a single trial. The fluctuating black line is a representation of the activity (X) for the modeled behavior, and this activity accumulates positively over time (with noise). Specifically, the noisy accumulation of X occurs by the model at every time step, $t = 1$ milliseconds (ms), by sequential independent samples from a Gaussian distribution with mean γ and standard deviation 1 (hence the term, sequential sampling model). Note that in this simulation, X begins at a neutral value of 0, and increases (with noise) over time with an average rate of 0.08 units/ms (γ), until it hits the necessary threshold value at 40 units (α). Upon reaching the threshold, the response is initiated. Parameter θ includes motor time for response execution (here abbreviated as TEA, Time External to the Accumulation process), and may also include time for low-level perceptual processing or encoding.

In the right plot of Fig. 1, many trials (e.g., a subject within an experimental design cell) are modeled with the same three parameters that simulated the single trial in the left plot. Note that these finishing times, from when the evidence accumulates to the necessary threshold, plus the TEA (θ),[2] model the RTs. These model-predicted RTs form a positive, right-skewed distribution.

In this canonical sequential sampling model, the resultant RT distribution is directly tractable by the probability density function (pdf) of the shifted Wald (SW) distribution, also known as the three-parameter inverse Gaussian distribution. The three sequential sampling parameters $\{\gamma, \alpha, \theta\}$ respectively quantify RT distribution *tail thickness*, *variance around the mode*, and *location* (onset). Luce (1986) discusses the importance of these RT distribution aspects for psychometric studies. Furthermore, this particular model likelihood has a closed-form solution in Eq. 5. We will henceforth refer to this SW model as a canonical sequential sampling model (SSM), and it will be used to test a baseline implementation of the proposed framework.[3]

Adapting the multiple regression approach to a cognitive model

The generalized formula for the hierarchical multiple regression approach, provided in "Process parameters as a function of a hierarchical multiple regression structure", are easily adapted to various data-driven process models. This is mainly achieved by specifying the likelihood $f(\cdot)$ in Eq. 1, and the number of parameters K in Eq. 4, to align with the proposed process model for the data. One may also consider simpler models than sequential sampling, such as signal detection models, binomial rate models, and item response theory models (see Lee & Wagenmakers, 2014, for such models, and other potentials). These kinds of models are relevant to our approach when a researcher seeks to account for response differences along several experimental conditions, participants, and items. However, complex models that are generally used for other purposes than measurement, such as neural networks, are likely too complex to adapt to the hierarchical multiple regression approach.

When implementing the proposed approach for the canonical SSM (used here as our running example), the multiple-covaried model consists of $K = 3$ parameters that are estimated per cell: γ, α, and θ. We apply this model to analyze performance data consisting of response times from correct responses. The likelihood function for the RT data, $f(\cdot)$, is simply the shifted Wald probability density function. Formally stated, the RT data likelihood function $f(\cdot)$ in Eq. 1 is the SW pdf, in which the y_{j_c} are the RTs, as RT_{j_c}, and $\Omega_c = \{\gamma_c, \alpha_c, \theta_c\}$, as

$$f(\mathrm{RT}_{j_c} \mid \Omega_c = \{\gamma_c, \alpha_c, \theta_c\}) = \frac{\alpha_c}{\sqrt{2\pi(\mathrm{RT}_{j_c} - \theta_c)^3}} \cdot \exp\left\{-\frac{[\alpha_c - \gamma_c(\mathrm{RT}_{j_c} - \theta_c)]^2}{2(\mathrm{RT}_{j_c} - \theta_c)}\right\}, \quad (5)$$

[2] For illustrative simplicity, here θ (TEA) is placed before the evidence accumulation begins (at $\theta = 200$ ms). However, whether θ is placed before, after, or split around the actual accumulation process (e.g., accounting for both concept/visual recognition and response execution time), all of these options are quantified equally (mathematically).

[3] The choice is also made for the sake of quantifying estimation advantages of the approach in a non-biased fashion. For example, some other popular sequential sampling variants (two-boundary) do not have a closed-form solution, and may use approximations to estimate the model (Navarro & Fuss, 2009). The results of interest could thus depend on the specific approximation used.

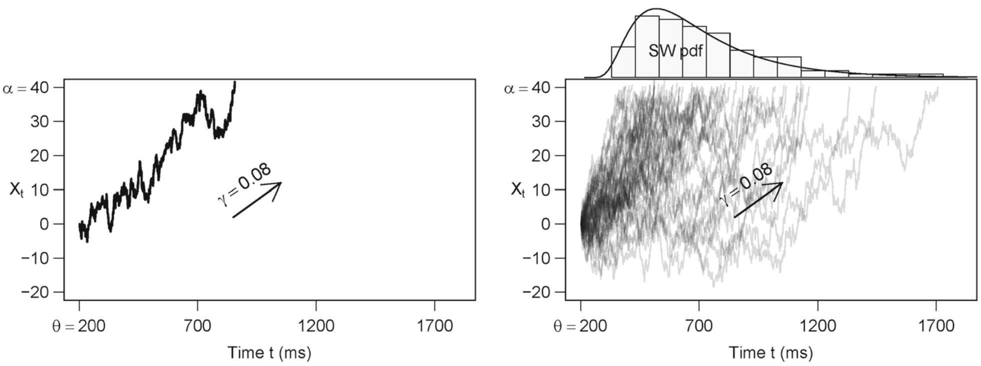

Fig. 1 Depiction of a canonical sequential sampling process as a cognitive-behavioral model, describing the RT data in the context of a latent quantity (e.g., signal) accumulating to threshold, α, at rate, γ, where θ accounts for the time lapsed outside of (around) this process. *Left*, a single trial is modeled with the parameters. *Right*, many trials (e.g., an experimental design cell) are modeled with the same parameter values, and these ultimately form a SW distribution shaped with the same signal accumulation parameters

with expected value $\alpha_c/\gamma_c + \theta_c$, and variance α_c/γ_c^3, for $\mathrm{RT}_{j_c} \in (\theta_c, \infty)$ and $\gamma_c, \alpha_c, \theta_c > 0$.

Bayesian estimation

In this section, a Bayesian estimation approach is developed for the multiple and covaried regression framework. We will apply it to the canonical SSM, and we refer to the augmented model as the Multi-Reg$_{\mathrm{SSM}}$. The Multi-Reg$_{\mathrm{SSM}}$ is summarized as a hierarchical cognitive model, in which the model's process parameters are hierarchically derived by a multiple and covaried regression structure. Furthermore, population distributions are applied at a layer above the hierarchical regressions by group (predictors, participants, items). The advantages of hierarchical population distributions have been discussed previously (see Rouder & Lu, 2005; Rouder, Lu, Speckman, et al., 2005; Rouder, Lu, Sun, et al., 2007). Readers more interested in the implementation results, rather than the technical Bayesian details, may proceed to "Application to simulated data".

First level: Multiple (and Covaried) regression that derives all cognitive parameters

When considering the potential approaches for estimating the Multi-Reg$_{\mathrm{SSM}}$ in the Bayesian framework, it is important to consider two essential mathematical properties of the model: (i) the sequential sampling parameters exist on the positive half-line, $\{\gamma_c, \alpha_c, \theta_c\} \in (0, \infty)$, and (ii) the regressions that hierarchically derive these parameters, share an error covariate structure (e.g., $\mathbf{\Sigma}$ as in Eq. 4) which models the correlations between process parameters. To satisfy (i), one can either implement an estimation algorithm that confines the regression sums in Eq. 4 to be always above 0, or alternatively, calculate these regression sums on the logarithmic scale which is constraint-free. To satisfy the covariate error-modeling of (ii), the three parameters can be modeled by a three-dimensional multivariate distribution in which the regression sums are the hierarchical means of the parameters, as M_{kc}, and $\mathbf{\Sigma}$ a 3 × 3 covariance matrix, handles the errors. From our simulation analyses, we found the method with the logarithmic scale to be practical and even advantageous for parameter comparisons (explained later in more detail), though subsequent work can be done to also develop optimal techniques for employing the alternative approach, i.e., the $M_{kc} > 0$ regression-sum-constraint approach.

Proceeding with the logarithmic approach, let $M_{kc} \in M_c = [M_{1c}, M_{2c}, M_{3c}]$ be the regression sums that hierarchically derive the three process model parameters $\{\gamma, \alpha, \theta\}$ for a given experimental cell. These regression sums will serve as the population means (e.g., process parameter values before error) as

$$\begin{aligned} M_{1c} &= \beta_\gamma X_c \\ M_{2c} &= \beta_\alpha X_c \\ M_{3c} &= \beta_\theta X_c\,. \end{aligned} \qquad (6)$$

By introducing the appropriate categorical coding of X_c, each $\beta_k X_c$ also includes potential person or item intercepts. Then let $\mathbf{\Sigma}$ be the 3 × 3 error covariate matrix that defines the noise around these sums on the logarithmic scale (e.g., for $\varepsilon_{1:3}$ in [3]). Then (i) and (ii) may be modeled by the following Bayesian priors,

$$\log\{\gamma_c, \alpha_c, \theta_c\} \sim \text{Multivariate Normal}(M_c, \mathbf{\Sigma}_{3\times 3})\,, \qquad (7)$$

where the logarithm of the set $\{\gamma_c, \alpha_c, \theta_c\}$ is modeled by the multivariate normal, and these parameter values on their natural scale can be easily obtained by taking the exponential. A notable advantage of this logarithmic scale approach concerns how the logarithmic locations, and modifications thereof of $\{\gamma_c, \alpha_c, \theta_c\}$, will correspond proportionally to their naturally scaled values, despite respectively existing in different magnitudes (e.g., tenths, tens, and hundreds on the natural scale). This feature will facilitate interpretation and comparison of these β weights in Eq. 6. For instance, although these regression weights $\{\beta_\gamma, \beta_\alpha, \beta_\theta\}$ are modeled to exist in comparable ranges (in respect to the multivariate normal) they will result in appropriately-scaled effect sizes of $\{\gamma_c, \alpha_c, \theta_c\}$ on their natural scale (see examples after Eq. 9).

Next, for these β values in Eq. 6, a natural prior distribution choice is the normal distribution. Though, keeping in mind that these β values include both a set of factor weights $\beta_{(f)}$ and potential person or item intercepts $\beta_{(i)}$, it is useful to distinguish for each set, appropriate prior mean μ and s.d. σ settings as follows:

$$\begin{array}{ll} \text{Factor Weights} & \text{Intercepts} \\ \beta_{\gamma w_{(f)}} \sim \text{Normal}(\mu_{f\gamma}, \sigma_{f\gamma}) & \beta_{\gamma w_{(i)}} \sim \text{Normal}(\mu_{i\gamma}, \sigma_{i\gamma}) \\ \beta_{\alpha w_{(f)}} \sim \text{Normal}(\mu_{f\alpha}, \sigma_{f\alpha}) & \beta_{\alpha w_{(i)}} \sim \text{Normal}(\mu_{i\alpha}, \sigma_{i\alpha}) \\ \beta_{\theta w_{(f)}} \sim \text{Normal}(\mu_{f\theta}, \sigma_{f\theta}) & \beta_{\theta w_{(i)}} \sim \text{Normal}(\mu_{i\theta}, \sigma_{i\theta}) \, , \end{array} \tag{8}$$

where $w \in \{1, \ldots, W\}$, W being the number of regressed factors (weights + intercepts).

Second level: Stochastic population distributions

A second level is formulated on the top of the model by-group (e.g., factor weights, intercepts) to pool information and constrain error in the regression weights themselves. This is done by using the population distribution approach previously mentioned. Specifically, for each parameter $k \in \{\gamma, \alpha, \theta\}$, the $\beta_{kw_{(f)}}$ factor weights are modeled by a hierarchical normal distribution with mean μ_{fk} and standard deviation σ_{fk}. Since as in our categorical regression coding, the first factor serves as baseline (that is, 0), this hierarchical modeling allows the factor effects to be predominantly positive or negative from baseline. Otherwise, if one used $\mu_{fk} = 0$, then the various $\beta_{kw_{(f)}}$ factors would be pushed by the prior to add to 0. A similar hierarchical modeling approach is also used for the $\beta_{kw_{(i)}}$ values that serve as the participant or item intercepts. Though in contrast, as intercepts which tend to locate the regression, greater prior mass is allocated to logarithmic ranges that correspond to the natural magnitudes of $\{\gamma, \alpha, \theta\}$ on the positive reals, as in Eq. 9.

