Psychon Bull Rev (2012) 19:363–394
DOI 10.3758/s13423-012-0236-8

THEORETICAL REVIEW

The structure of short-term memory scanning: an investigation using response time distribution models

Chris Donkin · Robert M. Nosofsky

Published online: 23 March 2012
© Psychonomic Society, Inc. 2012

Abstract A classic question in cognitive psychology concerns the nature of memory search in short-term recognition. Despite its long history of investigation, however, there is still no consensus on whether memory search takes place serially or in parallel or is based on global access. In the present investigation, we formalize a variety of models designed to account for detailed response time distribution data in the classic Sternberg (Science 153: 652–654, 1966) memory-scanning task. The models vary in their mental architectures (serial exhaustive, parallel self-terminating, and global access). Furthermore, the component processes within the architectures that make match/mismatch decisions are formalized as linear ballistic accumulators (LBAs). In fast presentation rate conditions, the parallel and global access models provide far better accounts of the data than does the serial model. LBA drift rates are found to depend almost solely on the lag between study items and test probes, whereas response thresholds change with memory set size. Under slow presentation rate conditions, even simple versions of the serial-exhaustive model provide accounts of the data that are as good as those of the parallel and global access models. We provide alternative interpretations of the results in our General Discussion.

Keywords Short term memory · Response time models · Math modeling and model selection · Serial position functions

In this article, we revisit and examine from a new perspective the Sternberg (1966) short-term memory-scanning paradigm, perhaps the most venerable of all recognition memory response time (RT) tasks. In the Sternberg paradigm, participants are presented with a brief list of study items (the memory set), followed by a test item (the probe). The task is to judge, as rapidly as possible while minimizing errors, whether the probe is a member of the memory set. Sternberg's classic result was that mean RT was an approximately linearly increasing function of memory set size. Furthermore, the slope of the mean RT function for positive probes (i.e., probes that are members of the memory set) was equal to the slope for negative probes. This pattern of results led Sternberg to suggest his classic serial-exhaustive model of short-term memory search.

Sternberg's (1966) article set into motion the modern study of memory-based information processing. Since the publication of his article, the original paradigm and variants of the paradigm have been tested innumerable times, and a wide variety of different mathematical models have been developed to account for performance in the task (for a review of many of these models, see Townsend & Ashby, 1983).

Despite the wide variety of different formal models of short-term memory search that have been considered, it is surprising that there have been relatively few attempts to contrast them by considering their ability to account for RT distribution data. Indeed, we are not aware of any studies that have engaged in competitive testing of fully parameterized versions of the models with respect to their ability to account for detailed forms of RT distribution data in the

Electronic supplementary material The online version of this article (doi:10.3758/s13423-012-0236-8) contains supplementary material, which is available to authorized users.

C. Donkin (✉)
Psychology Department, Mathews Building,
University of New South Wales,
Kensington, NSW, Australia
e-mail: Christopher.donkin@gmail.com

R. M. Nosofsky
Department of Psychological and Brain Sciences,
Indiana University,
Bloomington, IN, USA
e-mail: nosofsky@indiana.edu

Sternberg task. In light of major advances in the field in the development of formal RT models and methods for evaluating them, our main aim in the present research was to fill that gap. As is described more fully below, in addition to considering some of the major classes of models, a closely related goal was to determine the types of parameter variation within the models that seem crucial to capturing performance in the task.

Beyond evaluating the models in terms of their overall quantitative fit to the detailed RT distribution data, our goal was also to evaluate them with respect to their ability to account for focused qualitative effects in the data, such as how RT changes with set size or the serial position of a target within the memory set. Much past work has also conducted such focused, analytic contrasts to distinguish between alternative models of memory search. In general, however, such contrasts often pertained to relatively simple, baseline versions of the candidate models. In the present investigation, for purposes of generality and psychological plausibility, we tend to grant the models more parametric freedom. As will be seen, under these conditions, some of the previous qualitative tests that have been applied to distinguish the models are no longer diagnostic. Therefore, the complete RT distributions take on great importance, because they provide considerable constraint on the models' predictions. Thus, we adopt a dual-route attack in this research, by examining both focused qualitative predictions from the models and their quantitative fits to complete RT distributions.

Before turning to the candidate models and describing the key issues in greater detail, we first provide a brief review of some previous studies that did involve analysis of RT distribution data in the Sternberg task.

Previous examples of memory-scanning RT distribution data

The example that is perhaps closest to the present effort comes from Ratcliff's (1978) seminal article in which he introduced his multiple-channel diffusion model. In application to the Sternberg paradigm, Ratcliff assumed that presentation of the test probe evoked a set of parallel diffusion processes, with one diffusion process per memory set item. If any individual process reached a "match" criterion, the set of parallel diffusion processes would self-terminate, and the observer would respond that the probe was "old." By contrast, in cases in which all of the individual diffusion processes reached a "nonmatch" criterion, the observer would respond "new" following exhaustive processing of all the items on the list. Ratcliff demonstrated that the diffusion model was capable of yielding accurate quantitative fits to his rich sets of individual-participant RT distribution data. In the present article, we address several questions that were not part of Ratcliff's seminal study. First, Ratcliff did not contrast the quantitative fits of his model with any major competitors. Second, the version of the diffusion model that he fitted had a very large number of free parameters. For example, he allowed a separate drift rate parameter for each unique combination of set size and serial position of the positive probes. Questions therefore arise about whether more parsimonious accounts of the data may be available. Third, his evaluation of the model focused on its quantitative fit. There was little systematic evaluation of the extent to which the model accurately captured fundamental qualitative effects in the data.

In a second example, Hockley (1984) collected RT distribution data in a wide variety of cognitive tasks. Included in his battery was the Sternberg memory-scanning paradigm. Hockley's primary aim was to characterize how the overall form of the RT distributions changed with manipulations of fundamental variables such as memory set size. Hockley observed some major differences in the form of the RT distributions that arose across the different cognitive tasks. In his memory search task, for example, there was little change in the leading edge of the RT distributions (i.e., the shortest RTs) associated with different memory set sizes. This result posed problems, for example, for certain versions of serial-processing models. On the other hand, there were big changes in the leading edge in paradigms that involved visual search. Although Hockley's experiment provided valuable information for helping to constrain models, fully parameterized versions of the models were not considered in terms of their ability to account for the RT distribution data.

Another study that collected detailed RT distribution data in a memory-scanning task was that of Ashby, Tein, and Balakrishnan (1993). The goal of their study was to characterize a variety of nonparametric properties of the distributions that place strong constraints on alternative models. Following Ashby et al., a major goal of the present article is to consider many of these properties as well. In our view, a limitation of the Ashby et al. study regards its generality and the extent to which their task was representative of the standard memory-scanning paradigm. In particular, in their paradigm, the members of the memory set were presented simultaneously on the computer screen, whereas in the standard paradigm, the items are presented in sequential fashion. Under simultaneous-presentation conditions, there is little control over the types of strategies and processes that participants may use to encode the members of the memory set. Indeed, under Ashby et al.'s experimental conditions, items in the final spatial position of each study list often had the longest mean RTs (see Ashby et al., 1993, Fig. 5). By contrast, when there are serial-position effects in sequential-presentation versions of the task, items in the final serial position almost always have the shortest mean RTs (see the

Overall Research Plan section). Possibly, the eccentricity of the final items in Ashby et al.'s simultaneous-display paradigm led to less efficient visual encoding of those items.

Finally, Nosofsky, Little, Donkin, and Fific (2011) recently reported detailed RT distribution data in a memory-scanning task. They showed that a version of a "summed-similarity" exemplar model (an extended version of the exemplar-based random-walk [EBRW] model of Nosofsky & Palmeri, 1997) gave good quantitative accounts of the data. However, as was the case in Ratcliff's (1978) study, the aim was not to conduct comparisons with competing models. In addition, Nosofsky et al. evaluated only a few qualitative properties of their RT distribution data, and more in-depth analyses are needed.

Overall research plan

In this research, we decided to consider three main candidate models defined by different information-processing architectures. Within each architecture, a variety of different parametric assumptions were considered. The goal was both to investigate whether some architectures provided superior quantitative accounts of the RT distribution than did others and to evaluate the parsimony of the alternative accounts. To help interpret the resulting quantitative fits, we also evaluated them on a battery of their qualitative predictions.

The three main candidate models included a parallel self-terminating model, a serial-exhaustive model, and what can be described as a global-familiarity model. We chose these initial candidates for reasons of generality and historical influence and because all have received considerable support in various past applications.[1] The parallel self-terminating architecture is motivated directly by Ratcliff's (1978) seminal contribution. In addition, various researchers have suggested that qualitative aspects of their memory-scanning data best supported some version of a parallel self-terminating model or something closely akin to Ratcliff's model (e.g., Ashby et al., 1993; Hockley, 1984; McElree & Dosher, 1989). Of course, the serial-exhaustive model is strongly motivated by Sternberg's (1966, 1969) results. Finally, in recent work, Nosofsky et al. (2011) found that a summed-similarity exemplar model provided good accounts of a wide variety of memory-scanning data, including good fits to detailed RT distribution data. The exemplar model is in the same general family as the global-familiarity model that we will evaluate in this work.

