Journal of Mathematical Psychology 55 (2011) 416–429

Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Generalizing parametric models by introducing trial-by-trial parameter variability: The case of TVA

Mads Dyrholm [a,*], Søren Kyllingsbæk [a], Thomas Espeseth [b], Claus Bundesen [a]

[a] Center for Visual Cognition, Department of Psychology, University of Copenhagen, Øster Farimagsgade 2A, DK-1353 Copenhagen K, Denmark

[b] Center for the Study of Human Cognition, Department of Psychology, University of Oslo, Norway

ARTICLE INFO

Article history:
Received 24 February 2011
Received in revised form
16 August 2011
Available online 16 September 2011

Keywords:
Visual attention
Trial-by-trial variability
Model selection
VSTM capacity
TVA

ABSTRACT

We identify two biases in the traditional use of Bundesen's Theory of Visual Attention (TVA) and show that they can be substantially reduced by introducing trial-by-trial variability in the model. We analyze whole and partial report data from a comprehensive empirical study with 347 participants and elaborate on Bayesian model selection theory for quantifying the advantage of trial-by-trial generalization in general. The analysis provides strong evidence of trial-by-trial variation in both the VSTM capacity parameter and perceptual threshold parameter of TVA. On average, the VSTM capacity bias was found to be at least half an item, while the perceptual threshold parameter was found to be underestimated by about 2 ms.

1. Introduction

Most psychological experiments consist of a number of experimental conditions, each represented by a certain number of trials. The psychological state of a participant is normally assumed to be constant across trials within a given condition, which is sometimes a plausible simplifying assumption, but at other times a rough simplification (Van Zandt & Ratcliff, 1995). Many parameters are likely to show substantial variation from trial to trial, and such variation may induce a bias in the estimates for the mean values of the parameters if the parameters are assumed to be constant. By "bias" we mean a systematic tendency to misestimate on average. Two examples will clarify this issue: parameters t_0, the threshold of conscious perception, and K, the visual short-term memory (VSTM) span, in the Theory of Visual Attention (TVA; Bundesen, 1990).

TVA has been widely applied to the processing of briefly presented visual displays (for example, in whole and partial report tasks). Consider the processing of an array of elements exposed for a certain duration and then succeeded by an effective visual mask. The task of the participant is to report as many elements as possible from the display (whole report). For simplicity we assume that the participant performs the task without guessing.[1] Parameter t_0 is defined as the longest ineffective exposure duration of the stimulus display (Shibuya & Bundesen, 1988). Thus, information can be gained from the stimulus exposure if the exposure duration t is longer than t_0. At exposure durations shorter than t_0, the probability of reporting any element is zero. Parameter t_0 is normally assumed to be constant for any given stimulus. However, it seems likely that t_0 varies somewhat from trial to trial, rather than being strictly constant. The effect of such variation will be highly systematic. Correct reports will occur on some trials in which t_0 takes on a value which lies below its mean. Correct reports on such trials will drive the estimate of t_0 derived from the traditional model (the model assuming that t_0 is constant) down below the true mean value of t_0. Thus, the mean value of parameter t_0 will be underestimated. When t_0 is assumed to be constant, but

[1] In typical experiments applying TVA, the participants are instructed to refrain from blind guessing and report those targets, and only those targets, they are "fairly certain" they have seen. The data analysis is based on the simplifying assumptions that both blind guessing and perceptual confusions among stimuli (i.e., incorrect perceptual categorizations of stimuli) can be neglected. How far these assumptions are satisfied is sometimes evaluated by (a) use of catch trials on which the exposure duration of the targets is zero (i.e., the targets are omitted and only the masks are presented) and (b) analysis of the frequencies of erroneous reports. When both blind guessing and perceptual confusions can be neglected, the probability that an element x can be correctly reported equals the probability that (an appropriate categorization of) the element is encoded in VSTM (for a TVA-based analysis of processing of mutually confusable singly presented stimuli, see Kyllingsbæk, Markussen, & Bundesen, in press).

* Corresponding author.
E-mail addresses: mads.dyrholm@psy.ku.dk (M. Dyrholm), soeren.kyllingsbaek@psy.ku.dk (S. Kyllingsbæk), thomas.espeseth@psykologi.uio.no (T. Espeseth), claus.bundesen@psy.ku.dk (C. Bundesen).

0022-2496/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmp.2011.08.005

actually varies from trial to trial, estimates for t_0 will tend to be smaller than the real mean value of t_0.

The storage capacity of VSTM is also an important characteristic of the visual system. Traditionally, VSTM storage capacity in humans has been measured by asking the subject to report as many items as possible from a set of unrelated objects that are presented briefly (whole report; Cattell, 1885). Using letters as stimuli, Sperling (1960, 1963, 1967) found that the capacity was limited to about four or five independent items (see also Bundesen, 1990; Bundesen, Habekost, & Kyllingsbæk, 2005; Cowan, 2001; Shibuya & Bundesen, 1988). Parameter K is a measure of the storage capacity of visual short-term memory (VSTM): the maximum number of items that can be retained at the same time in VSTM. Normally K is assumed to vary only a little across trials. Non-integral values of K are interpreted as probability mixtures between two neighboring integers. For example, a value of 3.62 for K represents a mixture of values of 3 and 4 such that, on any trial, K equals 4 with a probability of 0.62. The variation of K is then given by the variation of a Bernoulli random variable and thus equal to $p(1-p)$ where p is the non-integer remainder of K (e.g. p equals 0.62 in the example where K equals 3.62). The variance of K is therefore minimal and equal to zero when K is an integer, and maximal at 0.25 when K has a non-integer remainder of 0.5.

However, K may possibly vary more strongly than this, and again, the effect of such variation should be highly systematic. In the traditional model, a participant with $K = 3.62$ cannot report more than 4 items correct from a stimulus display without guessing. If a participant obtains a score of 7 items correct just once, the traditional model implies that $K > 6$, regardless of the participant's performance on any of the remaining trials. Thus, the mean value of parameter K will be overestimated. Therefore, when K is assumed to be nearly constant, but actually varies substantially from trial to trial, estimates for K will tend to be greater than the real mean value of K.

Introducing trial-by-trial variation in a certain parameter is a way of generalizing a parametric model. The original model is a special case of the generalized model, the case in which the variance of the parameter in question is limited. If the parameter varies substantially from trial to trial, the generalization is likely to reduce the bias of the estimator for the parameter, as illustrated above. Generalization by introducing parameter variability can sometimes be referred to as construction of a 'hierarchical Bayesian model' (see e.g. Lee, 2011) where parameters of a 'parent distribution' are estimated, or as 'mixed models' in the classical literature (e.g. McCulloch & Searle, 2001). Recently, Rouder et al. (2008), Rouder, Morey, and Morey (2011) and Morey (2011) gave a detailed account, in context of hierarchical Bayesian models, of how estimates of VSTM capacity measured by Cowan's (2001) or Pashler's (1988) formula of change detection may be biased if trial-by-trial lapses in performance are not modeled.

We present computational formulas for fitting TVA to whole and partial report data by maximum likelihood procedures, both with and without assuming substantial trial-by-trial variation in parameters K and t_0. We give numerical examples by the analysis of whole and partial report data from a comprehensive empirical study ($N = 347$), and elaborate on the Bayesian model selection theory for quantifying the advantage of trial-by-trial generalization. Our results demonstrate the feasibility of obtaining a 'fine-grained view' of trial-by-trial variability which Rouder et al. (2008) suggested as a possible alternative to the all-or-nothing modeling of lapses in change detection. We start out by giving an introduction to TVA. For a more comprehensive introduction to TVA, see Bundesen and Habekost (2008).

2. A Theory of Visual Attention (TVA)

TVA enables probabilistic modeling of a subject's performance in tasks involving the categorization of elements from brief visual displays. Categorizations are defined as having the following form: "object x has a certain feature i", e.g., 'x is red' or 'x is a car'. The rate of processing, $v_x(i)$, of the categorization that an element x has a certain feature i (or, equivalently, belongs to a certain category i) is given by the *rate equation*

$$v_x(i) = \eta(x,i)\beta_i \frac{w_x}{\sum_{z\in S} w_z} \tag{1}$$

where $\eta(x,i)$ is the strength of the sensory evidence that x belongs to category i, β_i is the perceptual decision bias associated with category i, and the third term is the relative attentional weight of object x, w_x, divided by the sum of attentional weights across the set of all objects in the visual field, S. The attentional weights in the rate equation (1) are derived from pertinence values. Every category j for which membership of j can be used as a criterion for visual selection has a certain pertinence, π_j. The pertinence of category j is a measure of the importance of attending to objects that belong to category j. The attentional weight of object x is given by the *weight equation*

$$w_x = \sum_{j\in G} \eta(x,j)\pi_j \tag{2}$$

where G is the set of visual categories that can be ascribed pertinence, $\eta(x,j)$ is the strength of sensory evidence that object x belongs to category j, and π_j is the pertinence of category j. Thus, the pertinence of a given category j enters the sum with a weight equal to the strength of the sensory evidence that the object belongs to that category.

The subject is assumed to perceive a particular categorization if and only if that categorization is encoded into VSTM. The VSTM is assumed to be of limited storage capacity in the sense that it can only hold categorizations of K elements at any given time. Note that there is room for several categorizations of the same object in VSTM, the limit is in the number of objects not in the number of features (e.g. Bundesen, 1990; Luck & Vogel, 1997). The duration of the display must exceed a temporal threshold, t_0, in order for any categorization to take place. The amount by which the display duration must exceed the temporal threshold in order for a particular element to reach VSTM is assumed to be stochastic, and is traditionally modeled as coming from an exponential distribution.

Let p_E denote the probability that element x is encoded into VSTM. This probability is zero when $t \le t_0$, (the stimulus duration does not exceed t_0) and we need only derive it given $t > t_0$. On trials where the number of display elements does not exceed K, the storage capacity of VSTM is not a limiting factor. In that case we can derive p_E by simply considering the probability that the exponentially distributed encoding time is smaller than $t - t_0$

$$p_E | K \ge n(S), t > t_0 = 1 - \exp(-v_x[t - t_0]) \tag{3}$$

where $n(S)$ is the number of elements of the set S of displayed elements, and v_x is the rate with which the particular categorization takes place (or equivalently, v_x is the reciprocal of the average time it takes to encode the categorization). As an example, consider using single-element displays, where each trial shows an element at a location enumerated by x. Then a map of the subject's single-item processing rate may be represented by the set of rate parameters $\{v_x\}$ representing the different objects in the display. The rate parameters $\{v_x\}$ are fitted to the subject's responses across trials (hit versus miss trials) via (3).

In multi-element displays where the number of elements exceeds the VSTM storage capacity, there is a probability that the VSTM is filled up before a categorization of x occurs. The probability of encoding item x into VSTM is then derived by taking into account the possible combinations of elements reaching the VSTM. The notion of a 'power set', which is the set of all possible combinations of elements of a set (see definition and algorithm in Appendix A), can be used to express such combinatorics in a very elegant way:

$$\mathrm{p_E}|K < \mathrm{n}(S), t > t_0 = v_x \sum_{j=0}^{K-1} \sum_{J \in \mathcal{P}_j(\tilde{S})} \sum_{L \in \mathcal{P}(J)} (-1)^{|L|} \times \frac{1 - \exp(-[t - t_0]\nu)}{\nu} \tag{4}$$

where $\tilde{S} = S \backslash x$, $\mathcal{P}(J)$ is the power set of J, $\mathcal{P}_j(\tilde{S})$ is the subset of the power set of $\tilde{S}$ in which j elements of $\tilde{S}$ occur in each combination, and

$$\nu = \sum_{m \in S} v_m - \sum_{l \in J} v_l + \sum_{k \in L} v_k \tag{5}$$

see the derivation in Appendix B where multiple integrals over products are made analytically tractable by using the notion of power sets.