These hierarchical distributions of the intercepts model the population level information, in which μ_{ik} and σ_{ik} quantify the population mean and standard deviation pertaining to the group of participants or items involved in the experiment. Based on our simulation analyses, reasonable priors for these hierarchical parameters μ and σ are the following:

$$\begin{array}{ll} \text{Population weights} & \text{Population intercepts} \\ \mu_{f\gamma} \sim \text{Normal}(0, 0.25) & \mu_{i\gamma} \sim \text{Normal}(-2.0, 0.5) \\ \mu_{f\alpha} \sim \text{Normal}(0, 0.25) & \mu_{i\alpha} \sim \text{Normal}(3.0, 0.5) \\ \mu_{f\theta} \sim \text{Normal}(0, 0.25) & \mu_{i\theta} \sim \text{Normal}(5.5, 0.5) \\ \sigma_{fk} \sim \text{Gamma}(4, 20) & \sigma_{ik} \sim \text{Gamma}(4, 40) \, . \end{array} \tag{9}$$

These priors provide a good compromise between code that is generalizable to a variety of data sets (e.g., data on different magnitudes of milliseconds) and model estimation stability. As shown in Eq. 9, for the weight population parameters (left), one can use the same mean and s.d. priors for each of the three parameters $\{\gamma, \alpha, \theta\}$, since adjustments from a location on the log scale are similarly proportional for various magnitudes on the positive real scale. Then for the intercepts (right), which generally serve to locate the regression (e.g., such as a regression mean), it is useful to utilize priors which provide probability mass for reasonable hierarchical mean values for $\{\gamma, \alpha, \theta\}$ in the positive reals.

For example, to get a grasp of the various magnitudes in the second level settings in Eq. 9, suppose the intercepts for γ, α, and θ are respectively -2.0, 3.0, and 5.5, which on the natural scale are values 0.135, 20, and 244. Then an observed β weight of 0.1 (for an $x = 1$) results in a shift to 0.149, 22, and 270, and an observed β weight of 0.05 results in 0.142, 21, and 257. As for the intercept population mean priors, note that greater prior probability is placed around population mean intercepts on the natural scale for $\gamma = 0.135, \alpha = 20$ and $\theta = 244$. A movement of 1 standard deviation (here 0.5) above, provides that these population mean increase to $\gamma = 0.223$, $\alpha = 33$ and $\theta = 403$ on the natural scale. Thus, the prior provides enough flexibility to accommodate various data ranges.

On modifying the priors The suggested priors will allow the Multi-Reg$_{\text{SSM}}$ to handle a variety of RT data from different experiments. The settings described above have shown to provide appropriate model stability and Bayesian mixing performance in our simulations and empirical applications (e.g., with respect to Eq. 9, using categorical experimental factors as weights–left column, and persons and/or items as intercepts–right column). However, in cases when researchers are attempting to fit non RT-data with this model (e.g., accumulation over months, years, for examples see Chhikara, 1988; Folks & Chhikara, 1978), much longer RTs, or use notably different regression forms than discussed herein, the researcher is encouraged to calibrate these prior settings in order to achieve optimal Bayesian mixing. Particularly since the proposed Multi-Reg$_{\text{SSM}}$ is quite complex to fit, informative prior settings are recommended.

Covariance structure

So far, we have specified the Bayesian priors for all of the parameters except for the covariance matrix $\mathbf{\Sigma}$. A good prior to optimize the estimation for $\mathbf{\Sigma}$ is formulated by decomposing the covariance matrix into $\mathbf{R}$, the Cholesky factor of the correlation matrix underlying $\mathbf{\Sigma}$, and a diagonal matrix, $\mathbf{S}$, containing the scalars in which $\mathbf{C} = \mathbf{R} \times \mathbf{S}$ provides the Cholesky factor of the covariance matrix of $\mathbf{\Sigma}$. As $\mathbf{C}$ being the Cholesky factor of $\mathbf{\Sigma}$, then $\mathbf{\Sigma} = \mathbf{C} \times \mathbf{C}^{\mathrm{T}}$. Such a practice is recommended by the developers of the Bayesian inference software Stan (Stan Development Team, 2015b), in which the suggested priors (see p. 72) are the following distributions:

$$\begin{aligned} \mathbf{R} &\sim \text{LKJ Cholesky}(8.0) \\ s_{kk} &\sim \text{Cauchy}(0, 0.025)\,, \end{aligned} \tag{10}$$

in which s_{kk} are the diagonal values of $\mathbf{S}$, and LKJ Cholesky is a prior distribution for the Cholesky factors of correlation matrices, as developed by (Lewandowski et al., 2009). Since the values which occupy s_{kk} are bounded to be greater than zero, the prior in Eq. 10 for s_{kk} serves as a half-Cauchy prior. Note that this approach of estimating the reduced elements of $\mathbf{\Sigma}$ is a development from previous approaches which estimated the full covariance matrix using the Wishart distribution (see Gelman & Hill 2007). Finally, also note that we lower the scale of the Cauchy prior to 0.025 (from e.g., 2.50), since a number of real data analysis fits with the Multi-Reg$_{\mathrm{SSM}}$ have shown that the hierarchical regression weights (or rather the respective regression residuals) that derive $\{\gamma_c, \alpha_c, \theta_c\}$, tend to occupy a markedly smaller range than for example, regressed RT values and their residuals.

The process model and data likelihood

The model notation is concluded with the RT$_j$ values being modeled on their natural scale by the SSM likelihood function in Eq. 5. Thus the Multi-Reg$_{\mathrm{SSM}}$ parameters on their natural scale may simply be accessed by

$$\{\gamma_c, \alpha_c, \theta_c\} = \exp^{\log\{\gamma_c, \alpha_c, \theta_c\}} \tag{11}$$

and then

$$\mathrm{RT}_{j_c} \sim f(\gamma_c, \alpha_c, \theta_c)\,, \tag{12}$$

as in Eq. 5.

Advantages with a Bayesian implementation

There are notable advantages of implementing the methodology in the Bayesian framework. Firstly, the model parameters (e.g., the regression coefficients) are estimated simultaneously, which can improve fit performance. This is contrasted with some maximum likelihood or deviance minimization techniques where estimations of one or multiple parameters are serially-used to derive the other parameters (e.g., from method of moment equations). Secondly, a distribution of estimations is provided for each parameter that readily provides a measure of posterior uncertainty in the results, which is an aspect not readily available in frequentist approaches that provide point estimates. Thirdly, error in estimation can be constrained by the appropriate use of priors. Furthermore, the cognitive parameters are simultaneously modeled in the context of a covariance structure (a multivariate Gaussian distribution with $\mathbf{\Sigma}_{K \times K}$), which handles parameter intercorrelation. Finally, the estimation technique combines various advantages of hierarchical modeling (Lee, 2011) by nested population distributions (Rouder & Lu, 2005; Rouder, Lu, Speckman, et al., 2005; Rouder, Lu, Sun, et al., 2007) and a regression structure (Vandekerckhove et al., 2011).

Fitting approach

The fitting approach we have developed can be summarized as follows. Firstly, as a maximally data-driven mixture model application, the approach will estimate a drift rate, γ_c, threshold, α_c, and non-accumulation time, θ_c, simultaneously for every unique design cell c of an experiment. Furthermore, the corresponding population means for each of these design cells are estimated, and so is the covariation between cognitive parameters. These parameters, and particularly the population means, are hierarchically derived by the respective regression models in Eq. 6, which are calculated independently of one another, except for the shared error covariate structure in Eq. 7 that models the process parameter correlation.

These regression coefficients directly quantify the experimental predictor main effects (also covariates and interactions, if specified, although not considered in the current development) $\beta_{(f)}$, as well as participant, item, or trait effects $\beta_{(i)}$. Notably, this modeling of experimental main effects and parameter correlations, that pools information across the experimental cells, economizes observation numbers (as in Table 1) and these quantities are retrieved in one step rather than through post-hoc analyses (e.g., ANOVA). Furthermore, the Bayesian fitting approach estimates these predictors simultaneously (improving reliability), provides informative measures of uncertainty (e.g., parameter posterior distributions), and integrates the parameters over the uncertainty of all other parameters: e.g., individual level estimates are propagated to the group level estimates and vice-versa.

Defining an indicator matrix based on the regressed coefficients

Suppose an experiment contains N unique design cells and W factors (including potential person or item intercepts). Then, to identify the model, the standard coding of the factor matrix $\mathbf{X}_{N\times W}$ in a typical linear regression design is recommended. In this paper, we will demonstrate and apply categorical coding. For example, when one codes four experimental categorical conditions, then $\mathbf{X}$ for participant 1 may resemble the following: setting the first level as baseline (left) or last level (right),

$$\left[\begin{array}{ccccc} \text{lvl1} & \text{lvl2} & \text{lvl3} & \text{lvl4} & \text{p1} \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{array}\right] \quad \text{or} \quad \left[\begin{array}{ccccc} \text{lvl1} & \text{lvl2} & \text{lvl3} & \text{lvl4} & \text{p1} \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right]. \tag{13}$$

In this way, one experimental condition level is removed to serve as baseline, and “p1” is the intercept for participant 1, which will serve as the baseline performance of the participant (e.g., on experimental condition 1, or 4, respectively). Each participant will hence possess an intercept. Therefore each participant provides an additional column in $\mathbf{X}$, populated by 1’s for each unique experimental design cell, and 0’s for where there are other participants (thus adding also e.g., four rows in $\mathbf{X}$ of Eq. 13). Furthermore, although not shown in the current example, item intercepts may also be introduced into the regression. Then as for categorical variables (e.g., experimental conditions), these will be coded to possess $L-1$ levels, so as to not form additional intercepts.

With the categorical coding approach, an experiment which possesses one factor with three levels, one factor with two levels, ten participants, and no item intercepts, should have $N = 3 \times 2 \times 10 = 60$ unique experimental cells and $K = 2 + 1 + 10 = 13$ regression coefficients. Hence, $\mathbf{X}$ is 60×13 and is populated by 1’s and 0’s; each β in Eq. 4 has length 13; and 60 sets of $\{\gamma_c, \alpha_c, \theta_c\}$ values are estimated jointly. Finally, note that other kinds of covariates (continuous, ordered) may be included, such as participant age and so forth. Alternatively for $\mathbf{X}$, one may also consider using effects coding rather than categorical coding.

Bayesian sampler settings

Through regular testing of the model in the hierarchical Bayesian estimation (HBE) framework, we have found that typically six chains, 1000 samples each, 500 of which is warm-up (burnin/adaptation phase in Stan), and a thinning[4] of 5, resulting in $100 \times 6 = 600$ final samples for analysis, are reasonable settings for appropriate mixing of the model (for a review of sampling terms, see Gelman et al., 2004). They also produce a good compromise between model fit performance and exceedingly long model run times. With these settings, our fits to real data have taken generally between 6 to 48 h with Stan and RStan (Stan Development Team 2015a, b) software, depending on data size. Note that due to the high complexity of this three-tiered modeling, we have found that it is very important to observe the parameter traceplots (chain mixing plots) from the fit, since occasionally a few chains may have difficulty in appropriately converging. Another solution to reduce the probability of this occurrence is to further optimize the prior settings, such as in Eq. 8, according to the data and model chosen. Typically when estimating six chains, when one or a few chains have difficulty appropriate converging, this will also be reflected in many of the chain convergence diagnostic values, $\hat{R}$’s, being greater than 1.10. Alternatively, one may address convergence issues by increasing the number of samples and burn-in iterations.

Application to simulated data

In this section, we demonstrate the advantages of the approach using the Multi-Reg$_{\text{SSM}}$ as an example. These tests involve parameter recovery across different experimental designs using simulated data analyses. The subsequent section follows with an example application to experimental data.