Importantly, we implement each of the information-processing architectures by explicitly modeling each of the elementary match/mismatch decisions that take place within that architecture. For example, recall that in Ratcliff's (1978) approach, a separate diffusion process was used to model whether the probe matched or mismatched each individual item of the memory set. We used an analogous approach to implement the present parallel self-terminating, serial-exhaustive, and global-familiarity models.

In our view, combining the information-processing architectures with an explicit model of the elementary match/mismatch processes is a crucial step. For example, this form of integration provides a principled approach to constraining the assumed shapes of the elementary RT distributions associated with comparisons of the test probe with each individual item in the memory set. Given constraints on the form of these elementary distributions, it then becomes easier to contrast the alternative architectures on their predictions of the shapes of the overall composite RT distributions that arise from searching the entire memory set. For example, in a serial-exhaustive model, the composite distributions arise by summing the individual-item comparison times. By contrast, in a parallel model, in cases involving correct "no" responses to negative probes, the composite distributions arise from the maximum (slowest) of the individual item comparisons. This integration also makes principled predictions of error rates and speed–accuracy trade-offs in the memory-scanning task.

As is explained in detail in the next section, we chose to use a linear ballistic accumulator (LBA) approach (Brown & Heathcote, 2008), rather than a diffusion approach, for modeling the elementary match/mismatch decisions. Both approaches to modeling elementary decision processes have received considerable support in the literature, and we do not believe that our main conclusions are influenced by this specific choice. Because there is now a good deal of consensus that the LBA approach provides a reasonable description of the time course of elementary decision processes, it serves as a suitable building block for the information-processing architectures that we will investigate. Furthermore, simple analytic expressions are available for computing the likelihood of RT distributions predicted by the LBA model, thereby easing considerably the computational burden in the present investigations.

In this article, we investigate the performance of our integrated models in two main versions of the memory-scanning task. In the first case, we fitted the models to the RT distribution data collected recently by Nosofsky et al. (2011). In Nosofsky et al.'s conditions, the memory set items were presented at a fairly rapid rate, and there was

[1] We note that because the architecture (e.g., serial vs. parallel) and stopping rule (e.g., exhaustive vs. self-terminating) are independent, there are a number of logical combinations that we do not consider here. The particular combinations that we do consider are ones that have received considerable support in past research on short-term memory search. Of course, we believe that our RT distribution data will also be valuable for helping to evaluate modern alternatives that are not members of these three classes, such as the iterative-resonance model of Mewhort and Johns (2005) or dual-process accounts of short-term recognition (Oberauer, 2008).

only a short interval between presentation of the memory set and presentation of the test probe. Under such testing conditions, researchers often observe strong serial-position effects in the data, with shorter RTs associated with more recently presented test probes (e.g., Corballis, 1967; Forrin & Morrin, 1969; McElree & Dosher, 1989; Monsell, 1978; Nosofsky et al., 2011; Ratcliff, 1978). It is well known that this pattern of results poses challenges to the standard serial-exhaustive model (although the versions that we investigate here can, to some degree, predict serial-position effects). To preview, we do indeed find that despite considerable parametric freedom, the serial-exhaustive model fares much worse in accounting for these data than do the other candidate models.

Sternberg (1975) suggested that different processing strategies might come into play in short-term memory scanning depending on details of the procedure. For example, in cases involving fast presentation rates and short intervals between study and test, participants may adopt familiarity-based strategies. In Sternberg's original experiments, a much different procedure was used. In particular, he used slow presentation rates and a long study–test interval. Furthermore, participants were required to recall in order the entire memory set following their recognition judgment of the probe. Perhaps serial-exhaustive search is more prevalent under those types of testing conditions. To investigate these possibilities, we collected another set of RT distribution data in which we attempted to replicate, as closely as possible, the procedures described by Sternberg (1966). To preview, we find that under these alternative conditions, the serial-exhaustive model provides as good an account of the data as the global-familiarity and parallel self-terminating models (and an arguably more natural one). Before we present the data, however, we first describe the model architectures we will be testing.

The modeling framework

To begin, we should emphasize that the aim of our investigation was to discriminate among fairly general versions of the alternative architectures for short-term memory recognition. Therefore, the models we develop do not specify detailed cognitive mechanisms that underlie the overall process. For example, we make no specific assumptions about the ways in which items are maintained or retrieved or by which similarity comparisons are made. Instead, we adopt a descriptive approach in which the outcome of these detailed mechanisms is summarized in the form of the evidence accumulation parameters of the LBA decision process. Fast evidence accumulation, for example, would arise because of some (unspecified) combination of effective maintenance, retrieval, and similarity comparison. By adopting this approach, we seek to contrast versions of the information-processing architectures that are framed at a reasonably general level.

Linear ballistic accumulator

Within each architecture, we model the elementary decision processes using the LBA model (Brown & Heathcote, 2008).[2] The LBA model is based on an evidence accumulation framework, in which evidence for potential responses is accumulated until a threshold amount is reached. In the LBA, evidence is accrued at a linear rate and without noise (ballistically) toward a response threshold. Observed RT is a combination of the time taken for the decision process and the time taken for other aspects of RT not involved in the decision process, such as the time taken for the encoding of the stimuli and the motor response.

Consider first a single LBA accumulator, as depicted in Fig. 1a. Evidence in an LBA accumulator begins with a value randomly sampled from a uniform distribution whose maximum is set at a parameter A and whose minimum is set (without loss of generality) at zero. Evidence in that accumulator then increases linearly until a threshold, b, is reached. The rate of accumulation is sampled from a normal distribution with some fixed standard deviation, s. The mean rate of accumulation, however, depends on the stimulus (or in the case of short-term memory scanning, whether the probe is a target or a lure). In particular, the mean rate, v, will be large for an accumulator that corresponds to the correct response (e.g., for an "old" response when the probe is a target) and smaller for an incorrect response (e.g., a "new" response when the probe is a target). In the context of retrieval from memory, the rate of evidence accumulation reflects the quality of evidence for a match between the probe and the contents of memory. Although we take no stance on the details of the specific retrieval process, one could think of our process as similar to that assumed by Ratcliff (1978), but with evidence being accumulated in separate, independent counters, rather than a single diffusion process.

Model architecture

There are a variety of ways that LBA accumulators can be arranged in order to produce a decision (for analogous issues in the domain of multidimensional classification, see Fific, Little, & Nosofsky, 2010). We consider the three

[2] Our choice to use the LBA instead of other models of RT is unlikely to influence our results. Donkin, Brown, Heathcote, and Wagenmakers (2011) showed that the LBA and the diffusion model, for example, are able to mimic each others' predictions closely and that parameters that share interpretation map closely onto each other. Marley and Colonius (1992) provided further justification for the use of accumulator models to account for RTs and choice.