An essential assumption of TVA regarding the processing of multi-element displays is that only limited resources are available for processing and thus the processing rate v_x of element x depends on the other elements in the display. That is, v_x becomes a fraction of the total processing capacity such that the processing capacity is distributed to all elements of the display according to their relative attentional weights. Formally,

$$v_x = C \frac{w_x}{\sum_{z \in S} w_z} \tag{6}$$

where w_z is the attentional weight of element z, and C is the fixed limited processing capacity measured in elements per second or Hz. A typical maximum-likelihood fitting scenario involves inserting (6) into a likelihood function involving expressions like (3) and (4), then estimating $(K, C, \{w_z\}, t_0)$ from the subject's responses given a set of stimuli (cf. Kyllingsbæk, 2006). This type of analysis has been used in numerous patient studies based on TVA (see Bublak et al., 2005; Bublak, Redel, & Finke, 2006; Bublak et al., 2009; Duncan et al., 1999, 2003; Finke, Bublak, Dose, Müller, & Schneider, 2006; Finke et al., 2005, 2010, 2007; Gerlach, Marstrand, Habekost, & Gade, 2005; Habekost & Bundesen, 2003; Habekost & Rostrup, 2006, 2007; Habekost & Starrfelt, 2006, 2009; Peers et al., 2005; Redel et al., 2010; Starrfelt, Habekost, & Gerlach, 2010; Starrfelt, Habekost, & Leff, 2009).

In Sections 3 and 4 we generalize the K and t_0 parameters of TVA by introducing trial-by-trial parameter variability. The computational formulas for the first- and second order derivatives of $\mathrm{p_E}$ with respect to traditional and generalized TVA parameters are given in Appendix C. Second order gradient based maximum-likelihood fitting to whole and partial report data is completed via Appendices D–F. We use these equations in gradient descent optimization algorithms (Nielsen, 2006) which enable efficient estimation of the parameters when the model is fitted to experimental data. The equations have been implemented in a MATLAB® software package which is available through the website at location http://zappa.psy.ku.dk/libtva. The numerical examples that we provide in the following are reproducible via this software package.

3. Generalizing the VSTM storage capacity parameter of TVA

As previously mentioned, traditionally in TVA the VSTM storage capacity estimated for a particular subject with a particular type of stimulus material has been summarized by a single number K. The number has been interpreted as a probability mixture such that, for instance a value of 3.62 stands for a mixture of 3 and 4 with probabilities of 1−0.62 and 0.62, respectively (see, e.g., Shibuya & Bundesen, 1988). Following Hebb (1949), most researchers have understood the distinction between short- and long-term memory in neural terms: Traces in short-term memory are patterns of neural activation that may persist for a number of seconds due to reverberation (positive feedback loops). Traces in long-term memory are structural changes, such as long-lasting changes in synaptic efficiency caused by the reverberating patterns of neural activation. Given the Hebbian conception of short-term memory, the storage capacity of VSTM may be expected to vary considerably over time (cf. Bundesen et al., 2005; Usher & Cohen, 1999).

To generalize the fixed VSTM storage capacity parameter K we define a normalized histogram $\mathbf{m}$ such that the j'th element m_j represents the probability that K equals j on a given trial. The normalized histogram $\mathbf{m}$ thus represents the trial-by-trial probability mass function of K, and the model generalization is made by letting the histogram $\mathbf{m}$ replace the parameter K via the law of total probability

$$\mathrm{p_E}|\mathbf{m} = \sum_{K=1}^{\infty} [\mathrm{p_E}|K] \times m_K \tag{7}$$

where the sum runs from $K = 1$ because $\mathrm{p_E}|K$ equals zero when K is zero (m_0 effectively represents the probability of a lapse). Inserting the two cases (3) and (4) of $\mathrm{p_E}$ into (7) we get a finite sum index by K

$$\mathrm{p_E}|\mathbf{m} = \left[\sum_{K=1}^{\mathrm{n}(S)-1} v_x \Sigma \mathrm{p}(K|\mathbf{m}) \right] + [1 - \exp(-v_x \tau)] \mathrm{p}(K \geq \mathrm{n}(S)|\mathbf{m}) \tag{8}$$

where $v_x \Sigma$ equals the right hand side of (4) which can be computed efficiently as described in Appendix G. Thus (8) now enables us to compute the probability of a given item x entering VSTM given the distribution $\mathbf{m}$ of VSTM capacity. We propose to estimate $\mathbf{m}$ in a nonparametric manner using 5 degrees of freedom representing the probability mass function given $K \in [1, 6]$.

Although this generalized model is not parameterized by K, the subject can be characterized with a mean K-value by the standard formula for computing the mean of a discrete probability:

$$\mathrm{E}[K|\mathbf{m}] = \sum_{j=0}^{\infty} m_j \times j \tag{9}$$

which computes the expected K on any trial given $\mathbf{m}$. However, it may also be relevant to determine the maximum capacity for the subject as well as some characterization of the trial-by-trial deviation from this maximum. Let $(K_{\min}, K_{\max})$ denote the bounds of the substantial mass of $\mathbf{m}$. Then a shifted Binomial mass function over the interval $[K_{\min}, K_{\max}]$ with the probability parameter $\mathrm{p_{inc}} = (\mathrm{E}[K] - K_{\min})/(K_{\max} - K_{\min})$ offers a simple interpretation: the VSTM of the subject has a maximum capacity of $K_{\max}$, and there is an independent probability $1 - \mathrm{p_{inc}}$ that each of $K_{\max} - K_{\min} + 1$ VSTM slots is occupied by some task-irrelevant item. Since we have assumed that the subjects perform the task without guessing, the maximum likelihood estimate of $K_{\max}$ will naturally be no less than the maximum number of correct items reported by the subject on any trial.

3.1. Example: comparing the generalized K-model to the traditional model

The two histograms shown in Fig. J.1 represent distributions of estimated K values based on a sample of 347 subjects (see

Appendix H for demographic details) performing a mixture of whole- and partial report trials (324 trials total). In each trial a number of red and blue letters were shown briefly on a CRT monitor running at 100 Hz. The letter displays were followed by a display with six pattern masks covering the six possible stimulus locations and exposed for 500 ms. The task of the subjects was to report the red letters while ignoring the blue letters. Three types of displays were used: (1) whole report displays of six red letters with varying exposure duration ranging between 10 and 200 ms, (2) whole report displays with 2 red letters and a fixed exposure duration of 80 ms, and (3) partial report displays with 2 red letters and four blue letters also with a fixed exposure duration of 80 ms (for further details of the paradigm, see Vangkilde, Bundesen, & Coull, 2009, in press).

The left histogram in Fig. J.1 was produced using a traditional K-model, then forming the histogram of estimated K values, whereas the generalized model was used to form the right histogram of expected values $E[K]$. The traditional histogram reveals a peculiar pattern: subject frequency peaks right above integer values of K. This multi-modal nature of the traditional histogram to the left, with frequency peaks next to integer values of K, affirms a systematic error: the traditional estimate of a subject's K must be so large as to include the subject's best score (formally, a subject with a best score of Y correct targets implies a K estimate greater than Y minus one). Because the histogram to the right is closer to Normal, it underlines the fact that generalized modeling can help to eliminate such systematic errors, simply due to the allowance of trial-by-trial parameter variability. A likelihood ratio test on the sample confirms our rejection of the traditional model in favor of the generalized K-model ($\chi^2(1388) = 4380.85 \sim z = (4380.85 - 1388)/\sqrt{2 \times 1388} = 56.80, p \ll 0.0001$). The systematic error of the traditional model is an upward bias. It yields a higher estimate (mean 3.45, SEM = 0.06) than the generalized model (which has mean = 2.97, SEM = 0.05) which puts the bias of traditional estimation around the order of 0.5, assuming that the generalized estimates are unbiased. We cannot be sure that the generalized model is unbiased, but on the other hand we do not expect it to be biased in the opposite direction because it contains the traditional model as a special case, and hence the order of the traditional bias of 0.5 is likely to be a conservative estimate. In other words, it seems that traditional modeling of this type of data leads to estimates of K which on average are at least 0.5 higher than the true value.

Fig. J.2 shows the estimated trial-by-trial VSTM storage capacity histograms (each a generalized model vector $\mathbf{m}$) for the first twelve subjects. It is very intriguing to observe that the histograms are unimodal and sparse.

4. Generalizing the threshold for visual perception parameter of TVA

Traditionally in TVA, the threshold for visual perception for a particular subject with a particular type of stimulus material has been summarized by a single number t_0. If we instead assume that t_0 is distributed across trials according to a pdf with mean parameter $\mu_0 = \mathrm{E}[t_0]$ and deviation parameter $\sigma_0 = \sqrt{\mathrm{E}[(t_0 - \mu_0)^2]}$, then the generalization is made by substituting

$$\mathrm{p}_\mathrm{E}|\mu_0, \sigma_0 = \int_{-\infty}^{t} [\mathrm{p}_\mathrm{E}|t_0] \times \mathrm{p}(t_0|\mu_0, \sigma_0)\mathrm{d}t_0 \tag{10}$$

where the integration stops at t since if $t_0 > t$ then $\mathrm{p}_\mathrm{E}|t_0$ is zero. The integral (10) is a convolution integral,[2] and assuming that the prior $\mathrm{p}(t_0|\mu_0, \sigma_0)$ is Gaussian, the exponential processing assumption described above consequently becomes an ex-Gaussian processing assumption instead (named ex-Gaussian because it is an exponential convolved with a Gaussian; Luce, 1986, pp. 34–36). We assume the Gaussian prior because it is the most objective assumption when nothing but mean and variance are specified (cf. the MaxEnt principle, see Jaynes, 1957; Sivia, 1996), and it allows for the analytical solution of the convolution integral. The resulting p_E, now conditioned on μ_0 and σ_0 instead of t_0, is given by

$$\forall t : \mathrm{p}_\mathrm{E}|\mu_0, \sigma_0 = \begin{cases} \mathrm{erfc}(d)/2 - h \times \mathrm{erfc}(d + v_x\sigma_0/\sqrt{2}), & \dim(S) \le K \\ v_x\Sigma|\mu_0, \sigma_0, & \dim(S) > K \end{cases} \tag{11}$$

where $d \equiv (\mu_0 - t)/\sqrt{2\sigma_0^2}$, and $v_x\Sigma|\mu_0, \sigma_0$ is given by

$$v_x\Sigma|\mu_0, \sigma_0 = v_x \sum_{j=0}^{K-1} \sum_{J \in \mathcal{P}_j(\tilde{S})} \sum_{L \in \mathcal{P}(J)} (-1)^{|L|} \times \frac{\mathrm{erfc}(d)/2 - h \times \mathrm{erfc}(d + \nu\sigma_0/\sqrt{2})}{\nu} \tag{12}$$

$\mathrm{erfc}(\cdot)$ is the 'complementary error function' defined by the integral $\frac{2}{\sqrt{\pi}} \int_\cdot^\infty \exp(-t^2)\mathrm{d}t$ which has a well-known Taylor series expansion, and $h = \frac{1}{2}\exp(\nu[\mu_0 - t] + \nu^2\sigma_0^2/2)$. Thus again, the generalized p_E may be computed efficiently. The derivations are trivial using the rule derived in Appendix I; note the simple relationships between (12) and (4), and between (11) and (3).

The Gaussian prior is convenient in the way that it leads to the ex-Gaussian model. However, the theoretically unbounded support of the Gaussian speaks against the ex-Gaussian because it predicts a non-zero probability of encoding even in the exposure duration limit toward zero. Such misprediction may be negligible for small values of σ_0 (relative to μ_0), but a bounded prior is in principle desired. However, such extremum statistic may be very sensitive to guessing and we therefore defer investigations of the exact shape of the trial-by-trial distribution of t_0 to future context in which explicit guessing strategy modeling is taken into account.