The results we present are from several large simulated data analyses that consist of varying the complexity of the experiment, and the number of available observations per unique experimental design cell (observation sizes: 250, 125, 60, 30, 20, 10, 5). The simulation involves the analysis of 30 data sets per observation size. Each of these 30 simulated data sets had hierarchical data-generating parameters that were randomly drawn from distributions similar to those as in Eq. 9. That is, random sets of β_{kw} values were generated (e.g., for each of the experimental factors and levels, participants, etc.) that hierarchically derive each of the parameters in $\Omega_c = \{\gamma_c, \alpha_c, \theta_c\}$, as well as the random parameter covariance matrix $\mathbf{\Sigma}_{K\times K}$.

In these simulated analyses, the Multi-Reg$_{\text{SSM}}$ is fit, which estimates these regression weights, $\{\beta_{\gamma w}, \beta_{\alpha w}, \beta_{\theta w}\}$, the three SSM parameters $\{\gamma_c, \alpha_c, \theta_c\}$ for each experimental design cell c, and their covariation, $\mathbf{\Sigma}_{K\times K}$. Then, the recovery of the parameters and the fit of the observed data’s quantiles are calculated, both of which have been previously used to assess appropriate model fit. This large analysis is performed twice in two different contexts: firstly,

[4] Note that thinning is only recommended to avoid memory issues, and is not necessary.

for a three-factor (3 × 3 × 2 levels) experimental design that has ten participants, in which one can expect notable benefits in pooling cross-cell information by the Multi-Reg$_{\mathrm{SSM}}$ framework, and, secondly, for a single factor (two levels) experimental design with ten participants, in which one can expect similar performance to a regular hierarchical Bayesian implementation, since only participants can be pooled in this case.

Furthermore, to examine the advantages of the approach in the context of other fitting methods, we will compare performance in parameter recovery to two other principal methods: standard hierarchical Bayesian estimation (HBE), and maximum likelihood estimation (MLE). In summary, the Multi-Reg$_{\mathrm{SSM}}$ method is *(i)* a 'Cross-cell HBE Multi-Reg' approach, in which the recurrence of all experimental effects (participants, items, conditions) across cells is utilized in estimation, thus economizing the number of observations needed. It also models parameter covariation. We compare the results to *(ii)* a 'By-cell HBE Non-Reg' approach, which is a standard HBE implementation defined in Appendix B, in which only the recurrence of participants across cells is utilized, offering only partial observation economization. We also compare the results to *(iii)* a 'By-cell MLE' approach recently developed by Anders et al. (2016), which uses MLE / quantile-minimization (QM) to fit a non-hierarchical version that models no recurrence across cells (hence does not economize observation numbers), but has shown to fit data adequately (at also low numbers of observations, e.g., $N = 20$), and in a much more rapid amount of time than the other two methods (within a few minutes).

Table 1, from $N = 250$ to $N = 5$ observations, provides the average parameter recovery trend for the three methods *(i)*, *(ii)*, and *(iii)*, of $\{\gamma, \alpha, \theta\}$, across 30 data set simulations using a three-factor (3 × 3 × 2) experimental design that has 10 participants. By Table 1, it is evident that the Multi-Reg$_{\mathrm{SSM}}$'s 'Cross-cell HBE Multi-Reg' approach provides an advantage over the partial/full 'by-cell' approaches in terms of markedly improved parameter recovery, even with as few as $N = 5$ observations per unique experimental design cell. At all observation size levels, the Multi-Reg$_{\mathrm{SSM}}$ performed better, and only at $N = 125$ observations do the 'By-cell' approaches begin to provide similar results. The remarkable result is the Multi-Reg$_{\mathrm{SSM}}$ approach provides comparable performance using only $1/6$ of the observation numbers ($N = 5$) than the traditional hierarchical Bayesian implementation ($N = 30$), and furthermore $1/12$ of the observations than a maximum likelihood implementation ($N = 60$). That is, the same experiment that uses $3 \times 3 \times 2 \times 30 = 540$ observations per participant, could have been performed with only $3 \times 3 \times 2 \times 5 = 90$ observations per participant. Aside from improved parameter recovery, this suggests how more predictors, conditions, or covariates could be included in cognitive model analyses when using this approach, or how more data-demanding versions of SSMs may be enabled to fit the data.

Figure 2 contains a visual plot of the parameter recovery results for the $N = 20$, $N = 10$, and $N = 5$ cases of Table 1 for the Multi-Reg$_{\mathrm{SSM}}$. These plots can reflect if there are systematic trends that may not be captured by the simple Pearson r correlation statistic. One can see that the model recovers the generating parameter values consistently well, with almost no strong outliers or biases. Finally, the right column of Fig. 2 provides a residual distribution diagnostic check. In cases of appropriate model fit, Anders et al. (2016) found that the distribution of standardized residuals (divided by $\sigma = \sqrt{\alpha_c/\gamma_c^3}$ from Eq. 5) of predicted versus observed RT deciles, tends to follow an ordered trend in magnitude. One can see that these decile residual distribution modes and variances have an ordered tendency, and occupy values generally between 0.05 to 0.25.

Next, it is reasonable to infer that the advantage the 'Cross-cell' approach provides over the 'By-cell' approaches diminishes as experimental designs become more simple. This kind of result is demonstrated in Table 2 for *(i)*, *(ii)*, and *(iii)*, in which we simulate a simpler experimental design: a single factor (two levels) with ten

Table 1 Process model parameter recovery, average Pearson correlations

	Three-factor design (3 × 3 × 2 levels), with ten participants								
	Cross-cell HBE Multi-Reg			By-cell HBE Non-Reg			By-cell MLE		
	Base-level SSM Parameters			Base-level SSM Parameters			Base-level SSM Parameters		
Observations	γ	α	θ	γ	α	θ	γ	α	θ
$N = 250$	0.99	0.99	1.00	0.98	0.98	0.99	0.93	0.88	0.99
$N = 125$	0.99	0.98	1.00	0.97	0.94	0.99	0.88	0.77	0.99
$N = 60$	0.97	0.94	1.00	0.90	0.80	0.99	0.79	0.63	0.98
$N = 30$	0.95	0.90	0.99	0.85	0.72	0.96	0.70	0.52	0.96
$N = 20$	0.94	0.86	0.99	0.81	0.56	0.92	0.61	0.42	0.94
$N = 10$	0.91	0.79	0.98	0.71	0.31	0.86	0.50	0.32	0.92
$N = 5$	0.86	0.67	0.98	0.66	0.19	0.78	0.40	0.15	0.88

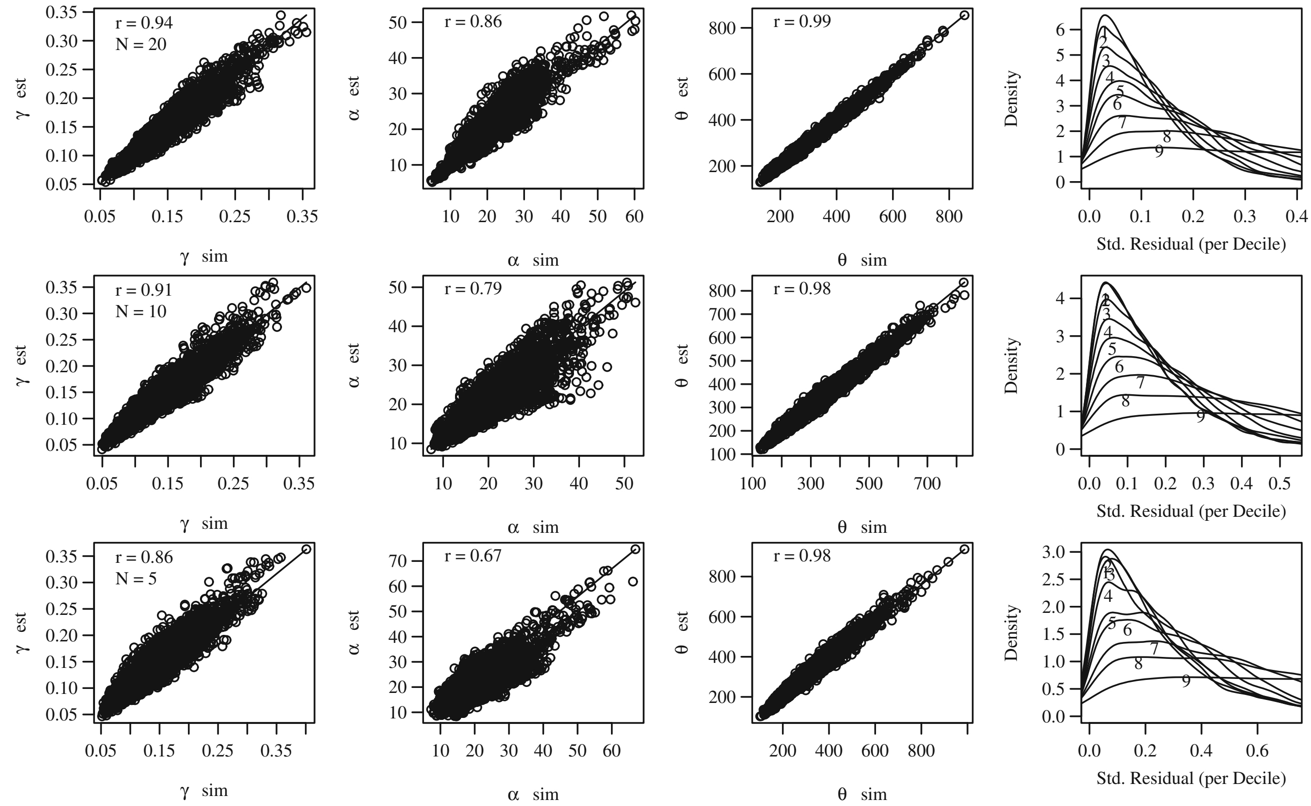

Fig. 2 Application of the proposed method on three sizes of data: each row in the plot corresponds to 30 sets of data, that respectively have $N = 20$, 10, and 5 observations per data set. The recovery corresponds to the results in Table 1, and consist of the posterior means of the parameters plotted against the generating value

participants. The recovery overall shows to be satisfactory in Table 2. However, comparing it to that of a more overlapping experimental design (three factors) as in Table 1, there is a notable reduction in the recovery strength, particularly for the cases of low numbers of observations per cell, e.g., $N = 5$ to $N = 30$. Then, comparing the recovery results of the Multi-Reg$_{\text{SSM}}$ with the other two methods, the advantage of the 'Cross-cell' approach of the Multi-Reg$_{\text{SSM}}$ is notably diminished. In this case, the improvement in recovery of the most difficult parameter to recover, α, is improved only by near 0.10 in the Pearson correlations, by using the 'Cross-cell' HBE approach versus the 'By-cell' HBE approach. Since the specifications are generally equal (only being able to pool participant information across cells), we

Table 2 Process Model parameter recovery, average Pearson correlations

	One-factor design (two levels), with ten participants								
	Cross-cell HBE Multi-Reg			By-cell HBE Non-Reg			By-cell MLE		
	Base-level SSM Parameters			Base-level SSM Parameters			Base-level SSM Parameters		
Observations	γ	α	θ	γ	α	θ	γ	α	θ
$N = 250$	0.99	0.97	1.00	0.98	0.98	0.99	0.93	0.88	0.99
$N = 125$	0.91	0.79	0.99	0.87	0.74	0.98	0.88	0.77	0.99
$N = 60$	0.86	0.72	0.98	0.78	0.64	0.96	0.79	0.60	0.97
$N = 30$	0.81	0.64	0.97	0.74	0.55	0.93	0.66	0.48	0.96
$N = 20$	0.75	0.56	0.95	0.67	0.46	0.92	0.62	0.40	0.95
$N = 10$	0.63	0.52	0.95	0.59	0.42	0.91	0.48	0.28	0.92
$N = 5$	0.51	0.32	0.92	0.47	0.24	0.87	0.35	0.18	0.86

speculate that the small advantages of the Multi-Reg$_{\mathrm{SSM}}$ over the standard HBE approach might be due to a modeling of the parameter covariation.