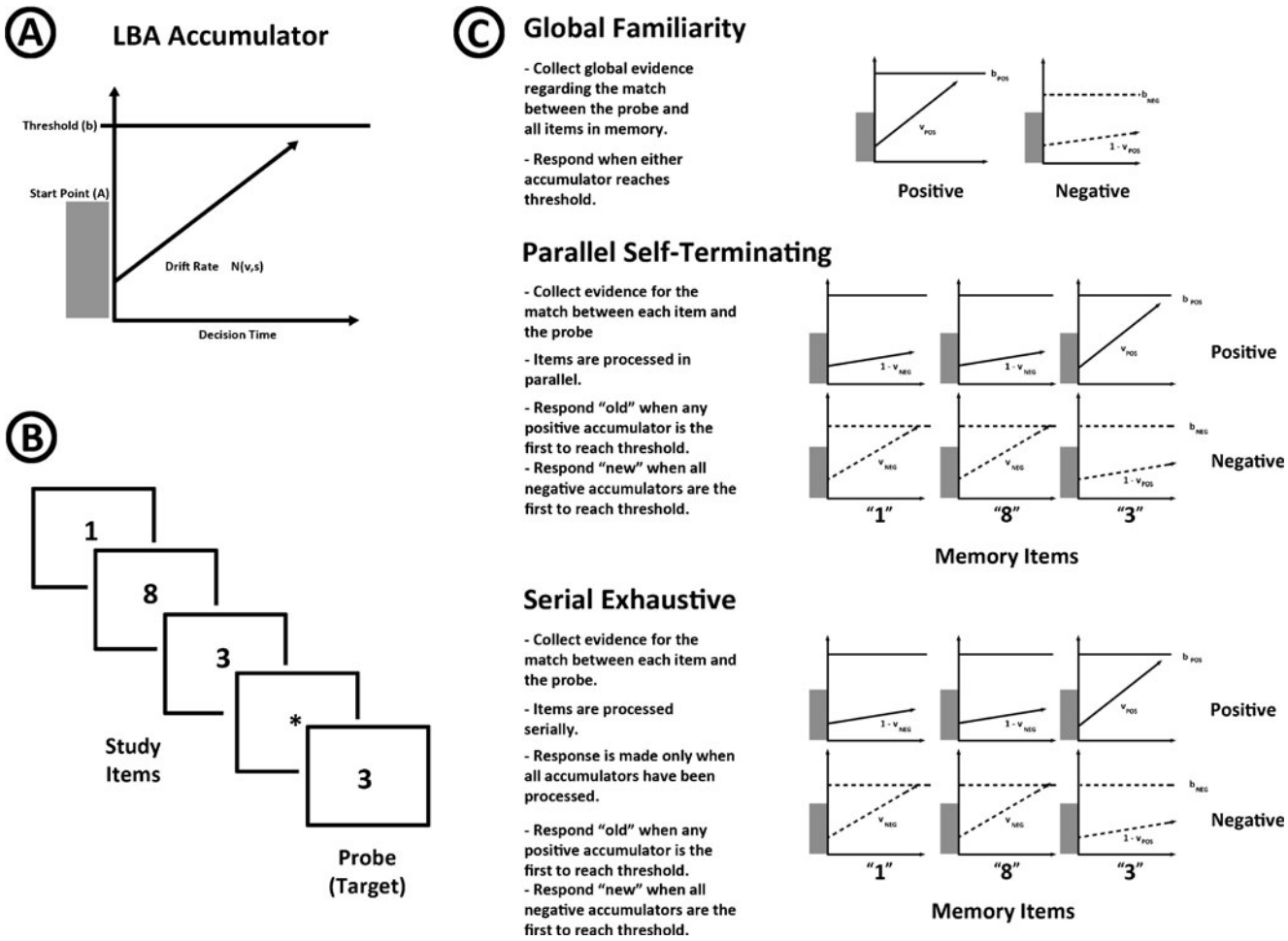

Fig. 1 **a** A single LBA accumulator. **b** A trial on which the digits 1, 8, and 3 are presented as study items and then the digit 3 is presented as a probe item. **c** The different model architectures for that trial are detailed: global familiarity (top), parallel self-terminating (middle), and serial exhaustive (bottom)

architectures that correspond to the aforementioned models of short-term memory scanning: global familiarity, parallel self-terminating, and serial exhaustive. We now discuss each of the model architectures in turn. Figure 1b, c provide a visual summary. Note that whereas the present section discusses only the general architectures, a subsequent Model Parameterizations section provides more detailed assumptions for specifying the accumulation rates and response thresholds in each architecture.

Global familiarity In a global-familiarity architecture, we assume that the recognition of a probe item is based on an overall match, or a global sense of similarity, between the probe and the representations of all memory set items. Although our implementation of the global-familiarity architecture is not tied to any particular instantiation, we envision a process that is similar to Nosofsky et al.'s (2011) EBRW model (see also Donkin & Nosofsky, in press). In that model, presentation of a probe leads to the activation of all items in memory in accord with their strength and their similarity to the probe. Highly activated items feed into a global evidence accumulation process for making "old" versus "new" decisions. Here, we model the global-familiarity process simply by assuming two LBA accumulators, one that accumulates evidence for a match between the probe and the items in short-term memory and another that collects evidence against a match between the probe and the items in short-term memory. An "old" response is triggered when the accumulator collecting positive evidence for a match is the first to reach threshold, and a "new" response is triggered when the accumulator collecting evidence against a match is the first to reach threshold.

Parallel self-terminating In a parallel self-terminating architecture, and unlike the global-familiarity architecture, we assume that the probe is compared individually with the memory representation of each of the items in the study list. All items are compared with the probe at the same time (in parallel), and a positive response is made immediately if a match between the probe and any item is detected (self-

terminating). We assume that the representation of each item in short-term memory is compared with the probe using two LBA accumulators, one collecting positive evidence (i.e., evidence for a match between the probe and the study item) and another accumulating negative evidence (i.e., against a match between the probe and the study item). An "old" response is triggered when any accumulator collecting positive evidence reaches threshold before all of the accumulators collecting negative evidence have reached threshold. A "new" response is triggered only when all of the accumulators collecting evidence against a match reach threshold before any positive accumulator reaches threshold.[3] As is illustrated by the slopes of the arrows in the middle row in Fig. 1c, the accumulator corresponding to a study item that is probed will accumulate positive evidence quickly, while negative evidence will be accumulated quickly for the remaining study items. When no study item is probed, all items in memory will accumulate negative evidence quickly.

Serial exhaustive In the serial-exhaustive architecture, we also assume, as in the parallel architecture, that the probe is compared individually with each study item. The difference is that the probe is compared with the items one at a time (i.e., sequentially), and a response is given only after all items have been compared (i.e., exhaustive). For simplicity, we assume that some fixed order of processing the items takes place, so that the mean drift rates associated with different lags do not vary across trials. We again implement this process by assuming a pair of LBA accumulators for each item in the study list, collecting positive and negative evidence for a match between the item and the probe. In the serial-exhaustive model, an "old" response is made if at least one of the positive accumulators reaches its threshold before the corresponding negative accumulator. A "new" response is made only if all negative accumulators reach threshold before their respective positive accumulators. In the serial-exhaustive model, however, each of the items is compared one after the other, and a response is triggered only after all items in short-term memory have been compared.[4]

[3] In our instantiation of a parallel self-terminating model, the "new" response is made only when all of the negative accumulators reach threshold before *any* of the positive accumulators. It is also possible to construct a parallel self-terminating model that makes a "new" response when each of the accumulators collecting negative evidence finishes before the accumulator collecting positive evidence for the same item. Eidels, Donkin, Brown, and Heathcote (2011) discussed this same issue and argued that the difference between the models is small, especially when accuracy is high, as it is in the short-term memory-scanning paradigm.

[4] The serial-exhaustive model has sometimes been questioned on grounds that it is implausible that search would continue once a match has been found. Sternberg (1975) noted various systems, however, in which an exhaustive stopping rule can actually give rise to a more efficient search process than can a self-terminating rule.

Model parameterizations

In addition to investigating the multiple model architectures, we consider a number of different parameterizations of the models. Within each of the model architectures, we fit a range of models in which drift rate and response threshold parameters are differentially allowed to vary across experimental conditions (such as the lag between a studied item and when it is probed, and the length of the study list). There are two reasons for investigating multiple parameterizations. The first is that the adequacy of an architecture may depend critically on the parameterization that is assumed. In addition, by investigating different parameterizations, we may discover which model provides the most parsimonious account of performance.

The second reason for investigating multiple parameterizations is that we can use the estimated parameters to better understand short-term memory-scanning performance. The parameters of the LBA have interesting psychological interpretations, and the way in which various empirical factors influence those parameters can be particularly revealing. For example, the rate of accumulation of evidence for the match between a study item and a probe provides an indication of the strength of the memory for that study item. Therefore, the way in which factors such as the number of items in the study list or the lag between study and probe influence the parameters of the LBA accumulators can be highly revealing of the processes underlying short-term memory.

We now present an overview of the considerations we made regarding how empirical factors may influence response thresholds and drift rates. The details of these parameterizations are provided in the Appendix.

Response threshold It seems likely that participants will set different thresholds for deciding whether there is a match or a mismatch between the probe and a study item. For example, participants may require less evidence to decide that a probe does not match a study item than to decide that the probe does match a study item. It is also possible that the length of the study list could influence the amount of evidence required to make a decision. Indeed, Nosofsky et al. (2011) found evidence that participants increase their response thresholds as the size of the study list grows. Note that, like Nosofsky et al., we assumed that when response thresholds were allowed to change with set size, they did so as a linear function of set size. (Our main conclusions are unchanged if more flexible functions are allowed.)

Drift rate We considered three different parameterizations for how empirical factors may influence the drift rates of the LBA processes. In the first, we assumed that drift rate (evidence for or against a match) is determined only by the lag of presentation between the probe and a study item.

In particular, we expected that the amount of evidence for the match between a study item and a positive probe would be strongest when the lag between study and probe items was one—that is, when the probe was the final item on the study list. We expected that evidence accumulation would slow as the lag between study and probe increased. This parameterization was motivated by previous findings that RTs for positive probes often vary strongly with their lag of presentation in the memory set (e.g., McElree & Dosher, 1989; Monsell, 1978; Nosofsky et al., 2011). As is detailed in the Appendix, drift rates associated with negative probes were also allowed to be influenced by lag of the study items.

In our second (more general) drift rate parameterization, we allowed for the possibility that the rate of evidence accumulation is *also* influenced by the size of the memory set. Again, we allowed for the possibility that drift rates are driven by the lag between study and probe. However, in this parameterization, we also allowed that there may be a systematic effect of memory set size on the drift rates as well. For example, the short-term memory store presumably has some capacity that may be strained as the size of the study list increases, thus reducing the rate at which study and probe items can be matched.

We also considered a third parameterization in which the rate of evidence accumulation was allowed to vary freely across all individual combinations of lag and study list length. This parameterization has the least constraint and allows for the possibility of complex interactions between set size and study–probe lag. We fit this version to ensure that no model architecture was stunted by the constraints placed on the evidence accumulation rate.

As is noted in the Appendix, all parameterizations made allowance for primacy effects on drift rate, because small primacy effects are often observed in the Sternberg memory-scanning task.