4.1. Example: comparing the ex-Gaussian with the traditional model

As in the previous example, we consider the data from the 347 subjects performing a mixture of whole- and partial report trials. The data are analyzed with both the traditional and the generalized t_0-model and the two corresponding histograms are shown in Fig. J.3. For both models the VSTM capacity was modeled with the traditional K-model. The histograms reveal a significant difference between the traditional and the generalized t_0-estimation: the traditional estimation yields a bimodal histogram, while the generalized model yields a more normal looking histogram. Fig. J.4 shows a scatter plot of the generalized estimates versus traditional estimates. It is clear that the generalized estimates deviate for subjects that have a traditional estimate just below 10, 20, or 50 ms. These numbers happen to be the three lowest exposure durations in the stimulus material. The phenomenon can be explained as a systematic error in the traditional model: if the subject has a mean t_0 above a certain stimulus duration, but sometimes perceives a stimulus at that duration due to trial-by-trial variability, then the traditional estimate will be driven below the stimulus duration. From Figs. J.3 and J.4 it is clear that the generalized t_0 model is better than the traditional TVA model for many subjects of our sample. The traditional TVA model yields a lower temporal threshold (mean = 14.3 ms, SEM = 0.7) compared to the ex-Gaussian TVA model (mean = 16.4 ms, SEM = 0.8), which puts the bias of the traditional estimator on the order of 2 ms on

[2] $[\mathrm{p}_\mathrm{E}|t_0]$ is a function of $t - t_0$.

average, depending on the particular exposure durations chosen in the paradigm in the present paper. A likelihood ratio test was in favor of the ex-Gaussian model ($\chi^2(347) = 463.59 \sim z = (463.59 - 347)/\sqrt{2 \times 347} = 4.43, p < 0.0001$). The scatter plot in Fig. J.4 indicates that the small difference in the average t_0 estimates between the traditional and the generalized t_0-model is driven only by a subset of the subjects. It is likely that a finer experimental sampling of the t axis will increase the significance as the generalized estimates versus traditional estimates are expected to diverge.

5. Selecting level of generalization

In our previous examples we used likelihood ratio tests to quantify the evidence of each individual trial-by-trial generalization as an alternative to the traditional TVA model. In this section, we consider the more general question of which parameter combination is the best. In the case of TVA we may ask: Should only t_0 be generalized? Should only K be generalized? Or, should both K and t_0 be generalized simultaneously? A likelihood ratio test will not suffice since it requires the null hypothesis to be a restricted version of the alternative (see for example McCulloch & Searle, 2001), which is not the case under this more general question. Instead we address the question quantitatively via Bayesian model selection. The Ockham's razor principle of Bayesian model selection trades off closeness of fit in favor of parsimony, and one of two competing models need not be a restricted version of the other.

To illustrate, let $\mathrm{p}(\mathrm{data}|\theta, \mathcal{M})$ denote the likelihood function of a parametric model that associates data with the parameter set θ. The symbol $\mathcal{M}$ represents all modeling assumptions except the actual values of the parameter vector θ. The Bayes factor for a model against another is defined as the ratio of marginal likelihoods where the marginal likelihood for model $\mathcal{M}$ is given by the integral

$$\mathrm{p}(\mathrm{data}|\mathcal{M}) = \int \mathrm{p}(\mathrm{data}|\theta, \mathcal{M})\mathrm{p}_\mathrm{A}(\theta|\mathcal{M})\mathrm{d}\theta \tag{13}$$

which can be approximated by the Laplace approximation (Kass & Raftery, 1995; Raftery, 1996). The Laplace approximation assumes that the mode of the likelihood function is sharply peaked compared with p_A (i.e. our prior knowledge is diffuse compared to the information gathered by the likelihood). The next assumption required by the Laplace approximation is that the mode of the likelihood function is Gaussian in shape, the analytical result is then obtained as

$$\mathrm{p}(\mathrm{data}|\mathcal{M}) \approx l^* \mathrm{p}_\mathrm{A}(\theta^*|\mathcal{M})(2\pi)^{\dim(\theta)/2} \det(\mathbf{H}^*)^{-1/2} \tag{14}$$

where $\dim(\theta)$ is the number of parameters, l^* is the value of the likelihood at the mode, and $\mathbf{H}^*$ is the Hessian matrix of second order derivatives of the negative log-likelihood function $-\log \mathrm{p}(\mathrm{data}|\theta, \mathcal{M})$ with respect to θ evaluated at the mode (see also Bishop, 1996; Raftery, 1996; Sivia, 1996). In fact, assuming that the mode of the likelihood function is shaped like a Gaussian with covariance matrix $\mathbf{C}$, then $\mathbf{H}^*$ will be symmetric and positive definite with the correspondence $\mathbf{H}^* = \mathbf{C}^{-1}$ (McCulloch & Searle, 2001; Sivia, 1996). That is, the $\det(\mathbf{H}^*)$ term in (14) may be used as an estimate of the precision of the likelihood mode.

Without specifying the exact value of $\mathrm{p}_\mathrm{A}(\cdot|\mathcal{M})$ we note that it depends on the number of parameters (the dimension of θ) and the ranges of individual parameters in θ; see for example Sivia (1996). Also note that $\mathrm{p}_\mathrm{A}(\cdot|\mathcal{M})$ does not scale with the number of trials and becomes increasingly negligible with an increasing number of trials. To conclude our illustration we consider the scenario of comparing two models (denoted by subscripts 0 and 1) with an equal number of parameters and $\mathrm{p}_\mathrm{A}(\theta_0^*|\mathcal{M}_0) = \mathrm{p}_\mathrm{A}(\theta_1^*|\mathcal{M}_1)$. The Bayes factor for $\mathcal{M}_1$ against $\mathcal{M}_0$ reduces to

$$\frac{\mathrm{p}(\mathrm{data}|\mathcal{M}_1)}{\mathrm{p}(\mathrm{data}|\mathcal{M}_0)} = \frac{l_1^*}{l_0^*} \times \left(\frac{\det(\mathbf{H}_0^*)}{\det(\mathbf{H}_1^*)}\right)^{1/2} \tag{15}$$

from which we see that the preference tends toward the model of greater likelihood (l_1^* vs. l_0^*), but due to the precision ratio ($\det(\mathbf{H}_0^*)$ vs. $\det(\mathbf{H}_1^*)$), this will only suffice if the precision of the likelihood mode is not too high (relative to the other model). If the likelihood precision is too high, the value of the likelihood is too sensitive to perturbation of parameter values in which case the model is expected to perform poorly on another data sample from the same population: the estimator variance is too high. Generalizing a parametric model by introducing trial-by-trial variability is likely to increase the estimator variance (increased mode precision) because, other things equal, an increase in the number of model parameters makes it possible to fit the observed data by a larger number of combinations of the model parameters. Thus, in order to lower the bias of an estimator working on a given sample, one must generally accept an increased variance. This dilemma is known as the bias–variance dilemma (Geman et al., 1992), and the Bayesian model selection strategy offers a principled balance between parsimony and closeness of fit to the particular data instance. Note that the determinant scales exponentially with $\dim(\theta)$, and the $(2\pi)^{\dim(\theta)/2}$ term in (14) represents robustness of the Bayesian model selection when comparing models with different numbers of parameters.

Further approximation can be made by considering the limit as the number of trials tends toward infinity. This leads to the Bayes Information Criterion (BIC)

$$l^* \det(\mathbf{H}^*)^{-1/2} \approx l^* N^{-\dim(\theta)/2} = \mathrm{BIC} \tag{16}$$

which is identical to the Schwarz (1978) criterion (see also Kass & Raftery, 1995; Raftery, 1996). The BIC is a convenient choice when the Hessian matrix is not available, but other alternatives exist, such as numerical integration via Markov Chain Monte Carlo (MCMC) procedures (see e.g. Ruanaidh & Fitzgerald, 1996), which may be more appropriate when it is hard to justify the assumption of an infinite number of trials.

5.1. Example: generalizing multiple parameters

In this example, we quantify which generalization of TVA parameters t_0 and K is the best by evaluating the integral in (13) on each of the 347 subjects performing a mixture of whole- and partial report trials. The Hessian matrix, $\mathbf{H}^*$, given by the second order derivatives of the likelihood function with respect to t_0, μ_0, C, and $\{w_x\}$ in Appendices C–E enable the use of the Laplace approximation for integrating the continuous parameters of the traditional TVA model and also for μ_0. However, Hessian entries involving K, $\mathbf{m}$ or σ_0 are not available and we use an MCMC procedure to integrate those dimensions numerically (see Appendix J for details about the MCMC procedure).

Table 1 shows the number of subjects for which each combination of K-model and t_0-model is best. About half of the subjects show evidence in favor of generalizing both K and t_0. About 40% of subjects benefit from generalizing only one or the other and

Table 1
The number of subjects for which the combination of K-model (column) and t_0-model (row) is the best.

	Traditional K	Generalized K	Total
Traditional t_0	29	62	91
Generalized t_0	88	168	256
Total	117	230	347

the traditional version of TVA was only favored for very few subjects (8%). Therefore, generalizing both TVA parameters K and t_0 by including the possibility of trial-by-trial variation of the parameters in the model, greatly improved the feasibility of the model in terms of both reduction of the bias of the parameter estimates and Bayesian selection criteria.

Fig. J.5 shows observed mean scores and model predictions for a subject for which the evidence is in favor of generalizing both t_0 and K. As can be seen, the generalization of K reduces the overshoot at the longest exposure, while the t_0 generalization adjusts for the observed S-shape around the shortest exposures.

6. Discussion and conclusions

We used the Theory of Visual Attention (TVA; Bundesen, 1990) as an example of a parametric model in which the major parameters are assumed to be nearly constant across trials. We tested the assumption for two parameters in TVA, the threshold for conscious perception, t_0, and VSTM capacity, K. The data from 347 subjects enabled us to investigate the distribution of parameter estimates in detail for patterns of biases. We found two distinct patterns for the two parameters: For parameter K, we found frequency peaks next to integer values of K, indicating systematic upward bias where K was erroneously pulled up to the highest number of letters reported by the subject, minus one. Furthermore, we found that parameter t_0 was underestimated by a downward bias toward values just below one of the shortest exposure durations used in the particular paradigm.

As expected, introducing random trial-by-trial variation of the two parameters significantly reduced both types of biases. For parameter K we introduced trial-by-trial variance by allowing K to vary freely, parameterized by a normalized histogram. The trial-by-trial variance for parameter t_0 was introduced by exchanging the single constant parameter by a normal distribution, given by its mean value and variance. Our results strongly support the principle of generalization by introduction of trial-by-trial parameter variability resulting in significantly reduced estimation bias (Morey, 2011). Specifically, in whole and partial report trials, the traditionally overestimated VSTM capacity was significantly reduced. The Gaussian trial-by-trial generalization that we applied to t_0 was modest on average, but it made a significant difference for many of the tested subjects. We do, however, expect a finer experimental resolution of the stimulus duration to yield unbiased estimation in subjects in general.

When comparing generalization of the K and t_0 parameters individually, we used likelihood ratio tests. However, this is not possible when comparing the effects mutually because one of these parameters is not a restricted version of the other. We instead used the Bayesian model selection theory to compare the significance of trial-by-trial parameter generalization in general. In Bayesian model selection a number of different parameters, and combinations hereof, can be generalized and the mutual significance of doing so can be compared. The analysis favored the generalization of both parameters K and t_0 for nearly half of the subjects. In contrast, the traditional model where the perceptual threshold, t_0, is constant and VSTM capacity, K, only varies between two neighboring integer values was favored for less than 10% of the subjects.