So far, we have only observed the recovery of the base-level SSM parameters at various experimental design complexities. However, we have not yet observed recovery of the regression weights, $\{\beta_\gamma, \beta_\alpha, \beta_\theta\}$, which hierarchically derive these cognitive parameters. Furthermore, it is worthwhile to note that these regression weights offer measurements of experimental effects, notably from a modeling that is aimed to disentangle experimental effect magnitudes from between-parameter correlations/covariation, e.g., as by $\boldsymbol{\Sigma}_{3\times3}$ from Eq. 7.

Table 3 provides the recovery results for $\{\beta_\gamma, \beta_\alpha, \beta_\theta\}$, in which the β_f subscript denotes regression weights for factors, and the β_i subscript denotes the participant intercepts. The recovery performance for the three-factor design related to Table 1 is provided on the left, and the one-factor design related to Table 2 on the right. The results indicate a strong recovery of the weights in the three factor design, and also an appropriate recovery for the single factor design, that each markedly improve with increasing observation sizes. Then, Fig. 3 provides visual plots of the regression weight (upper row) and intercept (bottom row) parameter recovery for the two lowest observation sizes, $N = 10$ and $N = 5$, for the three-factor experimental design case. Recovery is most tightly packed for weights related to the θ parameter, and secondly for the γ parameter (see also Fig. 2). The recovery of weights for the α parameter shows satisfactory trends, even at low numbers of observations in this three-factor design.

Finally, we have not yet observed recovery performance of the parameter inter-correlations that the modeling of $\boldsymbol{\Sigma}_{3\times3}$ from Eq. 7 may capture. There are a number of ways in which recovery of such inter-correlations may be measured. In Table 4, we provide the average absolute differences between Pearson r inter-parameter correlations, e.g., $|r_{\gamma_{est}\alpha_{est}} - r_{\gamma_{sim}\alpha_{sim}}| = \Delta_{r_{\gamma\alpha}}$ for each observation number size. An analysis of the results of Table 4 show a satisfactory recovery of the parameter inter-correlations that improves as observation numbers increase, and particularly for multi-factor experimental designs (left columns). At low observation numbers, the correlations consisting of the individual parameters that are most difficult (e.g., α, see Table 1) or easiest (e.g., θ) to recover, correspond to the magnitudes in the correlation recoveries.

For example, $\Delta_{r_{\gamma\theta}}$ is the most easily recovered, followed by $\Delta_{r_{\alpha\theta}}$, and then $\Delta_{r_{\gamma\alpha}}$. Then in respect to the single factor design, with ten participants, parameter inter-correlations are much more difficult to recover precisely at lower observation numbers $N = 5$ to $N = 20$, but notably becomes more appropriately on par near $N = 30$ observations and above. Augmentations also in the amount of participants may improve correlation recovery performance in these single factor designs.

Application to experimental data

In this section, the approach is demonstrated on a large data set involving a manual-gesture response task, in which 27 baboons (*Papio papio*) performed a visual search with contextual cues Goujon and Fagot (2013). The task consisted of searching for a visual target (the letter "T") that was embedded within configurations of distractors (letters "L"). The letters were either arranged predictively to locate the target (hence a contextual cue), or non-predictively (shuffled, without a cue). The baboons responded by touching the target on the display screen. The experimenters explored an animal model of statistical learning mechanisms in humans, specifically the ability to implicitly extract and utilize statistical redundancies within the environment for goal-directed behavior.

This data set was previously analyzed with the MLE/QM approach by Anders et al. (2016, e.g., the method compared

Table 3 Hierarchical parameter recovery (regression coefficients), average Pearson correlations

	Three-factor design						One-factor design					
	Factor weights			Intercepts			Factor weights			Intercepts		
Observations	$\beta_{f\gamma}$	$\beta_{f\alpha}$	$\beta_{f\theta}$	$\beta_{i\gamma}$	$\beta_{i\alpha}$	$\beta_{i\theta}$	$\beta_{f\gamma}$	$\beta_{f\alpha}$	$\beta_{f\theta}$	$\beta_{i\gamma}$	$\beta_{i\alpha}$	$\beta_{i\theta}$
$N = 250$	0.99	0.99	1.00	0.99	0.99	1.00	0.94	0.91	0.99	0.95	0.91	0.98
$N = 125$	0.99	0.95	1.00	0.99	0.99	1.00	0.91	0.87	0.98	0.91	0.81	0.99
$N = 60$	0.97	0.87	0.99	0.98	0.95	0.99	0.91	0.81	0.98	0.86	0.72	0.98
$N = 30$	0.91	0.84	0.99	0.97	0.93	0.99	0.88	0.72	0.96	0.84	0.63	0.97
$N = 20$	0.82	0.85	0.99	0.96	0.84	0.99	0.85	0.69	0.95	0.75	0.57	0.95
$N = 10$	0.92	0.79	0.99	0.92	0.78	0.98	0.59	0.55	0.94	0.68	0.58	0.96
$N = 5$	0.79	0.59	0.98	0.87	0.69	0.97	0.41	0.44	0.94	0.57	0.30	0.92

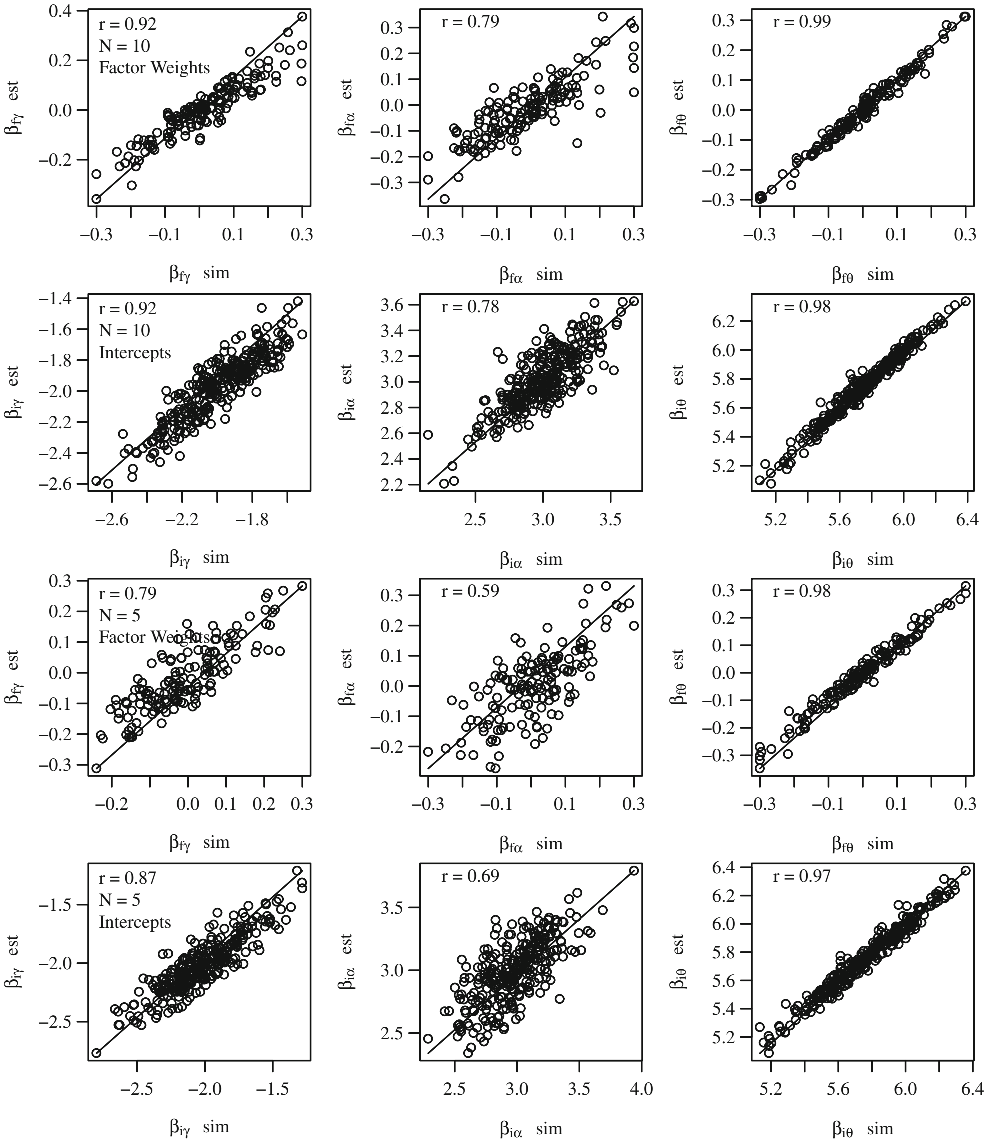

Fig. 3 Application of the proposed method on simulated data sizes of $N = 10$ (*top two rows*) and $N = 5$ (*bottom two rows*). Within each case, the *upper row* corresponds to the regression weights, while the *bottom row* corresponds to the participant intercepts. The recovery corresponds to the results in Table 3

Table 4 Process model parameter correlation recovery, $\Delta_{r_{\gamma\alpha}} = |r_{\gamma_{est}\alpha_{est}} - r_{\gamma_{sim}\alpha_{sim}}|$

	Three-factor design			One-factor design		
Observations	$\Delta_{r_{\gamma\alpha}}$	$\Delta_{r_{\gamma\theta}}$	$\Delta_{r_{\alpha\theta}}$	$\Delta_{r_{\gamma\alpha}}$	$\Delta_{r_{\gamma\theta}}$	$\Delta_{r_{\alpha\theta}}$
$N = 250$	0.06	0.04	0.05	0.18	0.13	0.21
$N = 125$	0.07	0.04	0.05	0.27	0.09	0.20
$N = 60$	0.09	0.06	0.08	0.32	0.15	0.19
$N = 30$	0.10	0.08	0.09	0.40	0.19	0.29
$N = 20$	0.18	0.07	0.14	0.58	0.19	0.37
$N = 10$	0.29	0.08	0.18	0.50	0.22	0.41
$N = 5$	0.30	0.12	0.20	0.66	0.24	0.68

in the right column of Tables 1 and 2) for the SSM, and the results can be compared here with the Multi-Reg$_{\text{SSM}}$ approach. As organized in the original publication, there are three meaningful partitions for this data set: the $C = 2$ non-predictive (control) vs. predictive contextual cue conditions; the $E = 40$ time-points (epochs) to observe training effects, in which every unit step in E consists of five blocks (each block contains 12 trials, and thus each E contains 60 trials); and the $P = 27$ individual baboons. The experiment hence consists of two factors (2×40) and 27 participants, leading to $N = 2 \times 40 \times 27 = 2160$ experimental design cells. However, only 2158 total cells were accessible since one baboon was absent from the experiment during the 36th epoch. The average RT distribution length (number of observations) per design cell is $\bar{L} = 30$, with standard deviation, $\mathrm{SD}(L) = 1.10$.

Model fit checks

In our Multi-Reg$_{\mathrm{SSM}}$ fit, we include these two learning conditions and 40 time points as regression factors, and use the last level of each factor as baseline (e.g., the right matrix in Eq. 13). Figure 4 provides the fit results. Beginning with the model goodness-of-fit checks, the right column of plots provides standard diagnostics (see Anders et al., 2016, for more detail). The top plot contains the deciles of all $N = 2158$ distributions fit with the Multi-Reg$_{\mathrm{SSM}}$. As can be seen, there is no systematic curvature in the plot and the SSM performs systematically well on the data set. The plot also captures the range of the data, and that there are about 4-6 of the 2158 cells fit in which their 9th decile (upper right of the plot) are notably underestimated by the Multi-Reg$_{\mathrm{SSM}}$. Then the middle plot provides the distribution of standardized residuals for each of the nine deciles (model-predicted RTs versus observed RTs) across the 2158 cells fit. Here it is shown that the fit optimally satisfies an ordering of distribution modes and variances. Finally, the bottom plot provides the sum standardized decile residuals, Δ, by cell, and its mean value, $\overline{\Delta} = 1.35$. Using the plot, one can also observe which cells are more poorly fit. Overall, $\rho_{\Delta\sigma}$ is small at -0.07, which supports Δ being a standardized residual statistic, as generally unbiased across varying observed RT cell distribution variances.