We present a total of 36 models: 12 different parameterizations of each of the three different model architectures. The 12 parameterizations within each model architecture come from a combination of 3 drift rate parameterizations and 4 response threshold parameterizations (see the Appendix for full details). Other parameters, including the start point variability and drift rate variability parameters of the LBA process, were held fixed across memory set size, lag, whether the probe was a target or a lure, and whether participants responded "old" or "new."

Nondecision time parameters For simplicity, in the case of the parallel self-terminating and global-familiarity models, we modeled the nondecision time component of RT using a single parameter, t_0, and this parameter was held fixed across all conditions. Because of historical precedent, in the case of the serial-exhaustive model, we allowed different nondecision times for targets and lures (Sternberg, 1975). Also, we found that fits of the serial-exhaustive model to the RT distribution data improved considerably when we made allowance for between-trial variability in the nondecision time. Thus, for the serial-exhaustive model, we modeled nondecision time as a log-normal distribution, with two separate mean nondecision times for targets and lures (T_{POS} and T_{NEG}) and a common log-normal scale parameter (S_T). (Note that in the special case in which $T_{POS} = T_{NEG}$ and $S_T = 0$, the nondecision time in the serial-exhaustive model is identical in form to what is assumed for the other models.) Finally, as is explained later in our article, to address past hypotheses advanced by Sternberg (1975) that involve encoding-time issues, we also fitted elaborated versions of the serial-exhaustive model in which nondecision time was allowed to vary across other conditions as well.

We start the investigation by fitting and comparing each of these models, using the data from Nosofsky et al.'s (2011) Experiment 2.

Nosofsky et al.'s experiment 2

Method

Nosofsky et al. (2011, Experiment 2) reported RT distributions for 4 participants, each of whom completed multiple sessions of short-term memory scanning. Memory set size ranged from one to five items, and all serial positions within each set size were tested. Two participants (1 and 2) completed 9 sessions (days) of 500 trials per session. Participants completed an equal number of trials in each memory set size condition, with the serial position probed on each trial selected randomly from within the relevant memory set. Two participants (3 and 4) completed 16 sessions of 300 trials in which each unique combination of serial position and set size was presented equally often. (Thus, for these 2 participants, smaller set sizes were tested less often than were larger ones.) Within each block of testing, the probe was a target item (a member of the study list) on half of the trials and was a lure item on the other half of the trials. Presentation orders were random within the constraints described above.

Trials were the same in both designs and began with a fixation cross presented for 500 ms. Study items were then presented for 500 ms, with a 100-ms break between stimuli. An asterisk was presented after the final study item for 400 ms, signifying that the next item presented was the probe. The probe remained on screen until a response was made, and feedback was presented for 1,000 ms, followed by a blank screen for 1,500 ms before the beginning of the next trial. The stimuli on each trial were randomly chosen from the 20 English consonant letters, excluding Y.

Results

Mean and standard deviation of response times The top row in each of Fig. 2a–d contains a plot of the mean and standard deviation of RTs for each individual participant as a function of memory set size and lag. The effect of set size on mean RTs was similar for all participants (first column). As has been observed in numerous past studies, mean RTs increased roughly linearly with set size, and the functions for targets and lures were roughly parallel to one another. As is shown in the second columns of Fig. 2, however, the slowdown with set size for positive probes is largely due to increasing lags between the probes and matching memory set items. That is, the slowdown occurs in large part because larger set sizes contain items that have a larger lag between study and probe. Nevertheless, holding lag fixed, there is also some evidence of

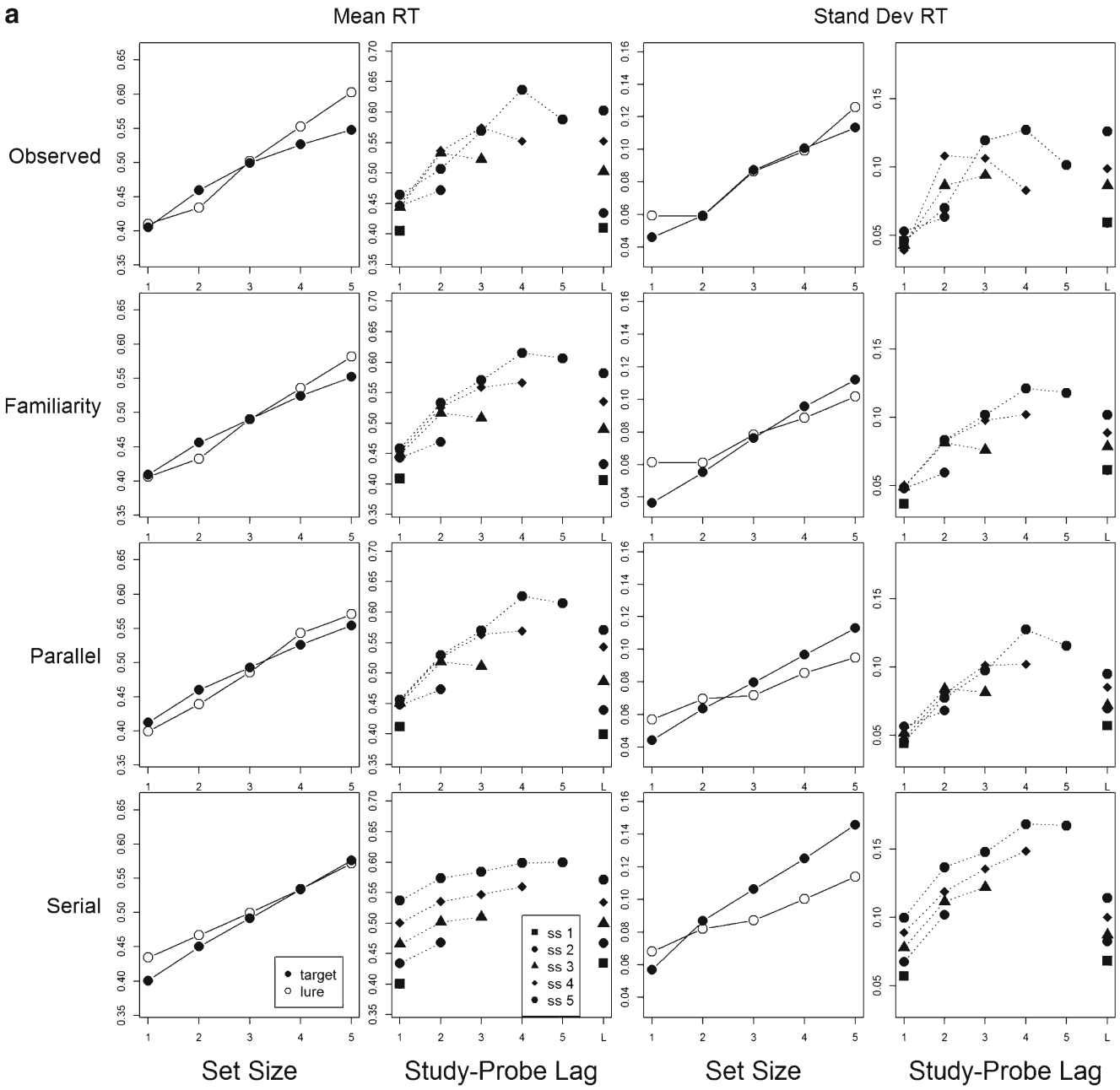

Fig. 2 Observed (top row) and predicted (second through fourth rows) means and standard deviations of response times (RTs) plotted as a function of set size only (first and third columns) and as a function of study–probe lag (second and fourth columns). In the study–probe lag plots, "X" refers to responses on lure trials; "ss" in the legend refers to set size. Model predictions come from models with drift rates that vary as a function of study–probe lag and study set size, and with response thresholds that vary across response and study set size (see the text for details)