As advocated by Morey (2011) and Rouder et al. (2008), ignoring possible variance on a trial-by-trial basis such as lapses in change detection may significantly bias parameter estimates of VSTM capacity, K. In contrast to estimation of K from change-detection experiments using either Pashler's (1988) or Cowan's (2001) formula, the use of whole- and partial report in combination with Bundesen's (1990) TVA enables the simultaneous estimation of several important and interacting visual parameters such as threshold for conscious perception, t_0, processing capacity, C, and VSTM capacity, K. The introduction of trial-by-trial variance in central parameters of the model (here t_0 and K) greatly improves the validity of the parameter estimates as measured by Bayesian model selection. That is not to say that alternative improvements may not lead to even better models. For example, we have dismissed the role of guessing, and we have also focused on only two parameters of the TVA model namely K and t_0. Therefore an untested alternative hypothesis is that instead of generalizing K and t_0, the data can be explained better by extending the model with a guessing strategy or by generalizing other TVA parameters such as C or w. However, in whole- and partial report as we have analyzed, the incremental hit-rate of a subject due to guessing can be approximated roughly as the product of three probabilities: p_1, the probability that the subject is unable to report a target via nonguessing; p_2, the probability that the subject chooses to guess; and p_3, the probability that the subject guesses correctly when he or she chooses to guess. This principle may serve as a route to modeling guessing, but we note that the functional form of $p_1p_2p_3$ may not be trivial (see for example Kyllingsbæk et al., in press). In our case we have told the subjects to refrain from guessing, and hence probability p_2 is small. Furthermore, the alphabet is relatively large, and hence p_3 is small. We therefore consider the combined probability negligible. While TVA parameters C and w could potentially benefit from generalization, we have focused on TVA parameters K and t_0 based on the histograms from 347 subjects that revealed t_0-oddities directly related to exposure duration and K-oddities related to integer values.

Acknowledgments

This research was supported by grants from the University of Copenhagen (M.D., S.K., & C.B.), the Sapere Aude Program of the Danish Council for Independent Research (M.D. & S.K.), the Danish Council for Strategic Research (S.K.), the Danish Research Council for the Humanities (S.K.), and the Research Council of Norway (T.E).

Appendix A. Power sets

Given a set S, the power set of S, denoted $\mathcal{P}(S)$, is the set of all possible subsets of S. For example,

$$S = \{A, B, C\} \Leftrightarrow$$
$$\mathcal{P}(S) = \{\emptyset, \{A\}, \{B\}, \{C\}, \{A, B\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}.$$

Let $\mathcal{P}_j(S)$ denote the power set with elemental cardinality-j constraint, that is $P_j(S)$ is the largest subset of $P(S)$ in which all elements (each being a subset of S) have cardinality equal to j. Clearly, $\mathcal{P}_j(S)$ can be obtained by forming $\mathcal{P}(S)$ and then filtering it, but since $\mathcal{P}(S)$ can be enormous compared to $\mathcal{P}_j(S)$, such filtering can demand a significant overhead. The online appendix contains an algorithm for running through $\mathcal{P}_j(S)$ efficiently.

Appendix B. Derivation of p_E, the probability that item x is encoded in VSTM

Let p_E denote the probability that a particular element is encoded into VSTM. In the following, x denotes the element in question. Let τ denote the *effective* stimulus duration $t - t_0$. Note that $S = S_T \cup S_D$, that is the set of all stimuli presented is the union of target stimuli and distractor stimuli. The following derivation is adapted from Kyllingsbæk and Habekost (2001).

$$p_E|K = \sum_{j=0}^{K-1} \sum_{J \in \mathcal{P}_j(S\setminus x)} \int_0^\tau v_x \exp(-v_x t) \prod_{k \in J} \times [1 - \exp(-v_k t)] \prod_{l \in S\setminus x\setminus J} \exp(-v_l t) \mathrm{d}t. \quad \text{(B.1)}$$

The term with products of the exponential functions can be collapsed into a single exponential function of the summed v values. Further, the term including $\prod 1-\exp(-v_x t)$ is transformed into a sum using the method of power sets $\prod_{i\in A}[1 - x_i] = \sum_{J\in\mathcal{P}(A)}(-1)^{\mathrm{n}(J)} \prod_{i\in J} x_i$

$$\mathrm{p_E}|K = v_x \sum_{j=0}^{K-1} \sum_{J\in\mathcal{P}_j(S\setminus x)} \int_0^\tau \exp\left(-\sum_{l\in S-J} v_l t\right) \times \sum_{L\in\mathcal{P}(J)} (-1)^{\mathrm{n}(L)} \exp\left(-\sum_{k\in L} v_k t\right) \mathrm{d}t. \tag{B.2}$$

The single term with exponential function is moved inside the $\sum$ operator.

$$\mathrm{p_E}|K = v_x \sum_{j=0}^{K-1} \sum_{J\in\mathcal{P}_j(S\setminus x)} \sum_{L\in\mathcal{P}(J)} (-1)^{\mathrm{n}(L)} \times \int_0^\tau \exp\left(-\sum_{l\in S\setminus J} v_l t\right) \exp\left(-\sum_{k\in L} v_k t\right) \mathrm{d}t. \tag{B.3}$$

Again the resulting product of the exponential functions is transformed into a single exponential function. Finally the expression is integrated yielding (4); see Appendix A for an algorithm for running through power sets $\mathcal{P}_j$.

Appendix C. Derivatives involving $\mathrm{p_E}$

Recall that $\mathrm{p_E}$ is the probability that a particular element is encoded into VSTM. In the following, x denotes the element in question, and z is used to denote arbitrary elements. Let τ denote the *effective* stimulus duration $t - t_0$.

C.1. General form of derivatives of $\mathrm{p_E}$ with respect to $\mathbf{v}$ when $\mathrm{n}(S) > K$

When $\mathrm{n}(S) > K$, differentiating $\mathrm{p_E}$ once with respect to the processing rates $\mathbf{v}$ follows the general form

$$\frac{\partial \mathrm{p_E}}{\partial v_z} = \delta(z-x)\Sigma + v_x \Gamma_z \quad \text{where } \Gamma_z \equiv \frac{\partial \Sigma}{\partial v_z} \tag{C.1}$$

where $\delta(\cdot)$ is the Dirac delta function which equals one when the argument is zero and equals zero when the argument is non-zero. Similarly for the second order derivative

$$\frac{\partial^2 \mathrm{p_E}}{\partial v_z \partial v_{z'}} = \delta(z-x)\Gamma_{z'} + \delta(z'-x)\Gamma_z + v_x \Xi_{z,z'} \tag{C.2}$$

where Γ and Ξ will be given in sections below.

C.2. Derivatives of $\mathrm{p_E}$ when $\mathrm{n}(S) \le K$ and assuming exponential processing

Assuming exponential processing, and that the number of stimuli does not exceed K, the first order derivatives of $\mathrm{p_E}$ with respect to the processing rates and temporal threshold are given by

$$\frac{\partial \mathrm{p_E}}{\partial v_z} = \delta(x-z)\tau \exp(-v_x\tau) \tag{C.3}$$

$$\frac{\partial \mathrm{p_E}}{\partial t_0} = -v_x \exp(-v_x\tau). \tag{C.4}$$

Similarly, the second order derivatives are given by

$$\frac{\partial^2 \mathrm{p_E}}{\partial v_z \partial v_{z'}} = -\delta(x-z)\delta(x-z')\tau^2 \exp(-v_x\tau) \tag{C.5}$$

$$\frac{\partial^2 \mathrm{p_E}}{\partial t_0 \partial t_0} = -v_x^2 \exp(-v_x\tau) \tag{C.6}$$

$$\frac{\partial^2 \mathrm{p_E}}{\partial t_0 \partial v_z} = \delta(x-z)[v_x\tau - 1]\exp(-v_x\tau). \tag{C.7}$$

C.3. Derivatives of $\mathrm{p_E}$ when $\mathrm{n}(S) > K$ and assuming exponential processing

Assuming exponential processing, and that the number of stimuli does exceed K, the first and second order derivatives of $\mathrm{p_E}$ with respect to the processing rates are given by the general form in Appendix C.1 using the following Γ and Ξ

$$\Gamma_z = \sum_{J,L} \Lambda_{z\notin J\setminus L}(-1)^{|L|}\left[\frac{\tau\exp(-\tau\nu)}{\nu} - \frac{1-\exp(-\tau\nu)}{\nu^2}\right] \tag{C.8}$$

and

$$\Xi_{z,z'} \equiv \frac{\partial \Gamma_z}{\partial v_{z'}} = \sum_{J,L} \Lambda_{z\notin J\setminus L}\Lambda_{z'\notin J\setminus L}(-1)^{|L|}\left[-\frac{\tau^2\exp(-\tau\nu)}{\nu} - \frac{2\tau\exp(-\tau\nu)}{\nu^2} + \frac{2-2\exp(-\tau\nu)}{\nu^3}\right] \tag{C.9}$$

where $\Lambda_{z\notin J\setminus L} = 1$ unless $z \in J\setminus L \iff \Lambda_{z\notin J\setminus L} = 0$. The first order derivative of $\mathrm{p_E}$ with respect to t_0 is given by

$$\frac{\partial \mathrm{p_E}}{\partial t_0} = v_x \frac{\partial \Sigma}{\partial t_0} \quad \text{where } \frac{\partial \Sigma}{\partial t_0} = -\sum_{J,L}(-1)^{|L|}\exp(-\tau\nu). \tag{C.10}$$

The second order derivative of $\mathrm{p_E}$ with respect to t_0 twice is given by

$$\frac{\partial^2 \mathrm{p_E}}{\partial t_0 \partial t_0} = -v_x \sum_{J,L}(-1)^{|L|}\nu\exp(-\tau\nu). \tag{C.11}$$

The second order derivatives of $\mathrm{p_E}$ with respect to $\mathbf{v}$ and t_0 are given by

$$\frac{\partial^2 \mathrm{p_E}}{\partial v_z \partial t_0} = \delta(z-x)\frac{\partial \Sigma}{\partial t_0} + v_x \frac{\partial \Gamma_z}{\partial t_0} \quad \text{where } \frac{\partial \Gamma_z}{\partial t_0} = \sum_{J,L} \Lambda_{z\notin J\setminus L}(-1)^{|L|}\tau\exp(-\tau\nu). \tag{C.12}$$

C.4. Derivatives of $\mathrm{p_E}$ when $\mathrm{n}(S) \le K$ and assuming ex-Gaussian processing

Assuming ex-Gaussian processing, and that the number of stimuli does not exceed K, the first order derivatives of $\mathrm{p_E}$ with respect to the processing rates are given by

$$\frac{\partial \mathrm{p_E}|\mu_0}{\partial v_z} = \delta(z-x)\frac{\partial - h\mathrm{erfc}(-c)}{\partial v_x} = \delta(z-x)\left[-\frac{\partial h}{\partial v_x}\mathrm{erfc}(-c) - h \times \frac{\partial \mathrm{erfc}(-c)}{\partial v_x}\right] \tag{C.13}$$

where

$$\frac{\partial h}{\partial v_x} = h \times [\mu_0 - t + v_x\sigma^2] \quad \text{and}$$

$$\frac{\partial \mathrm{erfc}(-c)}{\partial v_x} = -\frac{2\sigma_0}{\sqrt{2\pi}}\exp(-c^2). \tag{C.14}$$

The second order derivatives with respect to the processing rates are given by

$$\frac{\partial^2 \mathrm{p_E}|\mu_0}{\partial v_z \partial v_{z'}} = \delta(x-z)\delta(z'-x)\Bigg[-\frac{\partial^2 h}{\partial v_x \partial v_x}\mathrm{erfc}(-c) - 2\frac{\partial h}{\partial v_x}\frac{\partial \mathrm{erfc}(-c)}{\partial v_x} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial v_x \partial v_x}\Bigg] \tag{C.15}$$

where

$$\frac{\partial^2 h}{\partial v_x \partial v_x} = \frac{\partial h}{\partial v_x}[\mu_0 - t + v_x\sigma^2] + h\sigma^2 \quad \text{and}$$
$$\frac{\partial^2 \mathrm{erfc}(-c)}{\partial v_x \partial v_x} = -2c\frac{\sigma_0}{\sqrt{2}}\frac{\partial \mathrm{erfc}(-c)}{\partial v_x}. \tag{C.16}$$