Main Results

The left column of Fig. 4 provides the parameter main-effect results of the analysis. These include the posterior mean regression coefficients β and their 95% Bayesian credible intervals, for the two experimental factors: the contextual cue learning condition (left) and training time points (epochs, right). Each row respectively corresponds to $\{\beta_\gamma, \beta_\alpha, \text{and } \beta_\theta\}$, which hierarchically derive the SSM parameters, and provide direct inferences about the experimental factor main effects on the cognitive process, without a need for post-hoc analyses (e.g., ANOVA). The dotted line indicates the baseline (the last level of each factor), which should be used to interpret these regression parameter values.

Beginning with the effect of the contextual cue condition on visual search latency in the left column of Fig. 4,

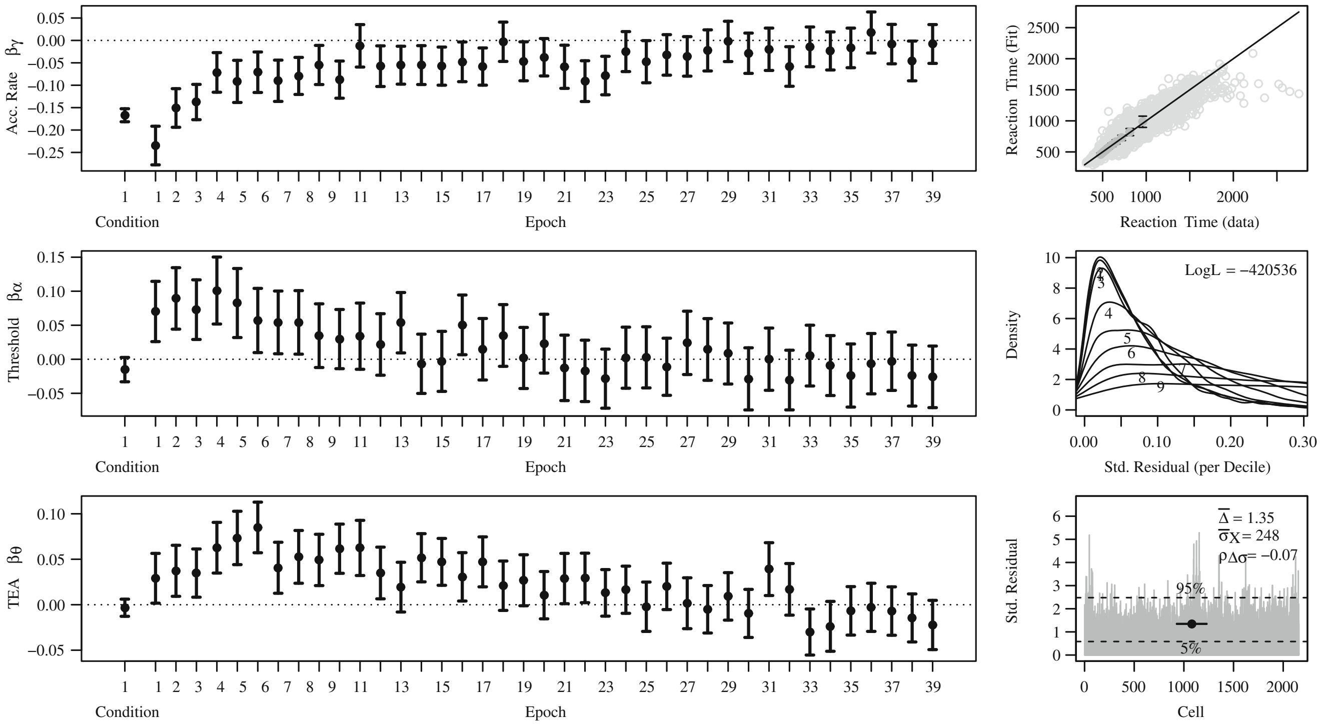

Fig. 4 The Multi-Reg$_{\mathrm{SSM}}$ fit to the baboon visual search task: (*left*) main-effect posterior mean parameter values with Bayesian credible intervals for each experimental factor; (*right*) model goodness-of-fit checks

the RT latencies are shown to be considerably faster due to a significant difference in the signal accumulation rate parameter, β_γ, when the cues are arranged in predictive patterns (baseline). Secondly, there is a small suggestive effect in reduced threshold (that could be interpreted as reduced response caution) in the control (no predictive cue) condition, as the narrow Bayesian interval overlaps the baseline. Finally, no effect was observed in the time external to accumulation, θ.

Next, regarding training effects on the visual search latencies, all parameters were affected in ways that support faster RTs with more training, yet in different patterns. Over the training interval, the signal accumulation rate (β_γ) increases rapidly between epochs 1-6 and then gradually settles between epochs 24 to 30. The response-triggering threshold (β_α) provides a steady decrease across training levels. The trend suggests that it may continue to improve with training beyond 40 epochs. Finally, non-accumulation time (β_θ) appears to show a slight increase between epochs $1 - 6$ before it begins a steady decreasing trend up to epoch 40.

Predicting missing data

On the topic of these main effects through the β_{kw} regressors, it is worthwhile to note that they may be used to predict the missing data cells in experiments. For example in this experiment, data is missing for one baboon in the 36th epoch (for both experimental conditions). We remark that since we have estimated the baboon's participant intercept for each of the parameters, the β_{kw} for the 36th epoch, and the β_{kw} for the control condition, these may be combined to accurately predict what its response times would have been for the missing epoch in each of the conditions.

Examining the cognitive parameters

Now we turn to examining how the hierarchical regression weights derive the base Multi-Reg$_{\text{SSM}}$ parameters, and how these Multi-Reg$_{\text{SSM}}$ parameters relate to the prior analysis results of the SSM with the MLE/QM method previously discussed. These two topics are respectively illustrated in Figs. 5 and 6. In each figure, the grey bars in the plots

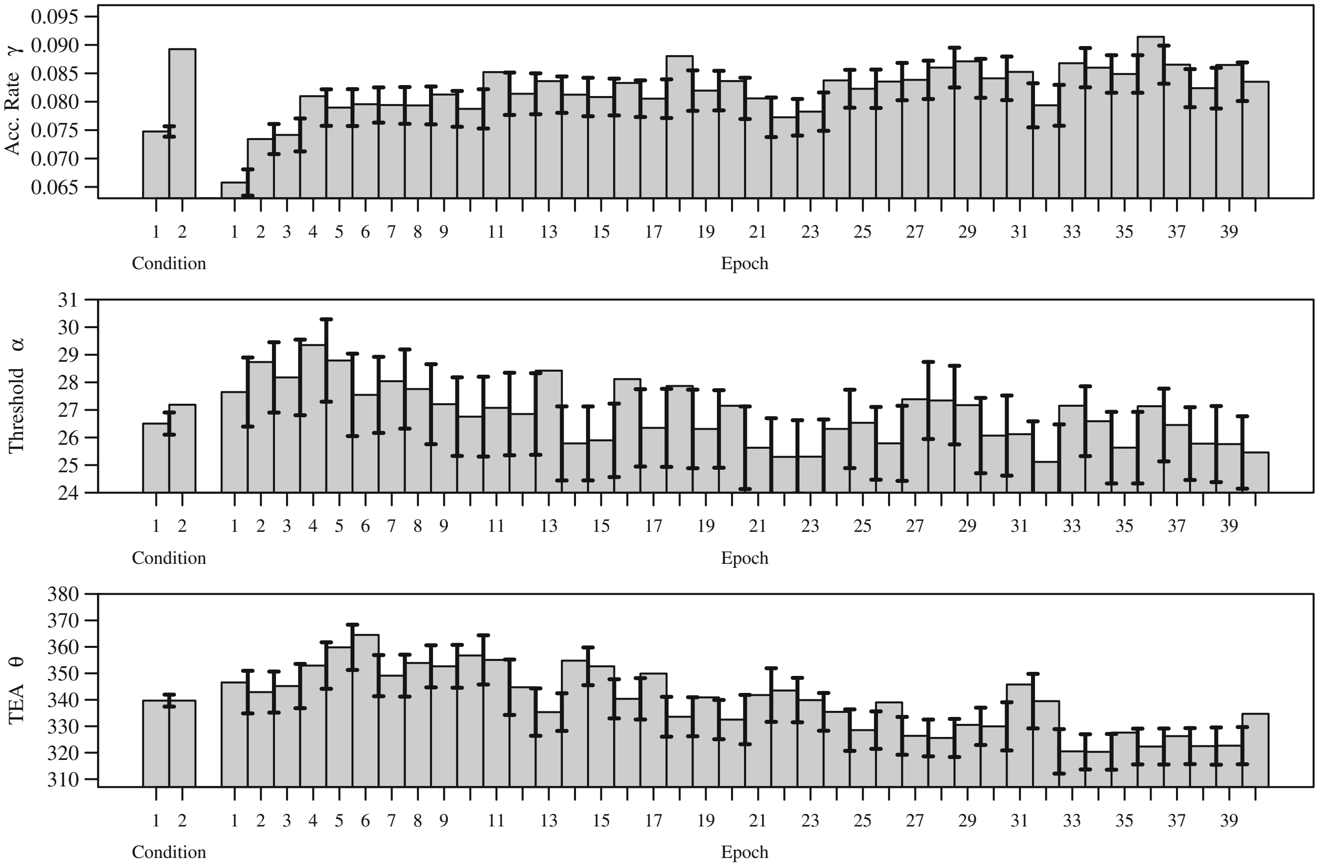

Fig. 5 The Multi-Reg$_{\text{SSM}}$'s process parameters that are hierarchically derived by the factor weights in Fig. 4, in combination with how they may be correlated to one another. Main-effect posterior mean parameter values are displayed with pairwise-difference (between posterior means) Bayesian 95% credible interval bars

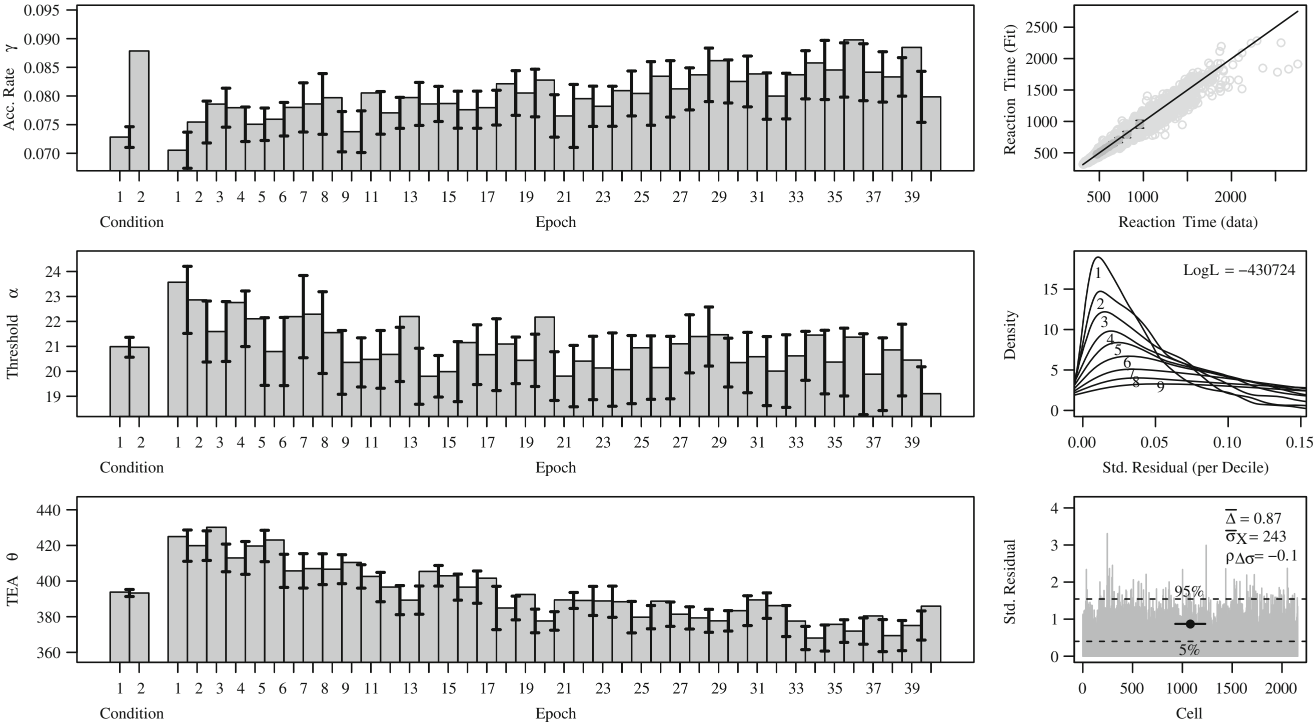

Fig. 6 The SSM fit to the manual response task by MLE/QM methods: (*left*) main-effect mean parameter values with pairwise-difference (within subject) standard error bars; (*right*) model goodness-of-fit checks. This figure is reproduced from the data fit in Anders et al. (2016)

are the main-effect mean process parameter values, and are calculated by the mean of within-subject posterior means for a given experimental level (as in also Anders, Riès, et al., 2015; Anders, Alario, et al., 2016). The interval bars for the Multi-Reg$_{\text{SSM}}$ represent the 95% Bayesian credible intervals for the pairwise-differences between adjacent experimental levels. The interval bars for the MLE/QM SSM represent the standard error of the mean, corrected by within-subject differences.