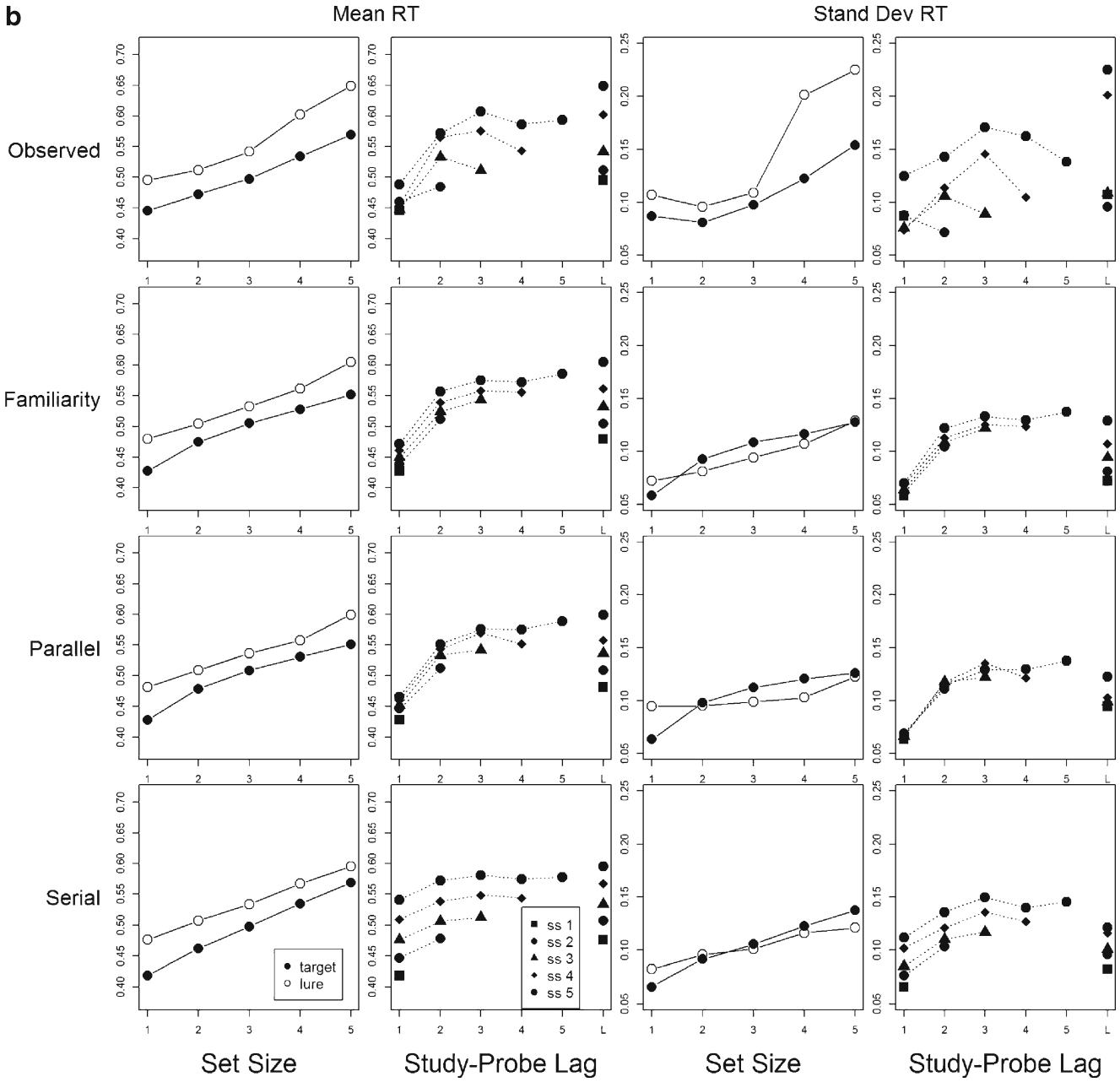

Fig. 2 (continued)

an effect of set size per se. For example, at lag 1, mean RTs get longer as set size increases.

The standard deviations in RTs, presented in the third and fourth columns of Fig. 2, tend to show a similar pattern as do the mean RTs, but they are much noisier. In general, the standard deviations increase with set size for both targets and lures. For the targets, this increase seems to be largely due to the influence of study–probe lag.

Response time distributions The RT distributions for correct responses for target and lure items are plotted for each participant in Fig. 3a (and again in 3b). The RT distributions are shown separately as a function of lag and set size for the targets (main columns 1–4) and as a function of set size for the lures (main column 5). The observed RT distributions (solid squares) in the figure are summarized using the .1, .3, .5, .7, and .9 quantiles, which represent the times at which 10%, 30%, 50%, 70%, and 90% of responses were made. The use of quantiles offers quick comparison between the RT distributions for different lags and set sizes. The height of the quantiles indicates the speed of responses, with central quantiles indicating the central tendency and the distance between quantiles indicating the spread and shape of the distribution. The lowest quantile within each distribution

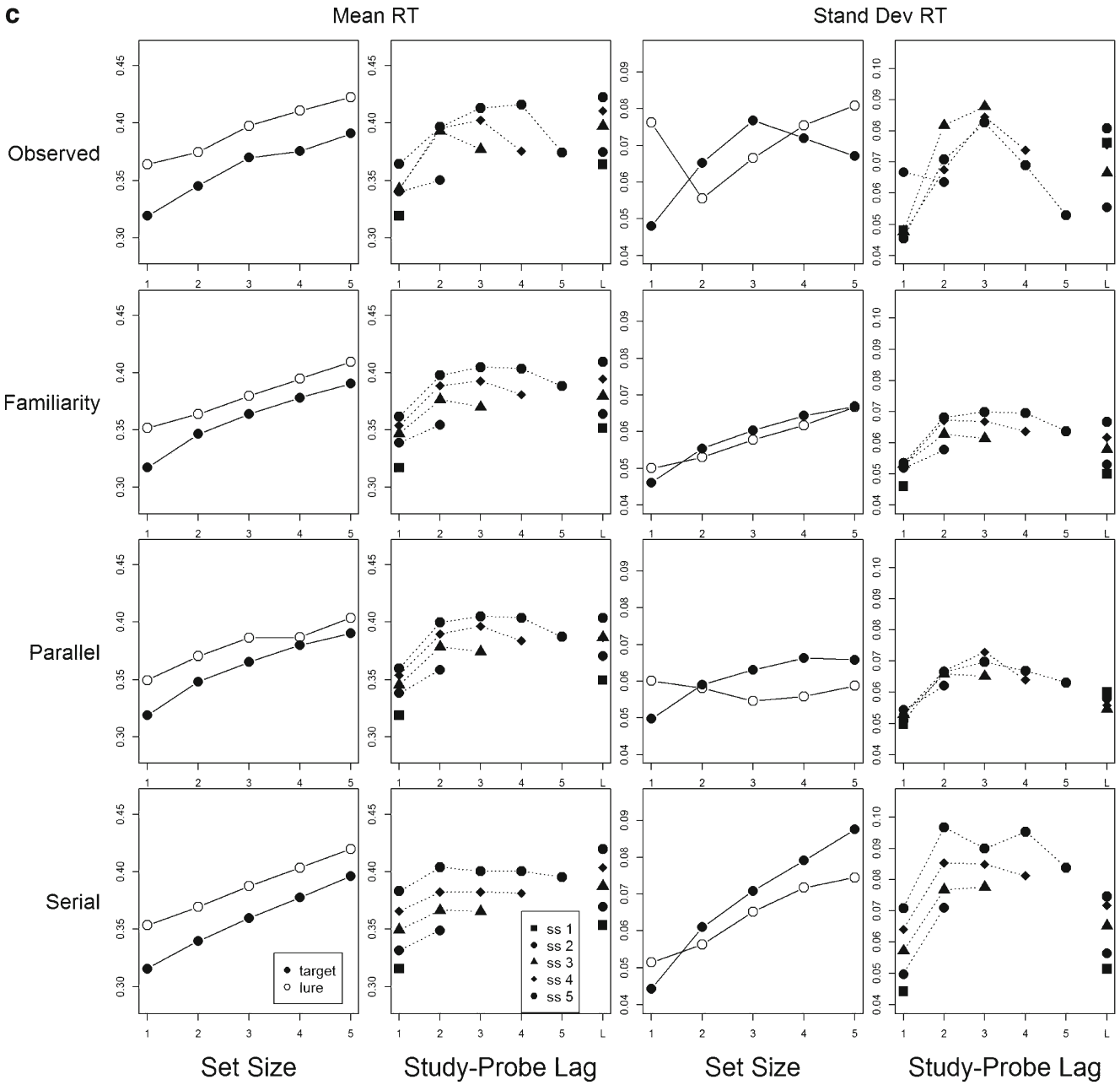

Fig. 2 (continued)

provides an indication of the leading edge of the distribution (i.e., the shortest RTs). The probability of making a correct response in each given lag and set size condition is reported underneath its respective set of quantiles. The proportion of correct responses is very high for most conditions, so we omit reporting the RT distributions for incorrect responses. Note, however, that models were fit to both correct and incorrect distributions, and all of those models that fit the data well also accounted for the speed of incorrect responses.

It is clear from Fig. 3 that the lag between the study item and test probe had a large effect on performance, as indicated by both a reduction in the proportion of correct responses and a lengthening of RTs as lag increased. That is, within each particular set size, there is a pattern in which quantiles increase in their location and spread as lag increases. For most participants, there is also a small effect of primacy: That is, within each set size, the item with the greatest lag (i.e., the item in the first serial position) is generally more accurate and faster than the preceding lag (i.e., the item in the second serial position). RTs also lengthen due to increases in the size of the study set, as reflected by the fact that all quantiles tend to increase in their vertical placement as one moves from left to right across the plot. This effect of set size is particularly evident for the lures.