The derivative with respect to μ_0 is given by

$$\frac{\partial \mathrm{p_E}|\mu_0}{\partial \mu_0} = \frac{1}{2}\frac{\partial \mathrm{erfc}(d)}{\partial \mu_0} - h\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - \frac{\partial h}{\partial \mu_0}\mathrm{erfc}(-c) \tag{C.17}$$

where

$$\frac{\partial \mathrm{erfc}(d)}{\partial \mu_0} = -\frac{2}{\sqrt{2\pi\sigma_0^2}}\exp(-d^2) \quad \text{and}$$
$$\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} = -\frac{2}{\sqrt{2\pi\sigma_0^2}}\exp(-c^2) \quad \text{and} \quad \frac{\partial h}{\partial \mu_0} = hv_x. \tag{C.18}$$

The second order derivative with respect to μ_0 is given by

$$\frac{\partial^2 \mathrm{p_E}|\mu_0}{\partial \mu_0 \partial \mu_0} = \frac{1}{2}\frac{\partial^2 \mathrm{erfc}(d)}{\partial \mu_0 \partial \mu_0} - 2\frac{\partial h}{\partial \mu_0}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \mu_0 \partial \mu_0} - \frac{\partial^2 h}{\partial \mu_0 \partial \mu_0}\mathrm{erfc}(-c) \tag{C.19}$$

where

$$\frac{\partial^2 \mathrm{erfc}(d)}{\partial \mu_0 \partial \mu_0} = \frac{-2d}{\sqrt{2\sigma_0^2}}\frac{\partial \mathrm{erfc}(d)}{\partial \mu_0} \quad \text{and}$$
$$\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \mu_0 \partial \mu_0} = \frac{-2c}{\sqrt{2\sigma_0^2}}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} \quad \text{and}$$
$$\frac{\partial^2 h}{\partial \mu_0 \partial \mu_0} = v_x\frac{\partial h}{\partial \mu_0}. \tag{C.20}$$

The second order derivatives with respect to processing rates and μ_0 are given by

$$\frac{\partial^2 \mathrm{p_E}|\mu_0}{\partial v_z \partial \mu_0} = \delta(z-x)\Bigg[-\frac{\partial^2 h}{\partial v_x \partial \mu_0}\mathrm{erfc}(-c) - \frac{\partial h}{\partial v_x}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - \frac{\partial h}{\partial \mu_0}\frac{\partial \mathrm{erfc}(-c)}{\partial v_x} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial v_x \partial \mu_0}\Bigg] \tag{C.21}$$

where

$$\frac{\partial^2 h}{\partial v_x \partial \mu_0} = h + \frac{\partial h}{\partial v_x}v_x \quad \text{and}$$
$$\frac{\partial^2 \mathrm{erfc}(-c)}{\partial v_x \partial \mu_0} = \frac{2c\sigma_0}{\sqrt{2}}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} \tag{C.22}$$

C.5. Derivatives of $\mathrm{p_E}$ *when* $\mathrm{n}(S) > K$ *and assuming ex-Gaussian processing*

Assuming ex-Gaussian processing, and that the number of stimuli does exceed K, the first and second order derivatives of $\mathrm{p_E}$ with respect to the processing rates are given by the general form in Appendix C.1 using the following Γ and Ξ

$$\Gamma_z|\mu_0, \sigma_0 = \sum_{j,J,L} \Lambda_{z\notin J\setminus L}(-1)^{|L|}\left[Y/\nu - I/\nu^2\right] \tag{C.23}$$

where I is given in (I.1) and

$$Y = \frac{\partial - h \times \mathrm{erfc}(-c)}{\partial \nu} = -\frac{\partial h}{\partial \nu}\mathrm{erfc}(-c) - h \times \frac{\partial \mathrm{erfc}(-c)}{\partial \nu} \tag{C.24}$$

where

$$\frac{\partial h}{\partial \nu} = h \times [\mu_0 - t + \nu\sigma^2] \quad \text{and}$$
$$\frac{\partial \mathrm{erfc}(-c)}{\partial \nu} = -\frac{2\sigma_0}{\sqrt{2\pi}}\exp(-c^2). \tag{C.25}$$

Further,

$$\Xi_{z,z'}|\mu_0, \sigma_0 = \sum_{j,J,L} \Lambda_{z\notin J\setminus L}\Lambda_{z'\notin J\setminus L}(-1)^{|L|}\Bigg[\frac{1}{\nu}\frac{\partial Y}{\partial \nu} - \frac{2Y}{\nu^2} + \frac{\mathrm{erfc}(d) - 2h \times \mathrm{erfc}(-c)}{\nu^3}\Bigg] \tag{C.26}$$

where

$$\frac{\partial Y}{\partial \nu} = -\frac{\partial^2 h}{\partial \nu \partial \nu}\mathrm{erfc}(-c) - 2\frac{\partial h}{\partial \nu}\frac{\partial \mathrm{erfc}(-c)}{\partial \nu} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \nu \partial \nu} \tag{C.27}$$

where

$$\frac{\partial^2 h}{\partial \nu \partial \nu} = \frac{\partial h}{\partial \nu}[\mu_0 - t + \nu\sigma^2] + h\sigma^2 \quad \text{and}$$
$$\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \nu \partial \nu} = -2c\frac{\sigma_0}{\sqrt{2}}\frac{\partial \mathrm{erf}(-c)}{\partial \nu}. \tag{C.28}$$

The first order derivative with respect to μ_0 is given by

$$\frac{\partial \mathrm{p_E}|\mu_0}{\partial \mu_0} = v_x\sum_{j,J,L}(-1)^{|L|}\frac{W}{\nu} \tag{C.29}$$

where

$$W = \frac{\partial I}{\partial \mu_0} = \frac{1}{2}\frac{\partial \mathrm{erfc}(d)}{\partial \mu_0} - h\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - \frac{\partial h}{\partial \mu_0}\mathrm{erfc}(-c) \tag{C.30}$$

where

$$\frac{\partial \mathrm{erfc}(d)}{\partial \mu_0} = -\frac{2}{\sqrt{2\pi\sigma_0^2}}\exp(-d^2) \quad \text{and}$$
$$\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} = -\frac{2}{\sqrt{2\pi\sigma_0^2}}\exp(-c^2) \quad \text{and}$$
$$\frac{\partial h}{\partial \mu_0} = h\nu. \tag{C.31}$$

The second order derivative with respect to μ_0 twice is given by

$$\frac{\partial^2 \mathrm{p_E}|\mu_0}{\partial \mu_0 \partial \mu_0} = v_x\sum_{j,J,L}(-1)^{|L|}\frac{\partial W}{\partial \mu_0}\Big/\nu \tag{C.32}$$

where

$$\frac{\partial W}{\partial \mu_0} = \frac{1}{2}\frac{\partial^2 \mathrm{erfc}(d)}{\partial \mu_0 \partial \mu_0} - 2\frac{\partial h}{\partial \mu_0}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \mu_0 \partial \mu_0} - \frac{\partial^2 h}{\partial \mu_0 \partial \mu_0}\mathrm{erfc}(-c). \quad \text{(C.33)}$$

The second order derivative with respect to processing rates and μ_0 are given by

$$\frac{\partial^2 \mathrm{p_E}|\mu_0}{\partial v_z \partial \mu_0} = \delta(x-z)\sum_{j,J,L}(-1)^{|L|}\frac{W}{\nu} + v_x \sum_{j,J,L}\Lambda_{z\notin J\setminus L}(-1)^{|L|}\left[\frac{\partial W}{\partial \nu}\Big/\nu - \frac{W}{\nu^2}\right] \quad \text{(C.34)}$$

where

$$\frac{\partial W}{\partial \nu} = -\frac{\partial^2 h}{\partial \nu \partial \mu_0}\mathrm{erfc}(-c) - \frac{\partial h}{\partial \nu}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} - \frac{\partial h}{\partial \mu_0}\frac{\partial \mathrm{erfc}(-c)}{\partial \nu} - h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \nu \partial \mu_0} \quad \text{(C.35)}$$

where

$$\frac{\partial^2 h}{\partial \nu \partial \mu_0} = h + \frac{\partial h}{\partial \nu}\nu \quad \text{and}$$

$$\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \nu \partial \mu_0} = \frac{2c\sigma_0}{\sqrt{2}}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0}. \quad \text{(C.36)}$$

C.6. Derivatives of $\mathrm{p_E}$ with respect to C and w

Derivatives of $\mathrm{p_E}$ with respect to C and w are obtained via chain rules which can be found in the online appendix.

Appendix D. Whole report equations

This section contains equations for computing the likelihood function and its derivatives assuming a trial of the whole report paradigm. The report set R contains the items from the stimulus set which the subject reported. Notation $\mathrm{n}(\cdot)$ is used to denote the cardinality of a set.

D.1. Likelihood function

Assuming exponential processing, the likelihood function is given by

$$\mathrm{p}(R|S) = \begin{cases} \mathrm{p}_1 \mathrm{p}(K \geq \mathrm{n}(R)|\mathbf{m}), & 0 < \mathrm{n}(R) = \mathrm{n}(S),\ t > t_0 \\ \mathrm{p}_1 \mathrm{p}(K > \mathrm{n}(R)|\mathbf{m}) + \mathrm{p}_2 \mathrm{p}(K = \mathrm{n}(R)|\mathbf{m}), & 0 < \mathrm{n}(R) < \mathrm{n}(S),\ t > t_0 \\ \mathrm{p}_1 \mathrm{p}(K > 0|\mathbf{m}) + \mathrm{p}(K = 0|\mathbf{m}), & 0 = \mathrm{n}(R) < \mathrm{n}(S),\ t > t_0 \\ 1, & 0 = \mathrm{n}(R) < \mathrm{n}(S),\ t < t_0 \\ 0, & 0 < \mathrm{n}(R) \leq \mathrm{n}(S),\ t < t_0 \end{cases} \quad \text{(D.1)}$$

where p_1 and p_2 are given below. For the ex-Gaussian model we get the likelihood function

$$\mathrm{p}(R|S) = \begin{cases} \tilde{\mathrm{p}}_1 \mathrm{p}(K \geq \mathrm{n}(R)|\mathbf{m}), & 0 < \mathrm{n}(R) = \mathrm{n}(S) \\ \tilde{\mathrm{p}}_1 \mathrm{p}(K > \mathrm{n}(R)|\mathbf{m}) + \tilde{\mathrm{p}}_2 \mathrm{p}(K = \mathrm{n}(R)|\mathbf{m}), & 0 < \mathrm{n}(R) < \mathrm{n}(S) \\ (\tilde{\mathrm{p}}_1 + \mathrm{p}_0)\mathrm{p}(K > 0|\mathbf{m}) + \mathrm{p}(K = 0|\mathbf{m}), & 0 = \mathrm{n}(R) < \mathrm{n}(S) \end{cases} \quad \text{(D.2)}$$

where $\tilde{\mathrm{p}}$ denotes convolution of p with a Gaussian. The contribution of the p_0 term to the gradient and Hessian is trivial to derive.