Firstly, in comparing the hierarchical predictors of the Multi-Reg$_{\text{SSM}}$'s $\{\beta_\gamma, \beta_\alpha$, and $\beta_\theta\}$ in Fig. 4 to the base-level parameters $\{\gamma, \alpha, \theta\}$ in Fig. 5, there is a generally strong correspondence between the results.[5] Note that in our simulation analyses, the base-level parameters (cognitive parameters) exhibited slightly better recovery performance than the hierarchical parameters (i.e. β coefficients), which can be a characteristic of many hierarchical models. Here, the only notable, but small difference between the results is in respect to the threshold α for the contextual cue condition. While the credible interval of β_α narrowly overlaps 0, the pairwise credible interval of α does not completely overlap the two condition levels. In both cases however, a potential effect on α is suggestive, though with low statistical power.

Interpretation Based on analysis of the cognitive parameters in Fig. 5 (γ, α, θ), it is clear that the presence of contextual cues allows for a much faster accumulation of information γ from the stimulus as to where the target is. The potential increase in α when there are contextual cues, suggests that the baboons may be more cautious to accumulate information from the predictive patterns (the cues) to locate the target, as compared to the control condition in which there is no information in the cues to locate the target. However, this slight delay in caution is overpowered by the much faster accumulation γ, so the RTs are still consistently faster during the contextual cue condition. Next, while the presence of contextual cues does not allow for a decrease in motor response time (modeled by θ), training over epochs clearly resulted in improvements. Training also improved the other parameters, which could be interpreted as over time, the baboons improve on processing the statistical redundancies in the environment, and this leads to

[5] Note that the dotted line of Fig. 4 is baseline, and thus respectively represents the second level of the contextual cue condition, or the last level of epoch.

faster accumulation of information γ from the stimulus, and less total information needed α from the stimulus in order to infer the location of the target.

Comparing the results to the MLE fitting method

Secondly, it may be interesting to compare these results with the previous method developed, the MLE/QM fitting method for the SSM by Anders et al. (2016, e.g., the method in the right column of Tables 1 and 2). The fit results using this method are contained in Fig. 6. The main differences observed are as follows. For the contextual cue condition, in contrast there is no suggested difference in threshold value between levels. Secondly for epoch (training effects), the Multi-Reg$_{\text{SSM}}$ suggested a logistic increasing trend of γ over time, and curved decreasing trends in α and θ that begin later (near epoch 6). In contrast, the 'by cell' MLE SSM approach suggests a linear improvement in γ over epochs, and curved decreasing trends in α and θ that begin immediately (near epoch 1).

The model fit diagnostics between the methods (the right columns of Figs. 4 and 6) provide interesting results as well. All three plots provide support that the MLE/QM method of fitting the SSM (which minimizes observed versus predicted quantiles), results in notably smaller quantile residuals than the Multi-Reg$_{\text{SSM}}$ which seeks to optimize the likelihood function. However, the Multi-Reg$_{\text{SSM}}$ achieves a markedly better log likelihood value. Thus, each method respectively won according to the criterion that it aimed to optimize. For example in respect to the quantile residuals between the two methods, MLE/QM$_{\text{SSM}}$ versus the Multi-Reg$_{\text{SSM}}$: the mean standardized residuals are $\overline{\Delta} = 0.87$ versus 1.35, and plot two displays the smaller standardized residual decile distribution modes achieved by the former (at 0.05 or below). Then in contrast, the MLE/QM$_{\text{SSM}}$ has a notably smaller log likelihood $-430, 724$ versus $-420, 536$.

This inspires future work for the best kinds of model diagnostic criteria to assess appropriate model fit, as quantile matching might not always represent the best parameter recovery. For example, see Tables 1 and 2 where the Multi-Reg$_{\text{SSM}}$ and standard by-cell Bayesian SSM recovered parameters better than the by-cell MLE/QM$_{\text{SSM}}$. However, it is also worth considering that quantiles as data points are more resilient to contaminant RT effects (Brown & Heathcote, 2003), which may adjust results in real data applications. However, we have performed simulation analyses like those in "Application to simulated data" using the quantiles as data points for the Bayesian models, and still the regular RTs used as data points achieved markedly better recovery performance.

Considering an interaction between factors

Lastly, one might be interested to fit this data with a Multi-Reg$_{\text{SSM}}$ that allows an interaction between the contextual cue condition, and the learning time points (epochs). We indeed fit such a model as well, but found minimal differences in the regression weights from Fig. 4, and the interaction β regression terms did not provide notably strong trends. Furthermore, the model fit checks were not very different from the right column in Fig. 4, and the log likelihood was only minimally improved at $-420, 529$ (interaction) versus $-420, 536$ (no interaction). Therefore, we retained the simpler model for the demonstration.

Discussion

We have demonstrated the advantages gained from nesting a multiple regression structure in a data-driven process model. These approaches are useful for analyzing experiments with multiple conditions, participants, and/or items of interest; and they are relevant for the models that would fit parameters along these experimental cells. Specifically, we developed a framework for how a full experimental design may be mapped into a multiple and covaried regression model that will maximally pool information from all recurrences of conditions between cells (factors and their levels, participants, and items). This information is used to hierarchically derive the process model parameters. The methodology allows for improved model parameter recovery at low numbers of observations, and consequently, allows for more experimental predictors to be simultaneously modeled. Simultaneous (joint) modeling of predictors may improve the predictive validity of a model, in contrast to separate analyses of predictors (e.g., conditions, participants, or items) which can cause misattribution errors (e.g., overestimation of effects, type I errors). For example, when certain predictors that significantly account for performance differences are not simultaneously modeled, a model may mistakenly attribute these performance differences to other predictors (Baayen, 2004; Baayen et al., 2008; Barr et al., 2013). Therefore, the proposed methodology may improve the cognitive-behavioral inferences made from experiments with data-driven process models.

The large simulation analyses included in "Application to simulated data" demonstrate that this methodology can provide a new standard over current practices in hierarchical modeling. The approach builds upon simpler nested regression structures previously proposed (Vandekerckhove et al., 2011), and also incorporates hierarchical population distributions (as discussed by Rouder & Lu, 2005; Rouder, Lu, Speckman, et al., 2005; Rouder, Lu, Sun, Speckman, et al., 2007). Specifically, our analyses demonstrated that for experiments having more than one factor (e.g., 2 levels), the methodology can achieve comparable performance to traditional hierarchical Bayesian modeling by using only a *fraction* of the observations. For example, comparable performance to standard hierarchical Bayesian modeling was achieved with only 1/6 of the observations, and with respect to standard maximum likelihood methods, only 1/12 of the observations. This is made possible when information can be pooled from the recurrences of conditions across the experimental cells (i.e. repeated measures). For example, these major advantages occurred in the simulation study for three-factor (e.g., 3 × 3 × 2 levels) experimental designs, but performance was otherwise similar to standard hierarchical methods for single factor experimental designs (2 level). We note that the designs were compared with equal numbers of observations per experimental cell.

In summary, the proposed multiple (and covaried) regression framework, used hierarchically for data-driven process models, can offer the following qualities: (i) an advantage in mixture modeling experimental data, by utilizing all recurring information across cells, (ii) markedly improved parameter recovery during low numbers of observations, allowing more predictors to be jointly modeled, (iii) pooling of information within groups (conditions, participants, and items) through modeling them with population distributions, which has been shown to improve performance, and (iv) pooling of information between experimental cells through the multiple regression design. Note that these nested regressions can also be used to incorporate covariate modeling (see Cavanagh et al., 2011; Frank et al., 2015). In addition, the framework offers (v) a direct modeling of experimental predictors for effects on cognitive parameters without a need for post-hoc analyses (e.g., ANOVA), (vi) the ability to easily predict missing data, based on having direct access to the predictors for each condition, participant, and item, and finally (vii) the potential to fit more complicated cognitive process models that require more data, since the approach economizes observations numbers.

As measurement and inference tools for experimental data, data-driven process models have been termed in some domains, as *cognitive psychometric* models (see Batchelder, 1998; Batchelder & Riefer, 1999; Riefer et al., 2002). We demonstrated the advantages of our approach using a canonical sequential sampling model (SSM), which is a family of models popularly used to account for performance differences in the time domain (e.g., reaction times). These models are not as thorough as multi-system or neural network models, but may be important empirical research tools.

Using simpler process models for empirical research is supportive of previous literature, with the notion that "less is more" when it comes to selecting a psychometric model for accurately estimating predictor and participant effects from experimental data. For example, van Ravenzwaaij, Donkin, and Vandekerckhove (2016) show that simpler SSMs with fewer parameters (e.g., the EZ-diffusion model, Wagenmakers, Van Der Maas, & Grasman, 2007) recovered the significant predictor effects in experiments better than their more complex counterparts, the Diffusion Decision Model (DDM, Ratcliff, 1978; Ratcliff & Smith, 2004; Ratcliff & McKoon, 2008). Such findings also highlight the growing differences between simple SSMs as apt measurement (quantitative, data-driven, psychometric) models, and others which are more suited for theoretical exploration (simulation testing, data-producing) of specific neural dynamics. More discussion on this topic is provided in Appendix C.

As discussed in "Process parameters as a function of a hierarchical multiple regression structure", our proposed framework can be easily applied to other SSMs, or other classes of models. With regard to other SSMs, the EZ-diffusion, DDM, LBA (Linear Ballistic Accumulator, Brown & Heathcote, 2008), and Q-/D-diffusion models (van der Maas, Molenaar, Maris, Kievit, & Borsboom, 2011) should be worthwhile candidates to further explore this framework. Moreover, it is worthwhile to note that a software package (in Python, Wiecki, Sofer, & Frank, 2013) currently exists for fitting predictors or covariates for the DDM. Several works using the package, and its tutorial, have generally emphasized simple regressions on one or two cognitive parameters in order to jointly model a cofactor or a trial-by-trial neural activity that could covary with that parameter. The research in the present paper develops upon this, and confirms with several large analyses, how a multiple and covaried regression structure on all cognitive parameters (mapping all experimental conditions) can improve information pooling, significantly improve parameter recovery at low numbers of observations, and allow more conditions to be jointly modeled. This is complementary to such covariate analyses, which may require additional observations. These performance advantages, and the capacity to model trial-by-trial covariates, may lead to nested regression structures becoming a new standard in hierarchical modeling. Though it is not clear if methods in current packages are appropriately calibrated to achieve model convergence with such advanced regression designs. Furthermore, parameter trade-off with the DDM should be an examined issue. Hence, current packages could test the Bayesian mixing performance (and parameter recovery) for full-fledged regression structures in these more complicated models, as well as implement the canonical SSM, which has various important empirical applications such as to visual search tasks, go/no-go tasks (driving simulations, cognitive load), lexical selection (picture naming/interference), saccades, general signal detection, and others. The code to implement the framework with the canonical SSM is provided in Appendix A.