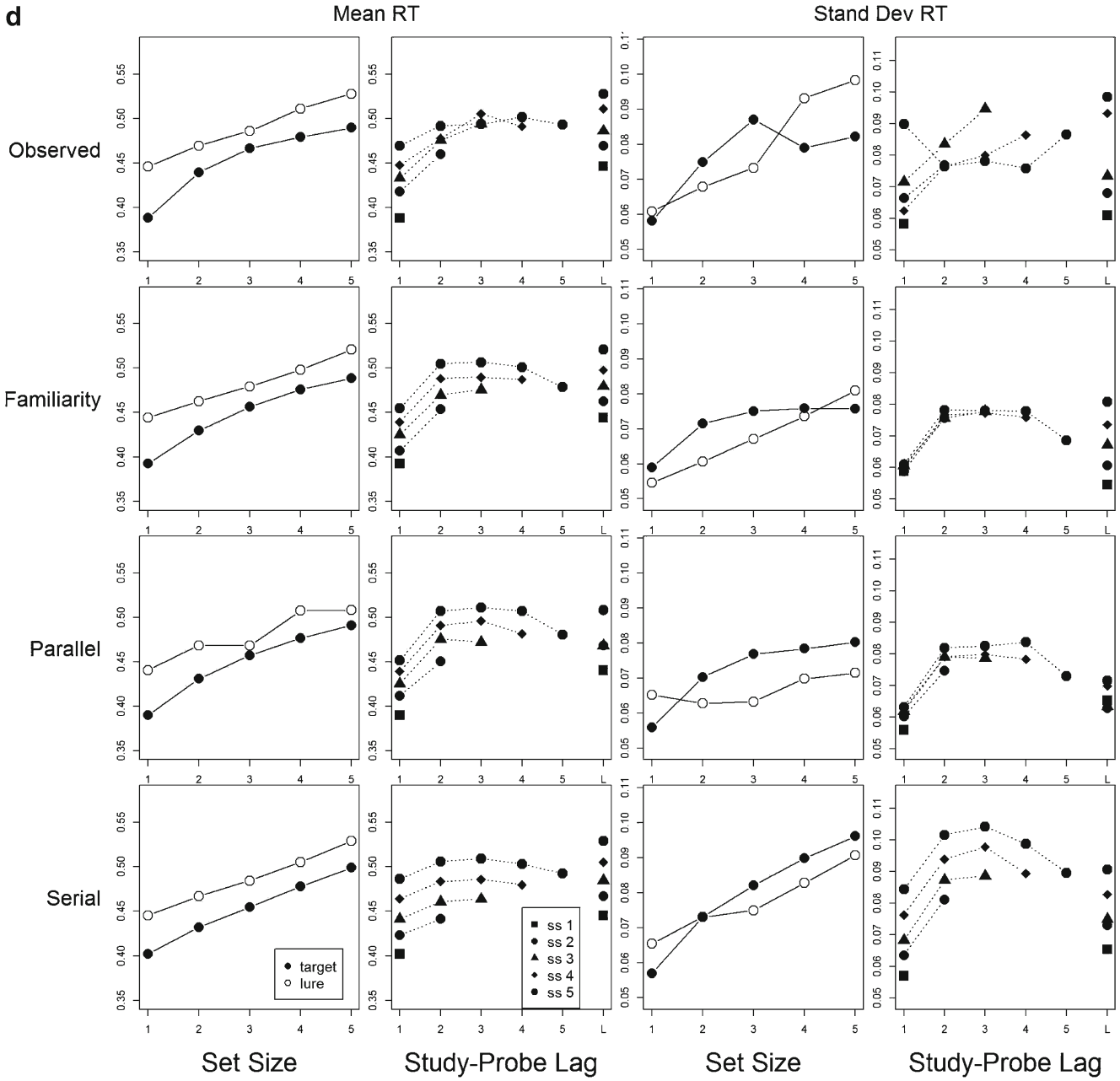

Fig. 2 (continued)

Importantly, in addition to an increase in the spread of the distributions, there is a slowdown in the entire distribution of RTs, due to increasing memory set size. This increase manifests itself as a shift in even the fastest quantile, an effect particularly evident when the probe was a lure.

We have now identified a wide array of qualitative patterns in the memory-scanning data that a successful model must account for. These patterns will provide a strong test of our candidate architectures for short-term memory scanning. First, however, we use model selection methods to identify which architectures and parameterizations provide the most parsimonious quantitative accounts of the complete RT distributions.

Model selection

We fit all 36 models (3 architectures × 12 parameterizations) to each of the 4 participants' full RT distributions for correct and incorrect responses. The global-familiarity and parallel self-terminating models were fit to the individual-trials data by using maximum likelihood estimation, and optimal parameters were found using SIMPLEX (Nelder & Mead, 1965). Closed-form analytic likelihood expressions exist for both the global-familiarity and parallel self-terminating architectures, but not for the serial-exhaustive architecture. The serial-exhaustive models, therefore, were fit by

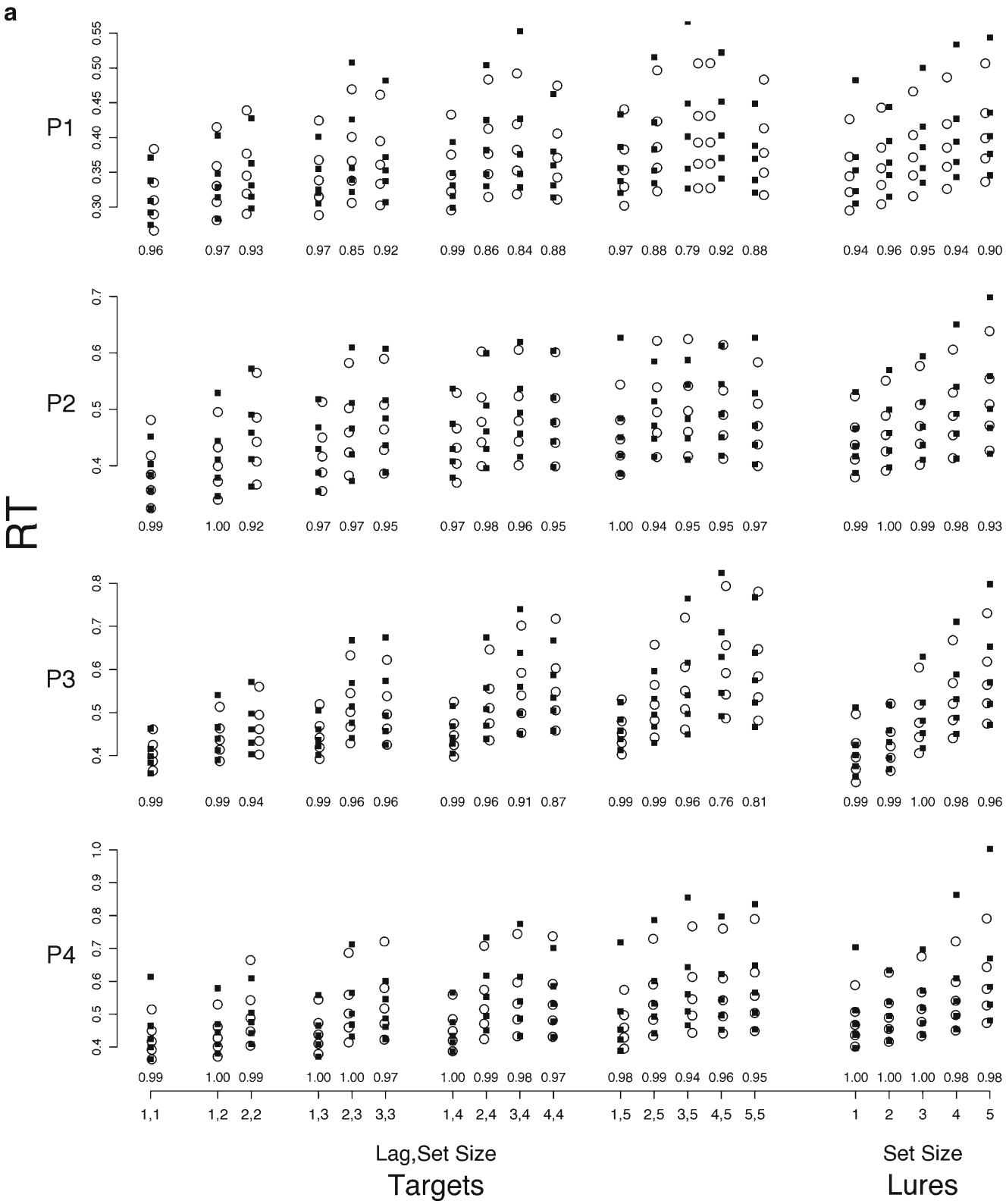

Fig. 3 Observed and predicted quantiles (black and gray points, respectively) for correct responses from the 4 participants reported in Experiment 2 of Nosofsky, Little, Donkin, and Fific (2011). Quantiles are plotted as a function of whether the probe was a target or a lure, the length of the study list (set size), and the lag between study item and probe (lag). The observed proportion of correct responses in each condition is also reported underneath their respective quantiles. Predictions are given for the global-familiarity architecture (**a**) (see the online supplement for predictions from the parallel self-terminating model) and for the serial-exhaustive architecture (**b**). The full parameterization used for all architectures allowed response thresholds to vary over study set size and evidence valence and drift rate to change over study–probe lag and set size. P1 refers to participant 1, P2 to participant 2, and so forth