D.2. Details of p_1

Assuming that K is not a limiting factor, the joint probability that R is encoded into VSTM while $S \setminus R$ is not, is given by

$$\mathrm{p}_1 = \sum_{J\in\mathcal{P}(R)}(-1)^{\mathrm{n}(J)}\exp(-\nu[t-t_0]) \quad \text{(D.3)}$$

where

$$\nu = \sum_{i\in J} v_i + \sum_{j\in S\setminus R} v_j. \quad \text{(D.4)}$$

The first order derivatives with respect to processing rates and temporal threshold are given by

$$\frac{\partial \mathrm{p}_1}{\partial t_0} = \sum_J (-1)^{\mathrm{n}(J)}\exp(-\nu[t-t_0])\nu \quad \text{(D.5)}$$

$$\frac{\partial \mathrm{p}_1}{\partial v_x} = -\sum_J (-1)^{\mathrm{n}(J)}\phi_x \exp(-\nu[t-t_0])[t-t_0] \quad \text{(D.6)}$$

where

$$\phi_z = (1_{z\in J} + 1_{z\in S\setminus R}). \quad \text{(D.7)}$$

The second order derivatives are given by

$$\frac{\partial^2 \mathrm{p}_1}{\partial v_x \partial v_z} = \sum_J (-1)^{\mathrm{n}(J)}\phi_x\phi_z \exp(-\nu[t-t_0])[t-t_0]^2 \quad \text{(D.8)}$$

$$\frac{\partial^2 \mathrm{p}_1}{\partial v_x \partial t_0} = \sum_J (-1)^{\mathrm{n}(J)}\phi_x \exp(-\nu[t-t_0])(1-[t-t_0]\nu) \quad \text{(D.9)}$$

$$\frac{\partial^2 \mathrm{p}_1}{\partial t_0 \partial t_0} = \sum_J (-1)^{\mathrm{n}(J)}\exp(-\nu[t-t_0])\nu^2. \quad \text{(D.10)}$$

Derivatives with respect to processing capacity and attentional weights are obtained via chain rules identical to those in Appendix C.6.

D.3. Convolution of p_1 with a Gaussian

The convolution is a trivial variation of the integral in Appendix I

$$\int_{-\infty}^{t}\mathrm{p}_1 \mathrm{d}t_0 = \sum_{J\in\mathcal{P}(R)}(-1)^{\mathrm{n}(J)}h \times \mathrm{erfc}(-c) \quad \text{(D.11)}$$

where h and c are given in Appendix I. The convolution modifies the derivatives, the first order derivatives are obtained via

$$\frac{\partial h \times \mathrm{erfc}(-c)}{\partial \mu_0} = \frac{\partial h}{\partial \mu_0}\mathrm{erfc}(-c) + h\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} \quad \text{(D.12)}$$

$$\frac{\partial h \times \mathrm{erfc}(-c)}{\partial v_x} = -Y\frac{\partial \nu}{\partial v_x} = -Y\phi_x \quad \text{(D.13)}$$

where Y is given in (C.24). The second order derivatives are obtained via

$$\frac{\partial^2 h \times \mathrm{erfc}(-c)}{\partial v_x \partial v_z} = -\phi_x\phi_z\frac{\partial Y}{\partial \nu} \quad \text{(D.14)}$$

$$\frac{\partial^2 h \times \mathrm{erfc}(-c)}{\partial \mu_0 \partial \mu_0} = \frac{\partial^2 h}{\partial \mu_0 \partial \mu_0}\mathrm{erfc}(-c) + 2\frac{\partial h}{\partial \mu_0}\frac{\partial \mathrm{erfc}(-c)}{\partial \mu_0} + h\frac{\partial^2 \mathrm{erfc}(-c)}{\partial \mu_0 \partial \mu_0} \quad \text{(D.15)}$$

$$\frac{\partial^2 h \times \mathrm{erfc}(-c)}{\partial v_x \partial \mu_0} = -\phi_x\frac{\partial W}{\partial \nu} \quad \text{(D.16)}$$

where ∂W, and ∂Y are given in Appendix C.5.

D.4. Details of p_2

When K is a limiting factor, the probability of R is derived by considering the probability given each element of R in turn closing the race

$$p_2(R) = \sum_{i\in R} v_i \sum_{J\in\mathcal{P}(R\setminus i)} (-1)^{n(J)} \frac{1-\exp(-\nu\tau)}{\nu} \tag{D.17}$$

where

$$\nu = v_i + \sum_{k\in S\setminus R} v_k + \sum_{j\in J} v_j. \tag{D.18}$$

The structure of this is identical to a case of p_E and the convolution and derivatives are thus trivially given.

Appendix E. Partial report equations

This section contains equations for computing the likelihood function and its derivatives assuming a trial of the partial report paradigm. Let R_T denote the report set. Let subscript T denote targets, and subscript D denote distractors.

E.1. Likelihood function

Assuming exponential processing, the likelihood function is given by

$$p(R_T|S) = \begin{cases} p_1 + p_2 + p_3, & t > t_0 \\ 1, & 0 = n(R) < n(S), t < t_0 \\ 0, & 0 < n(R) \le n(S), t < t_0 \end{cases} \tag{E.1}$$

where p_1, p_2, and p_3 are given below.

E.2. Details of p_1

p_1 is the probability of the report and that VSTM does not get filled. To derive it for generalized modeling, first the probability of encoding exactly the responded subset of the target stimuli is given by

$$f_1 = \prod_{i\in R_T} [1-\exp(-v_i\tau)] \prod_{j\in S_T\setminus R_T} \exp(-v_j\tau). \tag{E.2}$$

All combinations of distractors, up to filling the VSTM capacity minus 1, or up to all distractors encoded, are considered in f_2

$$f_2 = \sum_{k=0}^{\min[K-n(R_T)-1,nD]} \sum_{J\in\mathcal{P}_k(S_D)} \times \prod_{l\in J} [1-\exp(-v_l\tau)] \prod_{m\in S_D\setminus J} \exp(-v_m\tau) \tag{E.3}$$

then by the product of f_1 and f_2 (set to zero when $n(R_T) \ge K$) we get the traditional p_1 derived in Kyllingsbæk (2006). To generalize it, we transform f_1 and f_2 to sum form

$$f_1 = \sum_{G\in\mathcal{P}(R_T)} (-1)^{n(G)} \exp(-\nu_1\tau) \tag{E.4}$$

where

$$\nu_1 = \sum_{j\in S_T\setminus R_T} v_j + \sum_{g\in G} v_g \tag{E.5}$$

and

$$f_2 = \sum_{k=0}^{\min[K-n(R_T)-1,nD]} \sum_{J\in\mathcal{P}_k(S_D)} \sum_{L\in\mathcal{P}(J)} (-1)^{n(L)} \exp(-\nu_2\tau) \tag{E.6}$$

where

$$\nu_2 = \sum_{m\in S_D\setminus J} v_m + \sum_{l\in L} v_l. \tag{E.7}$$

The product of f_1 and f_2 is then transformed to a sum

$$f_1 \times f_2 = \sum_{k=0}^{\min[K-n(R_T)-1,nD]} \sum_{J\in\mathcal{P}_k(S_D)} \times \sum_{L'\in\mathcal{P}(J\cup R_T)} (-1)^{n(L')} \exp(-\nu_\times\tau) \tag{E.8}$$

where

$$\nu_\times = \sum_{j\in S_T\setminus R_T, S_D} v_j - \sum_{m\in J} v_m + \sum_{l\in L'} v_l. \tag{E.9}$$

Finally we get the generalized p_1 conditioned on $\mathbf{m}$ by marginalizing K

$$\begin{aligned} p_1|\mathbf{m} &= \sum_{K=n(R_T)+1}^{\infty} p(K|\mathbf{m}) \times f_1 \times f_2 \\ &= \sum_{k=0}^{nD} \{1 - \mathbf{c}[n(R_T)+k]\} \sum_{J\in\mathcal{P}_k(S_D)} \\ &\quad \times \sum_{L'\in\mathcal{P}(J\cup R_T)} (-1)^{n(L')} \exp(-\nu_\times\tau) \end{aligned} \tag{E.10}$$

where $\mathbf{c}$ is the cumulative sum vector of $\mathbf{m}$. Convolution of the above sum form is trivial by Appendix I. Gradient and Hessian entries are trivial.

E.3. Details of p_2

p_2 is the joint probability of the report and the VSTM gets filled and the last item is a target. For the traditional TVA model p_2 is zero if $n(R_T) = 0$ or if $n(R_T) < K - n(S_D)$, otherwise

$$p_2|K = \sum_{i\in R_T} v_i \sum_{L\in\mathcal{P}(R_T\setminus i)} \sum_{J\in\mathcal{P}_{K-n(R_T)}(S_D)} \times \sum_{M\in\mathcal{P}(J)} (-1)^{n(L)+n(M)} \frac{1-\exp(-\nu\tau)}{\nu} \tag{E.11}$$

where

$$\nu = v_i + \sum_{k\in S_T\setminus R_T} v_k + \sum_{j\in L} v_j + \sum_{m\in S_D\setminus J} v_m + \sum_{l\in M} v_l \tag{E.12}$$

as derived in Kyllingsbæk (2006). Here we derive it for generalized K modeling: $n(R_T) = 0 \Rightarrow p_2|\mathbf{m} = 0$, otherwise

$$\begin{aligned} p_2|\mathbf{m} &= \sum_{K=n(R)}^{n(R)+n(D)} p(K|\mathbf{m})[p_2|K] \\ &= \sum_{i\in R_T} v_i \sum_{q=0}^{n(D)} p(K = q + n(R)|\mathbf{m}) \\ &\quad \times \sum_{J\in\mathcal{P}_q(S_D)} \sum_{Q\in\mathcal{P}(J\cup R_T\setminus i)} (-1)^{n(Q)} \frac{1-\exp(-\tilde{\nu}\tau)}{\tilde{\nu}} \end{aligned} \tag{E.13}$$

where

$$\tilde{\nu} = v_i + \sum_{k\in S_T\setminus R_T} v_k + \sum_{k\in S_D\setminus J} v_k + \sum_{l\in Q} v_l \tag{E.14}$$

Convolution of the above sum form is trivial by Appendix I. Gradient and Hessian entries are trivial.

E.4. Details of p_3

p_3 is the joint probability of the report and that VSTM gets filled and the last item entering VSTM is a distractor. For the traditional TVA model p_3 is zero if $\mathrm{n}(R_T) = K$ or if $\mathrm{n}(R_T) < K - \mathrm{n}(S_D)$, otherwise

$$\mathrm{p}_3|K = \sum_{i\in S_D} v_i \sum_{L\in\mathcal{P}(R_T)} \sum_{J\in\mathcal{P}_{K-\mathrm{n}(R_T)-1}(S_D\setminus i)} \times \sum_{M\in\mathcal{P}(J)} (-1)^{\mathrm{n}(L)+\mathrm{n}(M)} \frac{1-\exp(-\nu\tau)}{\nu} \tag{E.15}$$

where

$$\nu = v_i + \sum_{k\in S_T\setminus R_T} v_k + \sum_{j\in L} v_j + \sum_{m\in(S_D\setminus i)\setminus J} v_m + \sum_{l\in M} v_l \tag{E.16}$$

as derived in Kyllingsbæk (2006). Here we derive it for generalized K modeling:

$$\begin{aligned}\mathrm{p}_3|\mathbf{m} &= \sum_{K=\mathrm{n}(R)+1}^{\mathrm{n}(R)+\mathrm{n}(D)} \Lambda_{K\neq\mathrm{n}(R)} \times \mathrm{p}(K|\mathbf{m}) \times [\mathrm{p}_3|K] \\ &= \sum_{i\in S_D} v_i \sum_{q=0}^{nD-1} \mathrm{p}(K = q+\mathrm{n}(R)+1|\mathbf{m}) \sum_{J\in\mathcal{P}_q(S_D\setminus i)} \\ &\quad \times \sum_{Q\in\mathcal{P}(J\cup R_T)} (-1)^{\mathrm{n}(Q)} \frac{1-\exp(-\nu\tau)}{\nu}\end{aligned} \tag{E.17}$$

where

$$\nu = \sum_{k\in S_T\setminus R_T} v_k + \sum_{m\in S_D\setminus J} v_m + \sum_{l\in Q} v_l. \tag{E.18}$$

Convolution of the above sum form is trivial by Appendix I. Gradient and Hessian entries are trivial.