While the proposed framework has already shown to be promising for data-driven process modeling techniques, there are several potential improvements left for future research. For example, although modeling potential for shorter and/or more complex experiments is heightened with the three-tiered hierarchical approach, model fitting time is likewise considerably lengthened. Currently model application times run near 24 h, even for a simple process model. This provides a challenge for sufficiently exploring the space of possible regression structure specifications, that for example Barr et al. (2013) may suggest examining. For instance, many different combinations of predictors, covariates, interactions, or even non-linear, quadratic regression equations could be explored for the best model fit.

While regression models pool information from all the recurrences of condition levels across experimental cells, it is made with the assumption that there are few interactions between conditions. Thus ideally, several versions of the hierarchical regression structure should be tested to select the hierarchical model that achieves the best quality of fit. Aside from this important note, future work could seek to (i) further refine estimation techniques (e.g., estimation algorithms, the Bayesian priors, such as for Eq. 9) of the framework, (ii) examine how the approach may function with extensions such as two- or three-way interactions, and non-linear regressions, (iii) implement and test extensions to other classes of cognitive models with the approach, or (iv) develop the multiple regression estimation to be calculated on parameter scales other than the logarithm (see "Bayesian estimation"), such as the natural scale (e.g., using other multivariate distributions than the Gaussian).

Author note We acknowledge funding by the European Research Council under the European Community's Seventh Framework Program (FP7/2007-2013 Grant agreement n° 263575), the Brain and Language Research Institute (Aix-Marseille Université : A*MIDEX grant ANR-11-IDEX-0001-02, LABEX grant ANR-11-LABX-0036), and the Institute of Language, Communication and the Brain (CONV grant ANR-16-CONV-0002). We thank the "Féderation de Recherche 3C" (Aix-Marseille Université) for institutional support.

Appendix A: Stan model code for the multi-reg$_{\mathrm{SSM}}$ (Cross-cell HBE multi-reg)

```
functions {
// log of the shifted Wald probability density function
   (RT value, vector of gamma, alpha, theta)
real sw_log(real RT, vector gat) {
return log(gat[2]/sqrt(2*pi()*((RT-gat[3])^3))) -
       ( (  (gat[2]-(gat[1]*(RT-gat[3])))^2)/(2*(RT-gat[3])) ) ;
  }
}

data{
int nobs; // number of RT observations (N related to Eq. 1)
int ncells; // number of unique experimental design cells
int ncoeff; // number of total coefficient weights
    (B, factor weights + intercepts)
int npers; // number of participants (intercepts)
int nfcoeff; // number of factor coefficient weights
    (factor weights: ncoeff-npers)
vector[nobs] RT; // the RT data, 1 RT per row in (Eq. 1)
int cellind[nobs]; // the cell ID for each RT observation
    (indexes for Eq. 12)
matrix[ncells, ncoeff] X; // factor level matrix in (Eq. 13)
}

parameters{
// these will be monitored and saved
matrix[ncoeff,3] B; // matrix of beta values
    (each column: gamma, alpha, theta)
vector[3] gat_log[ncells];//vector of gamma, alpha, theta by
    experimental cell in (Eq. 7)
cholesky_factor_corr[3] R; // 3x3 Cholesky factor of the correlation
    matrix in (Eq. 7)
vector<lower=0>[3] s; // scalar diagonal values for R to obtain C,
    or s_{kk} in (Eq. 10)
vector[3] mu_fac; // population factor means (left column, Eq. 9)
vector<lower=0>[3] sig_fac; // population factor sigmas
    (left column, Eq. 9)
vector[3] mu_int; // population intercept means (right column, Eq. 9)
vector<lower=0>[3] sig_int; // population intercept sigmas
    (right column, Eq. 9)
}

model{
vector[3] mu_log[ncells];// the hierarchical means in (Eq. 6)
matrix[3,3] C; // 3x3 Cholesky factor of the covariance matrix
    related to (Eq. 10)

// Second hierarchical level values in (Eq. 9)
mu_fac ~ normal(0, 0.25); // factor weights (left column)
sig_fac ~ gamma(4, 40);
mu_int[1] ~ normal(-2, 0.5); // intercepts (right column)
mu_int[2] ~ normal(3, 0.5);
mu_int[3] ~ normal(5.5, 0.5);
sig_int ~ gamma(4, 20);

// Priors for factor weights in (Eq. 8)
head(col(B,1), nfcoeff) ~ normal(mu_fac[1], sig_fac[1]);
head(col(B,2), nfcoeff) ~ normal(mu_fac[2], sig_fac[2]);
head(col(B,3), nfcoeff) ~ normal(mu_fac[3], sig_fac[3]);
// Priors for Person intercepts in (Eq. 8)
tail(col(B,1), npers) ~ normal(mu_int[1], sig_int[1]);
tail(col(B,2), npers) ~ normal(mu_int[2], sig_int[2]);
tail(col(B,3), npers) ~ normal(mu_int[3], sig_int[3]);
// Populations means in (Eq. 6, used in Eq.7)
for(i in 1:ncells){mu_log[i] = (X[i] * B)';}
// Prior values in (Eq. 12)
s ~ cauchy(0,0.025);
R ~ lkj_corr_cholesky(8);
// 3x3 Cholesky factor of the covariance matrix related to (Eq. 10)
C = diag_pre_multiply(s, R);
// MVN prior for gat on log scale in (Eq. 7)
gat_log ~ multi_normal_cholesky(mu_log, C);

// Likelihood in (Eq. 5)
for (j in 1:nobs) {RT[j] ~ sw(exp(gat_log[cellind[j]]));}
}

generated quantities {
// Simplified output of the desired quantities
vector[ncoeff] Bg; // regression weights for gamma (drift)
vector[ncoeff] Ba; // regression weights for alpha (threshold)
vector[ncoeff] Bt; // regression weights for theta (non-acc time)
vector[ncells] g; // gamma in positive reals (for each cell)
vector[ncells] a; // alpha in positive reals (for each cell)
vector[ncells] t; // theta in positive reals (for each cell)
matrix[3,3] covmat;// Cholesky factor of the covariance matrix
matrix[3,3] cormat; // correlation matrix
matrix[3,3] sigma; // covariance matrix
Bg = col(B,1);
Ba = col(B,2);
Bt = col(B,3);
for(i in 1:ncells){
g[i] = exp(gat_log[i,1]);
a[i] = exp(gat_log[i,2]);
t[i] = exp(gat_log[i,3]);
}
// See text near (Eq. 10)
cormat = R * R';
C = diag_pre_multiply(s, R);
sigma = C * C';
}
```

Appendix B: Hierarchical SSM model (By-cell HBE non-reg)

In this section, we develop a standard hierarchical Bayesian approach for the shifted Wald model (SWM), in which no multiple regression is involved, only hierarchical population distributions. This model is referred to as the 'By-cell HBE Non-Reg' model in Tables 1 and 2. When we apply the model in the simulated data analyses, we set the hierarchical parameters at the participant level.

Firstly, noting that all of the SWM parameters are located on the positive real line, a simple specification of their population distributions can be achieved with the gamma(α, β) distribution (where $\beta = 1/\text{scale}$):

$$\begin{aligned} \gamma_c &\sim \text{Gamma}(\mu_\gamma^2/\sigma_\gamma^2, \mu_\gamma/\sigma_\gamma^2) && \text{Participant Drift Rate} \\ \alpha_c &\sim \text{Gamma}(\mu_\alpha^2/\sigma_\alpha^2, \mu_\alpha/\sigma_\alpha^2) && \text{Participant Threshold} \\ \theta_c &\sim \text{Gamma}(\mu_\theta^2/\sigma_\theta^2, \mu_\theta/\sigma_\theta^2) && \text{Participant Non-Accumulation Time}\,, \end{aligned} \tag{14}$$

in which each experimental cell $c \in 1, \ldots, C$ has parameters $[\gamma_c, \alpha_c, \theta_c]$, drawn from a shared population distribution. In this way with $\alpha = \mu^2/\sigma^2$ and $\beta = \mu/\sigma^2$, the population parameters μ and σ^2 will respectively relate to the location and variance of the gamma-distributed individual parameters; and the distribution will resemble a normal distribution with support on the positive half-line (see Anders & Batchelder, 2013; Kruschke, 2011, for other examples).

Then, since the hierarchical parameters μ and σ are also distributed on the positive half-line, one may similarly consider the gamma distribution as a suitable prior.

$$\begin{aligned} \mu_\gamma &\sim \text{Gamma}(0.20^2/0.10^2, 0.20/0.10^2) & \sigma_\gamma &\sim \text{Gamma}(12.5, 250) \\ \mu_\alpha &\sim \text{Gamma}(20^2/10^2, 20/10^2) & \sigma_\alpha &\sim \text{Gamma}(9, 1.50) \\ \mu_\theta &\sim \text{Gamma}(600^2/300^2, 600/300^2) & \sigma_\theta &\sim \text{Gamma}(9, 0.03) \end{aligned} \tag{15}$$

These settings are reasonable for RT data in which most of the probability mass is within the range of 300 to 4000 ms. For example, much of our previous work using the model (Anders et al., 2016; Anders et al., 2015) for such RT ranges, has found drift rates to generally reside between .04 to .30, thresholds between 10 and 40 and non-accumulation times between 300 and 600 ms. Therefore, if one were to analyze much longer RTs, such as between 10,000 and 30,000 ms, one should augment these ranges.

Given these priors and hierarchical settings, the Bayesian framework is completed by the multiplication of these priors and the model likelihood. The SWM possesses the following closed-form likelihood:

$$f(\gamma_c, \alpha_c, \theta_c \mid \text{RT}_{j_c}) = \frac{\alpha_c}{\sqrt{2\pi(\text{RT}_{j_c} - \theta_c)^3}} \cdot \exp\left\{-\frac{[\alpha_c - \gamma_c(\text{RT}_{j_c} - \theta_c)]^2}{2(\text{RT}_{j_c} - \theta_c)}\right\}, \tag{16}$$

with expected value $\alpha_c/\gamma_c + \theta_c$, and variance α_c/γ_c^3, for $\text{RT}_{j_c} > \theta_c$.

Appendix C: Supplementary information regarding sequential sampling models and their selection

Varieties of extended sequential sampling models The main varieties of sequential sampling models may be classified by (i) the single accumulator versus multiple accumulator distinction and, within these classes, by (ii) modified individual parameter conditions or trends (e.g., of $\{\gamma, \alpha, \theta\}$, or X_t) within or between trials (e.g., see Busemeyer & Townsend, 1992). For example within trials, a few of these modifications may consist of the decay of X_t, a diminishing threshold α over time (Hawkins, Forstmann, Wagenmakers, Ratcliff, & Brown, 2015), substages with different γ values (Diederich & Busemeyer, in review; Holmes, Trueblood, & Heathcote, 2016), or X_t reversing directions (Busemeyer and Townsend, 1993; Diederich & Busemeyer, 2006), which is a form of an Ornstein–Uhlenbeck process (Busemeyer & Townsend, 1992).

The difference in (i) using single versus multiple accumulators, may primarily concern how one prefers to model X_t for the observed response in the context of n-alternatives. The currently most popular single accumulator model is the Drift Diffusion Model (DDM, Ratcliff, 1978; Ratcliff & McKoon, 2008), which is appropriate for $n = 2$ alternative forced choice (2AFC) tasks, and is modeled by X_t moving to an upper or lower boundary. In this single accumulator framework, $n > 2$ choices would be modeled by the movement of X_t in an n-dimensional object to one of the thresholds (being the object's sides, Stroock & Varadhan, 2016, see also Smith 1979). Typically in this case, the augmented evidence for one choice provides reduced evidence against the other choices. In contrast, the multiple accumulator approach models the activation of each alternative with a separate, single boundary accumulator (as in Fig. 1); these accumulators race, and the first accumulator to meet its threshold is modeled as the performed behavior. In this case, it is not a forced property of the model that evidence in favor of one alternative is reduced for the others

(LaBerge, 1962; Usher, Olami, & McClelland, 2002; Brown & Heathcote, 2008), but it may be an added parameter (Usher & McClelland, 2001).