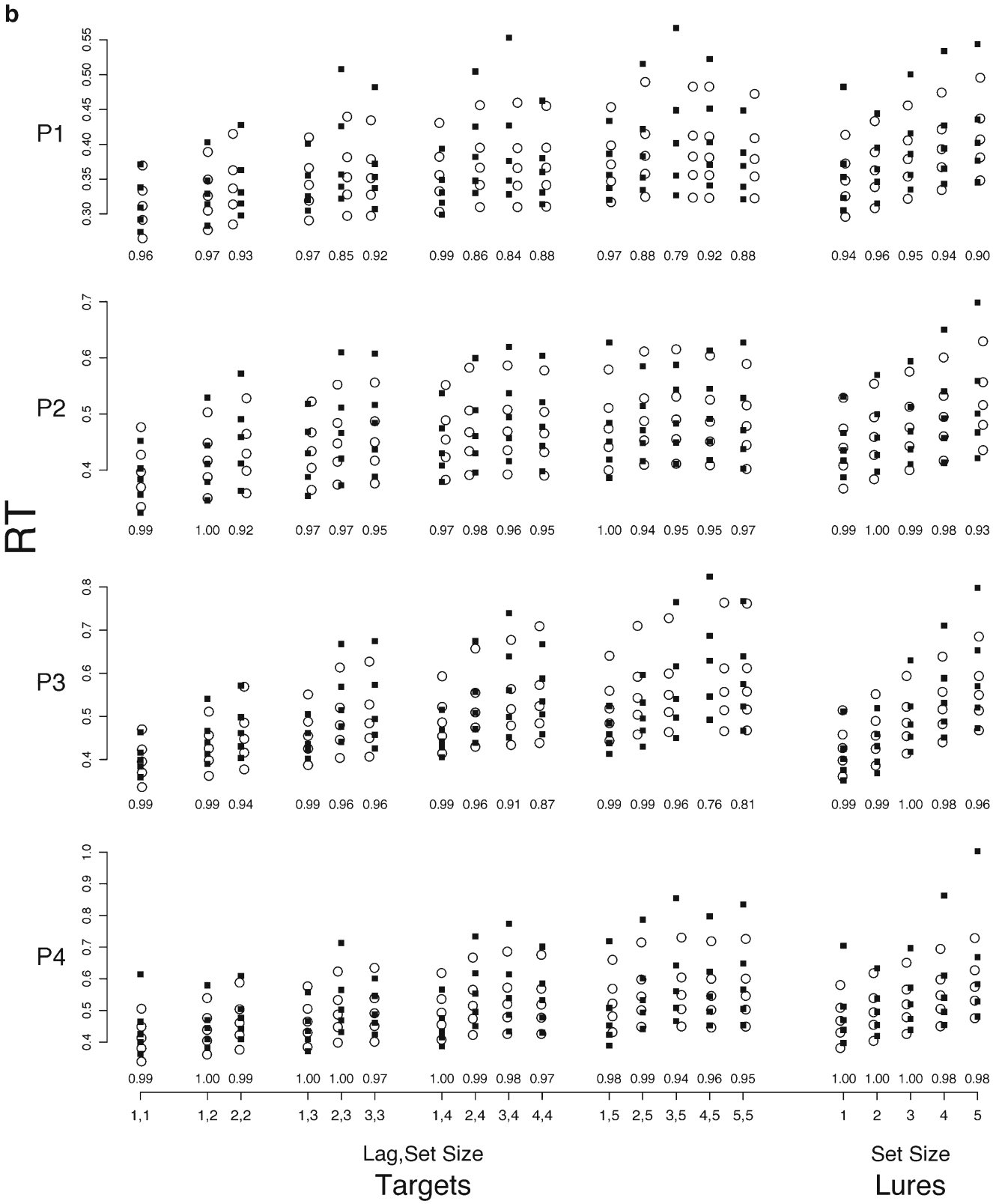

Fig. 3 (continued)

simulation. We simulated 20,000 trials for each combination of serial position and set size to produce predicted RTs for the serial-exhaustive models. Observed and simulated RTs for correct and incorrect responses within each condition were divided into 50-ms bins, from 150 to 3,000 ms, and their agreement was evaluated using the multinomial log-likelihood function (Eq. 9 in Nosofsky et al., 2011). Furthermore, as is explained below, to allow meaningful comparisons among all three classes of models, we also fitted members of the class of global-familiarity and parallel models, using these same simulation techniques. In all cases, extensive efforts were made to ensure that any differences in the fit quality of the serial-exhaustive models were not due to the fitting method used. For example, all models were fit with multiple start points, which most often converged to the same part of the parameter space. (Furthermore, when we fit the global-familiarity and parallel self-terminating models using the simulation method, we found that they converged closely to the same best-fitting parameters as were estimated using the analytic solutions.)

We used the Bayesian information criterion (BIC) to determine which architectures and parameterizations provided the most parsimonious account of the data (Schwarz, 1978; for a discussion of using BIC to evaluate models of RT, see Donkin, Brown, & Heathcote, 2011). BIC is made up of two parts: The first is a measure of how well the model fits the data, and the second is a penalty term based on the number of parameters in the model. In this way, BIC selects models that are only as complex as is necessary to explain the major patterns in the data, while penalizing models that incorporate unnecessary parameters.

BIC was calculated for all 36 models separately for each participant, using the following formula: $\mathrm{BIC} = -2 \times \log L + k \times \log N$, where k is the number of parameters in the model, N is the total number of data points used to fit the model to the data, and $\log L$ is the log-likelihood of the best set of parameters given the observed data. The term $-2 \times \log L$ represents quality of fit and becomes smaller as the model fit improves, whereas $k \times \log N$ is a penalty term that causes BIC to increase as the number of free parameters increases. The best model is the one that yields the smallest BIC value.

Recall that analytic likelihoods for the serial-exhaustive models were not possible and were, instead, calculated on the basis of simulation. Therefore, to compare the three architectures, we took the best-fitting model from each of the global-familiarity and parallel self-terminating architectures and refit them in the same way as that in which the serial-exhaustive models were fit, via simulation. In Table 1, we report the ΔBIC values that come from this procedure. ΔBIC values were calculated by taking the difference between each model's BIC and the model with the smallest BIC value. A ΔBIC value of 0, therefore, represents the best-fitting model. Examination of the table reveals that the ΔBIC values for even the best serial-exhaustive model were much larger than those for the global-familiarity and parallel self-terminating models for all 4 participants. Later in this section, we consider an elaborated version of the serial-exhaustive model that makes allowance for encoding-time differences depending on the lag with which a positive probe was presented (cf. Sternberg, 1975, p. 12). As will be seen, even this elaborated serial-exhaustive model performs far worse than do the alternative architectures. Therefore, we now focus on the results from the global-familiarity and parallel self-terminating models.

Table 1 ΔBIC values based on simulation for each of the global-familiarity (GF), parallel self-terminating (PST), and serial-exhaustive (SE) architectures for each of the 4 participants in Nosofsky, Littlel, Donkin, and Fific's (2011) Experiment 2

	ΔBIC				*k*
	P1	P2	P3	P4	
GF	0	0	0	26	19
PST	27	56	18	0	19
SE	146	124	689	518	21
SE w/enc	117	85	645	353	24

Note. For all architectures, we report the best-fitting version of the model and the number of parameters. We also include the ΔBIC values for the serial-exhaustive model that includes encoding time that varies with study–probe lag (SE w/enc)

Figure 4 shows the ΔBIC values (calculated using the analytic methods) for each of the 24 models from the global-familiarity and parallel self-terminating architectures for each of the 4 participants in Nosofsky et al.'s (2011) Experiment 2. Note first that the best-fitting versions of the parallel and global-familiarity models provide roughly equally good accounts of the data, although particular parameterizations within each architecture fare better than others. We organize our discussion of the results of our model fitting into sections regarding, first, the response threshold parameterizations and, second, drift rate parameterizations.

The results for the response threshold parameterizations are clear-cut. Regardless of architecture, models with response thresholds that changed across study set size *and* were different for positive and negative match accumulators were generally preferred over models that did not allow for these forms of flexibility. To see this result, note that the fourth cluster of bars within both model architectures (labeled EV + SS in the figure) tends to have ΔBIC values closest to 0 for both the global-familiarity and parallel models. Indeed, the best models according to BIC were almost always those in which response thresholds varied both for set size and for positive and negative match accumulators.