Appendix F. Optimization and error-bars

If the mode of the likelihood is approximately Gaussian, then the logarithm of the likelihood will be approximately quadratic. Assuming that the mode of the likelihood function $\mathrm{p}(\mathrm{data}|\theta)$ is Gaussian shaped with covariance matrix $\mathbf{C}$, then the Hessian matrix $\mathbf{H}^*$ of second order derivatives of the negative log-likelihood function $-\log \mathrm{p}(\mathrm{data}|\theta)$ with respect to θ, evaluated at the mode, will be symmetric and positive definite with the correspondence of $\mathbf{H}^* = \mathbf{C}^{-1}$ (McCulloch & Searle, 2001). This means that estimation variances are provided on the diagonal of $(\mathbf{H}^*)^{-1}$. By taking the square root, one can obtain the standard-error of estimate ('error-bars'; see also Sivia, 1996) which follows the $N^{-1/2}$ scaling of the standard-error of the mean in the conventional statistics due to $\mathbf{H}$ being a sum of single-trial contributions.

We use a 'damped Newton' method for robust optimization even when the Hessian is not positive definite (Nielsen, 2006). The damped Newton step regulates the search direction so that it is always within ninety degrees of the gradient descend direction. Continuous and positive parameters, C, w, and α are optimized in the domain of their logarithms. Optimization of parameters for which we do not have second order log-likelihood derivatives, is simply done by sweeping the space of possible combinations in a joint manner. For each point of the sweep, the before mentioned second order optimization procedure takes place. Optimization of the generalized K-model $\mathbf{m}$ is not done by sweeping. Instead we use a 'shaving' procedure, initialized with a uniform $\mathbf{m}$ which is then shaped iteratively by trimming individual elements. Again, this is done in joint optimization with all other parameters.

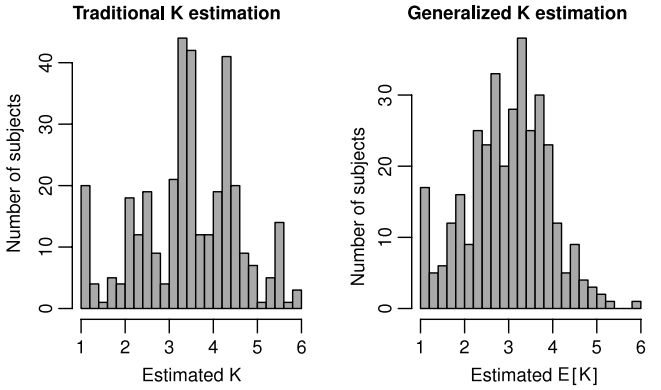

Fig. J.1. Estimated VSTM storage capacity for 347 subjects. Left: Using a traditional K-model. Right: Using a generalized K-model with trial-by-trial variability.

Appendix G. Efficient computation of sum over K

By noting that the upper limit of the sum index j inside Σ is dependent on the sum index K, the sum over K can be moved inside the sum over j as a cumulative scaling yielding the *efficient* result

$$\mathrm{p}_\mathrm{E}|\mathbf{m} = v_x\widetilde{\Sigma} + [1-\exp(-v_x\tau)]\mathrm{p}(K \geq \mathrm{n}(S)|\mathbf{m}) \tag{G.1}$$

where

$$\widetilde{\Sigma} = \sum_{j=0}^{\mathrm{n}(S)-2} \mathrm{p}(j < K < \mathrm{n}(S)|\mathbf{m}) \sum_{J\in\mathcal{P}_j(\tilde{S})} \sum_{L\in\mathcal{P}(J)} (-1)^{|L|} \frac{1-\exp(-\tau\nu)}{\nu} \tag{G.2}$$

where the scaling $\mathrm{p}(j < K < \mathrm{n}(S)|\mathbf{m})$ is readily given from the cumulative equivalent of $\mathbf{m}$. Furthermore, if there exists a number j^* such that $\mathrm{p}(j^* < K|\mathbf{m}) = 0$, that is, if all probability mass is assumed $K \leq j^*$, the summation over j can be terminated early by checking that $\mathrm{p}(j < K \leq \mathrm{n}(S) - 1|\mathbf{m})$ goes to zero. Hence, for example, when computing p_E for the traditional integer mixture, due to early termination, the number of terms in the sum of $\widetilde{\Sigma}$ is exactly the same as had p_E been computed using the less general Σ.

Appendix H. Demographics for TVA sample from Norwegian Cognitive NeuroGenetics (NCNG) sample

All participants read an information sheet and signed a statement of informed consent approved by the Regional Committee for Medical and Health Research Ethics (South–East Norway) (Project ID: S-03116). Three hundred and forty seven persons (234 females) in the age range 19–81 (Mean = 50.4, SD = 17.1) participated. All participants were recruited by advertisements in a local newspaper to take part in a larger community based study on the genetics of cognition. All subjects were interviewed and screened for neurological or psychiatric diseases known to affect the central nervous system, and history of substance abuse. Any person with a history of treatment for any of the above was excluded from further participation. The participants were administered the Vocabulary and Matrix reasoning subscales of the Wechsler Abbreviated Scale of Intelligence (Wechsler, 1999) to estimate general cognitive abilities. Participants included in the study performed within an estimated full scale IQ range of 88–148 (Mean = 120.3, SD = 10.4). The participants had 14.5 years of education on average (Range = 9–22, SD = 2.3) and 320 were right handed.

Appendix I. The convolution integral

The convolution integral is evaluated

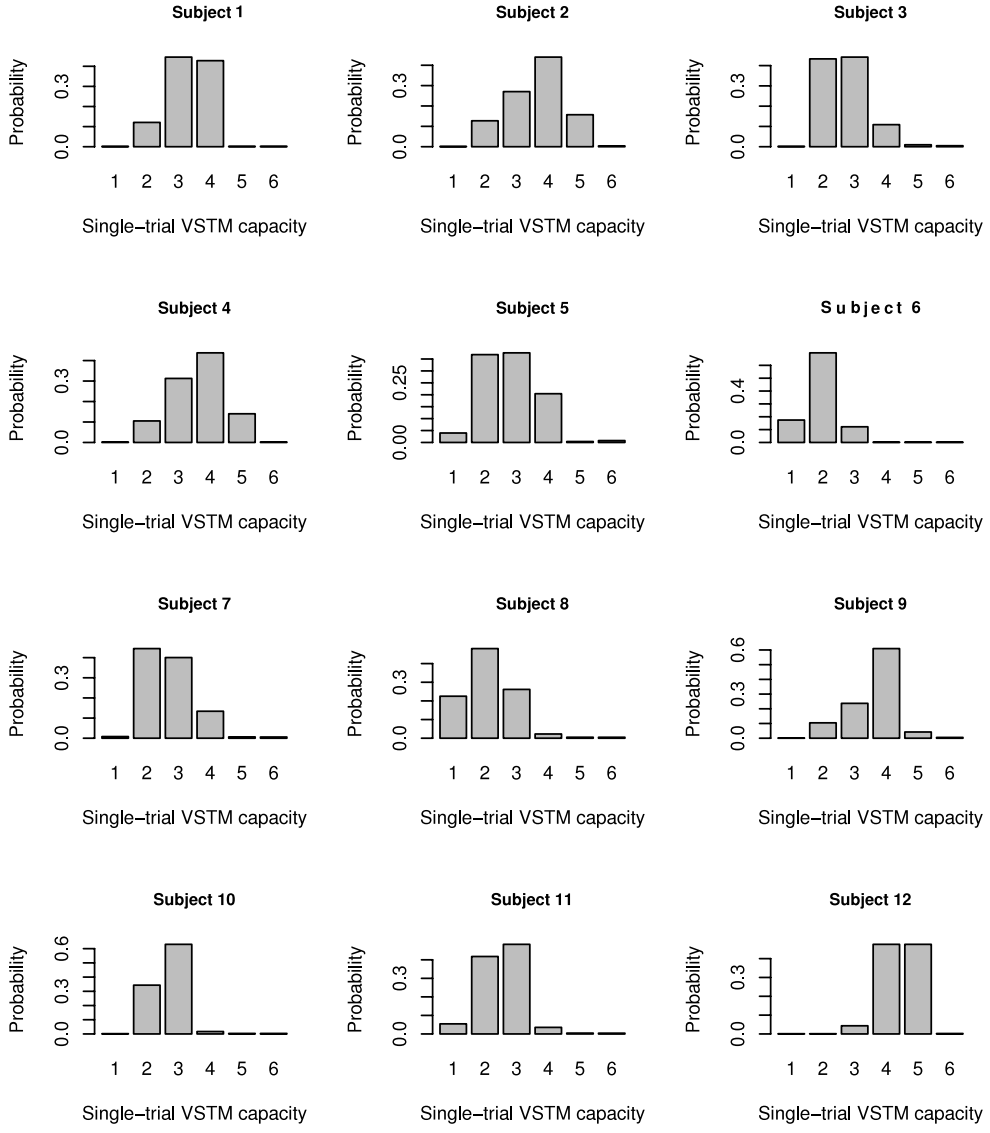

Fig. J.2. Estimated single-trial VSTM storage capacity density function for the first 12 subjects. The probabilities are estimated via the generalized VSTM storage capacity model having four degrees of freedom over the traditional VSTM storage capacity model.

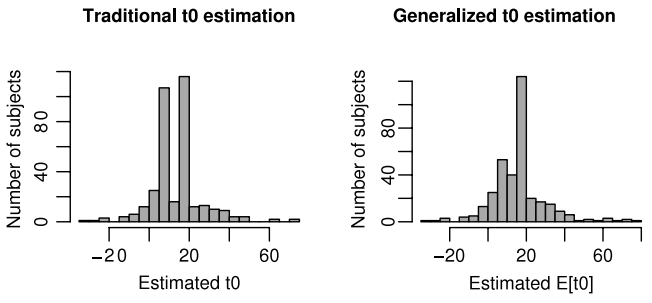

Fig. J.3. Estimated threshold for visual perception for 347 subjects. Left: Using a traditional t_0-model. Right: Using a generalized t_0-model with trial-by-trial variability.

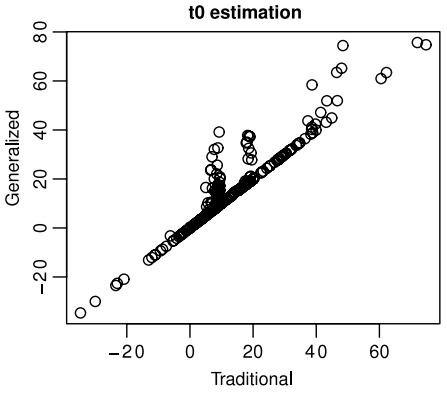

Fig. J.4. Traditional versus generalized estimation of the threshold for visual perception for 347 subjects. For most subjects the generalized estimate equals that of the traditional model. The deviations are structured in clusters displayed vertically just below 10, 20, 50 ms.

$$I = \int_{-\infty}^{t} [1 - \exp(-[t - t_0]\nu)]\mathrm{N}(t_0|\mu_0, \sigma_0)\mathrm{d}t_0$$
$$= \mathrm{erfc}(d)/2 - h \times \mathrm{erfc}(d + \nu\sigma_0/\sqrt{2}),$$
$$d = -\tau_\mu/\sqrt{2\sigma_0^2}, \tau_\mu = t - \mu_0 \quad \text{(I.1)}$$

where $d = (\mu_0 - t)/\sqrt{2\sigma_0^2}$. Thus, convolution of $1 - \exp(-\tau\nu)$ yields the substitution rule

$$1 - \exp(-\tau\nu) \leftarrow I = \mathrm{erfc}(d)/2 - h \times \mathrm{erfc}(d + \nu\sigma_0/\sqrt{2}). \quad \text{(I.2)}$$