Why not directly use a model that quantifies the activation of all alternatives, or of inhibition? Except for carefully designed experiments where only two or three alternative behaviors are possible, trying to model ≥ 4 alternatives, or even inhibition/decay along several conditions of an experiment, currently leads to models that are too complex to effectively serve as measurement tools. Hence, although these are interesting models, they correspond less to the objectives of the analysis developed herein. Such models are currently stronger as data-producing models for theoretical exploration. For example, to our knowledge there have been no publications yet that estimate from the data the activation of four or more alternatives along several experimental conditions. Furthermore, in respect to recovering *inhibition* parameters, a recent publication shows that even with advanced fitting methods, unrealistic experimental observation numbers are needed: that is *10,000* or more trials from a single subject, in a three alternative case (Miletić, Turner, Forstmann, & van Maanen, 2017, for the Leaky Competing Accumulator Model).

The canonical SSM is a solid psychometric model that can measure many predictors from the data. It is easily generalizable to many kinds of experiments. It can use predictors in regression modeling to potentially infer additional accumulation dynamics (e.g., decay over trials, interference from conditions). It does not have significant parameter trade-off. It has a closed-form solution as in Eq. 5. For all these reasons, it can serve as a useful data-driven process model for RT data. However, while this canonical SSM is indeed practical and generalizable to a wide range of experimental designs, we note that it is not the ideal model when more specialized sequential sampling models, which more closely describe the experimental task, are capable of being fit to the data. For example, for two-alternative forced choice tasks (2AFC), using our multiple regression approach on a two-boundary (DDM), or two-accumulator model (Brown & Heathcote, 2008, Race / Linear Ballistic Accumulator), would be preferred.

References

Anders, R., Alario, F.-X., & Van Maanen, L. (2016). The shifted Wald distribution for response time data analysis. *Psychological Methods*, 21.

Anders, R., & Batchelder, W. H. (2012). Cultural consensus theory for multiple consensus truths. *Journal for Mathematical Psychology*, *56*, 452–469.

Anders, R., & Batchelder, W. H. (2013). Cultural consensus theory for the ordinal data case. *Psychometrika*, *80*, 151–181.

Anders, R., Riès, S., van Maanen, L., & Alario, F.-X. (2015). Evidence accumulation as a model for lexical selection. *Cognitive Psychology*, *82*, 57–73.

Anderson, J. R. (1996). ACT: A simple theory of complex cognition. *American Psychologist*, *51*, 355.

Baayen, R. H. (2004). Statistics in psycholinguistics: A critique of some current gold standards. *Mental Lexicon Working Papers*, *1*, 1–47.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*, 390–412.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*, 255–278.

Batchelder, W. H. (1998). Multinomial processing tree models and psychological assessment. *Psychological Assessment*, *10*, 331.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review*, *6*, 57–86.

Brown, S., & Heathcote, A. (2003). QMLE: Fast, robust, and efficient estimation of distribution functions based on quantiles. *Behavior Research Methods, Instruments, & Computers*, *35*, 485–492.

Brown, S., & Heathcote, A. (2008). The simplest complete model of choice response time: Linear ballistic accumulation. *Cognitive Psychology*, *57*, 153–178.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Busemeyer, J. R., & Townsend, J. T. (1992). Fundamental derivations from decision field theory. *Mathematical Social Sciences*, *23*, 255–282.

Busemeyer, J. R., & Townsend, J. T. (1993). Decision field theory: A dynamic-cognitive approach to decision making in an uncertain environment. *Psychological Review*, *100*, 432.

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*, 1462–1467.

Chhikara, R. (1988). *The Inverse Gaussian Distribution: Theory, Methodology, and Applications* volume 95. CRC Press.

Cohen, J. (1968). Multiple regression as a general data-analytic system. *Psychological Bulletin*, *70*, 426.

Cohen, Y., & Cohen, J. Y. (1988). Analysis of variance. *Statistics and data with R: An applied approach through examples*, pp. 463–509.

Dehaene, S. (2008). Conscious and nonconscious processes: Distinct forms of evidence accumulation. *Better Than Conscious* pp. 22–49.

Diederich, A., & Busemeyer, J. R. (2006). Modeling the effects of payoff on response bias in a perceptual discrimination task: Bound-change, drift-rate-change, or two-stage-processing hypothesis. *Perception & Psychophysics*, *68*, 194–207.

Diederich, A., & Busemeyer, J. R. (in review). Multi-stage sequential sampling model of multi-attribute decision making.

Everitt, B. S. (1981). *Finite mixture distributions*. Wiley Online Library.

Folks, J., & Chhikara, R. (1978). The inverse Gaussian distribution and its statistical application–a review. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 263–289.

Frank, M. J., Gagne, C., Nyhus, E., Masters, S., Wiecki, T. V., Cavanagh, J. F., & Badre, D. (2015). fMRI and EEG predictors of dynamic decision parameters during human reinforcement learning. *The Journal of Neuroscience*, *35*, 485–494.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*, 2nd ed. Boca Raton, FL.: Chapman & Hall/CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and hierarchical/multilevel models*. Cambridge, UK: Cambridge University Press.

Gerstein, G. L., & Mandelbrot, B. (1964). Random walk models for the spike activity of a single neuron. *Biophysical Journal*, *4*, 41–68.

Goujon, A., & Fagot, J. (2013). Learning of spatial statistics in nonhuman primates: Contextual cueing in baboons (*Papio papio*). *Behavioural Brain Research*, *247*, 101–109.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics*. New York: Wiley.

Hawkins, G. E., Forstmann, B. U., Wagenmakers, E.-J., Ratcliff, R., & Brown, S. D. (2015). Revisiting the evidence for collapsing boundaries and urgency signals in perceptual decision-making. *The Journal of Neuroscience*, *35*, 2476–2484.

Holmes, W. R., Trueblood, J. S., & Heathcote, A. (2016). A new framework for modeling decisions about changing information: The piecewise linear ballistic accumulator model. *Cognitive Psychology*, *85*, 1–29.

Howell, D. C. (2012). *Statistical methods for psychology*. Cengage Learning.

Iversen, G. R., & Norpoth, H. (1987). *Analysis of variance*. 1. Sage.

Jolliffe, I.T. (2002). *Principle component analysis*, 2nd Edn. New York: Springer-Verlag.

Kelly, S. P., & O'Connell, R. G. (2013). Internal and external influences on the rate of sensory evidence accumulation in the human brain. *The Journal of Neuroscience*, *33*, 19434–19441.

Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R and BUGS*. New York: Academic Press.

LaBerge, D. (1962). A recruitment theory of simple behavior. *Psychometrika*, *27*, 375–396.

Laming, D. R. J. (1968). *Information theory of choice-reaction times*. Academic Press.

Lazarsfeld, P. F. (1959). *Latent structure analysis volume 3*. NY: McGraw-Hill.

Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.

Lee, M. D., & Wagenmakers, E.-J. (2014), *Bayesian cognitive modeling: A practical course*. Cambridge University Press.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*, 1989–2001.

Van der Linden, W. J., & Hambleton, R. K. (1997), *Handbook of modern item response theory*. Springer.

Luce, R. D. (1986). *Response times: Their role in inferring elementary mental organization*. Oxford University Press.

van der Maas, H. L., Molenaar, D., Maris, G., Kievit, R. A., & Borsboom, D. (2011). Cognitive psychology meets psychometric theory: On the relation between process models for decision making and latent variable models for individual differences. *Psychological Review*, *118*, 339.

Miletić, S., Turner, B. M., Forstmann, B. U., & van Maanen, L. (2017). Parameter recovery for the leaky competing accumulator model. *Journal of Mathematical Psychology*, *76*, 25–50.

Navarro, D. J., & Fuss, I. G. (2009). Fast and accurate calculations for first-passage times in Wiener diffusion models. *Journal of Mathematical Psychology*, *53*, 222–230.

O'Connell, R. G., Dockree, P. M., & Kelly, S. P. (2012). A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nature Neuroscience*, *15*, 1729–1735.

Oravecz, Z., Anders, R., & Batchelder, W. H. (2015). Hierarchical Bayesian modeling for test theory without an answer key. *Psychometrika*, *80*, 341–364.

Pike, R. (1973). Response latency models for signal detection. *Psychological Review*, *80*, 53.

Ratcliff, R. (1978). A theory of memory retrieval. *Psychological Review*, *85*, 59.

Ratcliff, R., Gomez, P., & McKoon, G. (2004). A diffusion model account of the lexical decision task. *Psychological Review*, *111*, 159.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*, 873–922.

Ratcliff, R., & Smith, P. L. (2004). A comparison of sequential sampling models for two-choice reaction time. *Psychological Review*, *111*, 333.

Ratcliff, R., Thompson, C. A., & McKoon, G. (2015). Modeling individual differences in response time and accuracy in numeracy. *Cognition*, *137*, 115–136.

Ratcliff, R., Van Zandt, T., & McKoon, G. (1999). Connectionist and diffusion models of reaction time. *Psychological Review*, *106*, 261–300.

van Ravenzwaaij, D., Donkin, C., & Vandekerckhove, J. (2016). The EZ diffusion model provides a powerful test of simple empirical effects. *Psychonomic Bulletin & Review*, pp. 1–10.

Riefer, D. M., Knapp, B. R., Batchelder, W. H., Bamber, D., & Manifold, V. (2002). Cognitive psychometrics: Assessing storage and retrieval deficits in special populations with multinomial processing tree models. *Psychological Assessment*, *14*, 184.

Rouder, J. N., & Lu, J. (2005). An introduction to Bayesian hierarchical models with an application in the theory of signal detection. *Psychonomic Bulletin & Review*, *12*, 573–604.

Rouder, J. N., Lu, J., Speckman, P., Sun, D., & Jiang, Y. (2005). A hierarchical model for estimating response time distributions. *Psychonomic Bulletin & Review*, *12*, 195–223.

Rouder, J. N., Lu, J., Sun, D., Speckman, P., Morey, R., & Naveh-Benjamin, M. (2007). Signal detection models with random participant and item effects. *Psychometrika*, *72*, 621–642.

Rouder, J. N., Morey, R. D., & Pratte, M. S. (2013). Hierarchical Bayesian models. *Practice*, *1*, 10.

Scheibehenne, B., & Pachur, T. (2015). Using Bayesian hierarchical parameter estimation to assess the generalizability of cognitive models of choice. *Psychonomic Bulletin & Review*, *22*, 391–407.

Smith, P. (2016). Diffusion theory of decision making in continuous report. *Psychological Review*.

Stan Development Team (2015a). RStan: The R interface to Stan, version 2.8.0.

Stan Development Team (2015b). *Stan Modeling Language Users Guide and Reference Manual, Version 2.8.0*.

Stone, M. (1960). Models for choice-reaction time. *Psychometrika*, *25*, 251–260.

Stroock, D. W., & Varadhan, S. S. (1979). Multidimensional diffusion processes, volume 233 of grundlehren der mathematischen wissenschaften [fundamental principles of mathematical sciences].

Townsend, J. T., & Ashby, F. G. (1983). *Stochastic modeling of elementary psychological processes*. CUP Archive.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review*, *108*, 550.

Usher, M., Olami, Z., & McClelland, J. L. (2002). Hick's law in a stochastic race model with speed–accuracy tradeoff. *Journal of Mathematical Psychology*, *46*, 704–715.

Vandekerckhove, J. (2014). A cognitive latent variable model for the simultaneous analysis of behavioral and personality data. *Journal of Mathematical Psychology*, *60*, 58–71.

Vandekerckhove, J., Tuerlinckx, F., & Lee, M. D. (2011). Hierarchical diffusion models for two-choice response times. *Psychological methods*, *16*, 44.

Wagenmakers, E.-J., Van Der Maas, H. L., & Grasman, R. P. (2007). An EZ-diffusion model for response time and accuracy. *Psychonomic Bulletin & Review*, *14*, 3–22.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the drift-diffusion model in Python. *Frontiers in Neuroinformatics*, *7*, 14.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.