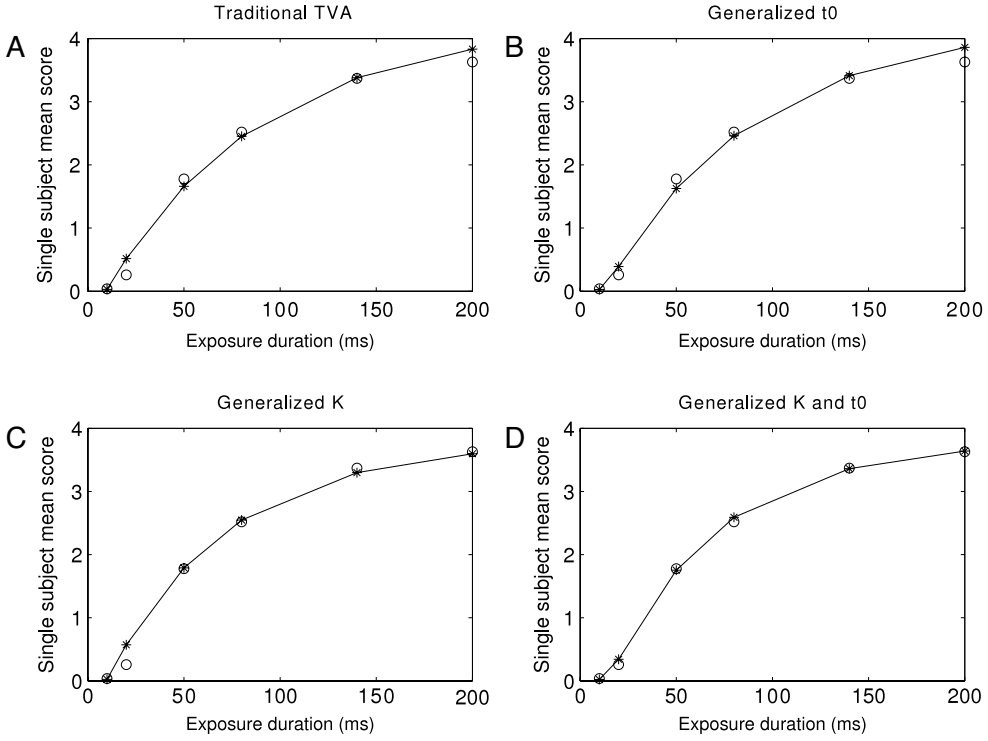

Fig. J.5. Mean score over the whole-report trials (6 elements in the stimulus display) for a single subject for which the evidence is in favor of both K and t_0 generalization. Observed data are plotted as '$\circ$'. Model predictions are shown as '$*$' connected with straight lines. (A) The traditional TVA model can be seen to not follow the S-shape of the observations at the shortest exposures, while the predicted mean score is too high at the longest exposure. (B) The generalized t_0 model does a better job of S-shaping at the rise of the curve. (C) The generalized K model corrects the prediction at the longest exposure. (D) Generalizing both t_0 and K yields an S-shape at the shortest exposures and better prediction at the longest exposure.

Appendix J. Numerical integration by Markov Chain Monte Carlo method

We use a Markov Chain Monte Carlo (MCMC) procedure for numerical integration (a good introduction is given in Ruanaidh & Fitzgerald, 1996). Our procedure is initialized by maximizing the likelihood, then 1000 MCMC samples are drawn using the Metropolis–Hastings accept/reject rule (Hastings, 1970; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953). For each MCMC sample, a feasible proposal is generated by rejection-sampling. The first 500 MCMC samples are discarded to erase the memory of initialization, the integral is then computed as the average of the remaining 500 samples.

References

Bishop, C. M. (1996). *Neural networks for pattern recognition*. Oxford, UK: Oxford University Press.

Bublak, P., Finke, K., Krummenacher, J., Preger, R., Kyllingsbæk, S., Müller, H. J., et al. (2005). Usability of a theory of visual attention (TVA) for parameter-based measurement of attention II: evidence from two patients with frontal or parietal damage. *Journal of the International Neuropsychological Society*, *11*, 843–854.

Bublak, P., Redel, P., & Finke, K. (2006). Spatial and non-spatial attention deficits in neurodegenerative diseases: assessment based on Bundesen's theory of visual attention (TVA). *Restorative Neurology and Neuroscience*, *24*, 287–301.

Bublak, P., Redel, P., Sorg, C., Kurz, A., Förstl, H., Müller, H. J., et al. (2009). Staged decline of visual processing capacity in mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging*,.

Bundesen, C. (1990). A theory of visual attention. *Psychological Review*, *97*(4), 523–547.

Bundesen, C., & Habekost, T. (2008). *Principles of visual attention*. New York: Oxford University Press.

Bundesen, C., Habekost, T., & Kyllingsbæk, S. (2005). A neural theory of visual attention: bridging cognition and neurophysiology. *Psychological Review*, *112*(2), 291–328.

Cattell, J. M. (1885). Über die Zeit der Erkennung und Benennung Schriftzeichen, Bildern und Farben. *Philosophische Studien*, *2*, 635–650.

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–114. discussion 114–185.

Duncan, J., Bundesen, C., Olson, A., Humphreys, G., Chavda, S., & Shibuya, H. (1999). Systematic analysis of deficits in visual attention. *Journal of Experimental Psychology: General*, *128*, 450–478.

Duncan, J., Bundesen, C., Olson, A., Humphreys, G., Ward, R., Kyllingsbæk, S., et al. (2003). Dorsal and ventral simultanagnosia. *Cognitive Neuropsychology*, *20*, 675–701.

Finke, K., Bublak, P., Dose, M., Müller, H. J., & Schneider, W. X. (2006). Parameter-based assessment of spatial and non-spatial attentional deficits in Huntington's disease. *Brain*, *129*, 1137–1151.

Finke, K., Bublak, P., Krummenacher, J., Kyllingsbæk, S., Müller, H. J., & Schneider, W. X. (2005). Usability of a theory of visual attention (TVA) for parameter-based measurement of attention I: evidence from normal subjects. *Journal of the International Neuropsychological Society*, *11*, 832–842.

Finke, K., Dodds, C. M., Bublak, P., Regenthal, R., Baumann, F., Manly, T., et al. (2010). Effects of modafinil and methylphenidate on visual attention capacity: a TVA-based study. *Psychopharmacology*, *210*, 317–329.

Finke, K., Schneider, W. X., Redel, P., Dose, M., Kerkhoff, G., Müller, H. J., et al. (2007). The capacity of attention and simultaneous perception of objects: a group study of Huntington's disease patients. *Neuropsychologia*, *45*, 3272–3284.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, *4*, 1–58.

Gerlach, C., Marstrand, L., Habekost, T., & Gade, A. (2005). A case of impaired shape integration: implications for models of visual object processing. *Visual Cognition*, *12*, 1409–1443.

Habekost, T., & Bundesen, C. (2003). Patient assessment based on a theory of visual attention (TVA): subtle deficits after a right frontal-subcortical lesion. *Neuropsychologia*, *41*, 1171–1188.

Habekost, T., & Rostrup, E. (2006). Persisting asymmetries of vision after right side lesions. *Neuropsychologia*, *44*, 876–895.

Habekost, T., & Rostrup, E. (2007). Visual attention capacity after right hemisphere lesions. *Neuropsychologia*, *45*, 1474–1488.

Habekost, T., & Starrfelt, R. (2006). Alexia and quadrant-amblyopia: reading disability after a minor visual field deficit. *Neuropsychologia*, *44*, 2465–2476.

Habekost, T., & Starrfelt, R. (2009). Visual attention capacity: a review of TVA-based patient studies. *Scandinavian Journal of Psychology*, *50*, 23–32.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

Hebb, D. O. (1949). *Organization of behavior*. New York: Wiley.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *The Physical Review*, *106*(4), 620–630.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*(430), 773–795.

Kyllingsbæk, S. (2006). Modeling visual attention. *Behavior Research Methods*, *38*, 123–133.

Kyllingsbæk, S., & Habekost, T. (2001). Item based fitting of whole report data. *Technical report*. University of Copenhagen.
Kyllingsbæk, S., Markussen, B., & Bundesen, C. (2011). Testing a Poisson counter model for visual identification of briefly presented, mutually confusable single stimuli in pure accuracy tasks. Journal of Experimental Psychology: Human Perception and Performance, in press (doi:10.1037/a0024751).
Lee, M. D. (2011). How cognitive modeling can benefit from hierarchical Bayesian models. *Journal of Mathematical Psychology*, *55*, 1–7.
Luce, R. D. (1986). *Response times: their role in inferring elementary mental organization*. New York: Oxford University Press.
Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279–281.
McCulloch, C. E., & Searle, S. R. (2001). *Generalized, linear, and mixed models*. Wiley.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087–1092.
Morey, R. D. (2011). A Bayesian hierarchical model for the measurement of working memory capacity. *Journal of Mathematical Psychology*, *55*, 8–24.
Nielsen, H. B. (2006). IMMOPTIBOX—a Matlab toolbox for optimization and datafitting. Available from http://www.imm.dtu.dk/hbn/immoptibox/.
Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*, 369–378.
Peers, P. V., Ludwig, C. J., Rorden, C., Cusack, R., Bonfiglioli, C., Bundesen, C., et al. (2005). Attentional functions of parietal and frontal cortex. *Cerebral Cortex*, *15*, 1469–1484.
Raftery, A. E. (1996). Approximate Bayes factors and accounting for model uncertainty in generalized linear models. *Biometrika*, *83*, 251–266.
Redel, P., Bublak, P., Sorg, C., Kurz, A., Förstl, H., Müller, H. J., et al. (2010). Deficits of spatial and task-related attentional selection in mild cognitive impairment and Alzheimer's disease. *Neurobiology of Aging*, Epub ahead of print.
Rouder, J. N., Morey, R. D., Cowan, N., Zwilling, C. E., Morey, C. C., & Pratte, M. S. (2008). An assessment of fixed-capacity models of visual working memory. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(16), 5975–5979.
Rouder, J. N., Morey, R. D., Morey, C. C., & Cowan, N. (2011). How to measure working memory capacity in the change detection paradigm. *Psychonomic Bulletin and Review*,.
Ruanaidh, J., & Fitzgerald, W. (1996). *Numerical Bayesian methods applied to signal processing*. New York: Springer.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*(2), 461–464.
Shibuya, H., & Bundesen, C. (1988). Visual selection from multielement displays: measuring and modeling effects of exposure duration. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(4), 591–600.
Sivia, D. S. (1996). *Data analysis: a Bayesian tutorial*. New York: Oxford University Press.
Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs*, *74*(11).
Sperling, G. (1963). A model for visual memory tasks. *Human Factors*, *5*, 19–31.
Sperling, G. (1967). Successive approximations to a model for short term memory. *Acta Psychologica*, *27*, 285–292.
Starrfelt, R., Habekost, T., & Gerlach, C. (2010). Visual processing in pure alexia: a case study. *Cortex*, *46*, 242–255.
Starrfelt, R., Habekost, T., & Leff, A. P. (2009). Too little, too late: reduced visual span and speed characterize pure alexia. *Cerebral Cortex*, *19*, 2880–2890.
Usher, M., & Cohen, J. D. (1999). Short term memory and selection processes in a frontal-lobe mode. In D. Heinke, & G. W. Humphreys (Eds.), *Connectionist models in cognitive neuroscience* (pp. 78–91). Springer-Verlag.
Vangkilde, S., Bundesen, C., & Coull, J. (2009). Differential effects of nicotine on discrete components of visual attention. In *Proceedings of the international neuropsychological society mid-year meeting* (p. 53).
Vangkilde, S., Bundesen, C., & Coull, J. T. (2011). Prompt but inefficient: nicotine differentially modulates discrete components of attention. Psychopharmacology, in press (doi:10.1007/s00213-011-2361-x).
Van Zandt, T., & Ratcliff, R. (1995). Statistical mimicking of reaction time data: single process models, parameter variability and mixtures. *Psychonomic Bulletin & Review*, *2*(1), 20–54.
Wechsler, D. (1999). *Abbreviated scale of intelligence*. San Antonio, TX: The Psychological Corporation.