Turning to the drift rate parameterizations, the most consistent pattern was that models in which the rate of evidence

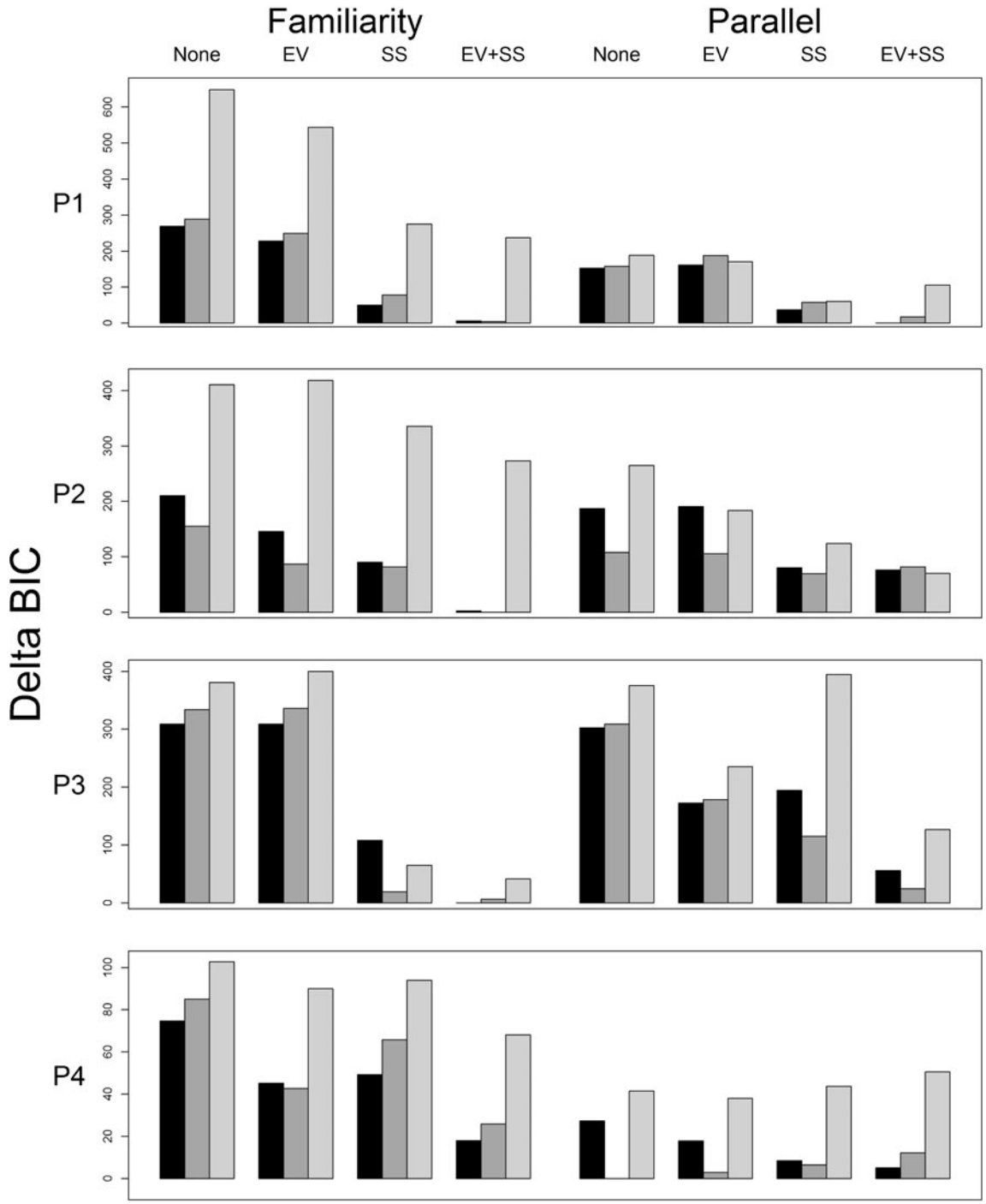

Fig. 4 ΔBIC values are plotted for each of the 24 global-familiarity and parallel self-terminating models fit to each of the 4 participants reported in Experiment 2 of Nosofsky, Little, Donkin, and Fific (2011). Smaller values of ΔBIC indicate models that are more parsimonious. The models vary according to their architecture: global familiarity (the first four clusters of bars) and parallel self-terminating (second four clusters). The different response threshold parameterizations implemented, represented by each cluster of three bars, were the following: none, one response threshold for all conditions; EV, separate thresholds depending on the evidence valence (whether evidence was collected for a positive or a negative match between study item and probe); SS, linearly increasing thresholds based on study set size; and EV + SS, a combination of the former two parameterizations. The three different drift rate parameterizations, represented by black, dark gray, and light gray bars, respectively, were the following: lag, drift rate changed due to the lag between study and probe; lag + SS, –drift rate changed due to lag and as a linear function of set size; and free, drift rate was estimated freely across all combinations of set size and serial positions. P1 refers to participant 1, P2 to participant 2, and so forth

accumulation was allowed to vary freely over all individual combinations of memory set size and lag (lightest gray bars) were not favored by BIC. These free drift rate models suffered because the data did not require the degree of flexibility permitted by their numerous free parameters. The models whose drift rates were determined by study–probe lag (and primacy) or a systematic combination of lag and set size were preferred over the free model. However, it is hard to distinguish between the two constrained drift rate parameterizations (i.e., *lag only* or *lag plus set size*).

In sum, the quantitative fits indicate that the best global-familiarity and parallel self-terminating models yield essentially identical accounts of the RT distribution data, with the serial-exhaustive model providing a decidedly worse account. The evidence also indicates that the magnitude of the response thresholds varied with memory set size (and with the positive or negative valence of the evidence accumulator). Finally, models that assume that drift rate is a systematic function of lag (and perhaps set size as well) provide more parsimonious accounts of the data than do ones that allow drift rate to vary freely across all unique combinations of lag and set size.

Before evaluating the models on their ability to capture the fundamental qualitative effects in the data, we first briefly examine the specific pattern of best-fitting parameter values from the preferred model versions. These parameter values are reported in Table 2. We focus our discussion on the global-familiarity and parallel models. First, consistent with our expectations, for all participants, drift rates for positive probes tend to decrease as a function of the lag between study and probe. In other words, the strength of evidence for a match between study and probe items tends to decrease as lag increases. Second, once one takes lag into account, there is no evidence that increases in set size lead to any *decreases* in drift rate toward the correct response thresholds. (If anything, in those cases in which the *lag plus set size* model yielded superior BIC values than did the *lag only* model, ΔS was slightly positive, not negative.) As we

Table 2 Parameter estimates for each of the global-familiarity (GF), parallel self-terminating (PST), and serial-exhaustive (SE) architectures for each of the 4 participants in Nosofsky, Little, Donkin, and Fific's (2011) Experiment 2

	P1			P2			P3			P4		
	GF	PST	SE	GF	PST	SE	GF	PST	SE	GF	PST	SE
s	0.22	0.26	1.15	0.27	0.32	1.21	0.18	0.24	1.19	0.15	0.17	1.15
T_{POS}	0.25	0.26	0.36	0.22	0.22	0.38	0.16	0.19	0.29	0.15	0.16	0.38
T_{NEG}			0.39			0.44			0.33			0.42
S_T			0.001			0.001			0.001			0.001
A	0.09	0.13	0	0.08	0.12	0	0.06	0.09	0	0.07	0.08	0
b_{POS}	0.23	0.24	0.08	0.27	0.32	0.1	0.14	0.13	0.03	0.19	0.22	0.05
b_{NEG}	0.19	0.18	0.06	0.32	0.33	0.05	0.17	0.17	0.03	0.28	0.26	0.03
Δb_{POS}	0.009	0.021	0	0.01	0.02	0	0.01	0.01	0	0.02	0.02	0
Δb_{NEG}	0.04	0.02	0	0.01	0	0	0.008	0	0	0	0	0
v_1	1.05	1.04	2.66	1.11	1.24	3.38	0.68	0.81	2.12	0.69	0.80	2.78
v_2	0.79	0.76	1.96	0.85	0.95	2.52	0.56	0.63	1.45	0.59	0.66	1.90
v_3	0.70	0.66	1.82	0.80	0.88	2.36	0.54	0.61	1.48	0.58	0.65	1.90
v_4	0.57	0.53	1.41	0.81	0.88	2.33	0.54	0.61	1.57	0.60	0.67	1.99
v_5	0.51	0.46	1.37	0.77	0.82	2.32	0.54	0.64	1.78	0.65	0.70	2.55
u_1	0.75	0.80	2.00	1.09	1.05	1.96	0.74	0.76	1.81	0.85	0.78	2.14
u_2	0.84	1.28	2.63	1.04	1.23	2.41	0.73	0.87	2.42	0.80	0.83	2.40
u_3	0.78	0.66	2.67	0.99	1.06	2.49	0.71	0.72	2.31	0.76	0.92	2.63
u_4	0.78	0.60	2.61	0.95	1.02	2.28	0.70	1.02	2.42	0.71	0.66	2.38
u_5	0.77	0.55	2.51	0.88	0.83	2.40	0.69	0.59	2.31	0.67	1.15	2.30
P	1.17	1.19	1	1	1.04	1	1.09	1.08	1	0.99	1.04	1
ΔS	0.02	0.07	−0.004	0	0.02	−0.001	0.003	0.04	0.001	0.04	0.02	0

Note. The parameter estimates are the maximum likelihood estimates based on fits of the model in which response thresholds changed across study set size and evidence valence and drift rate changes systematically across study–probe lag and set size. Note that for the serial-exhaustive model, T_{POS}, T_{NEG}, and S_T refer to the means and standard deviation of a normal distribution that is then exponentiated to produce the log-normal distribution. Remaining parameters are as follows: s and A represent between-trial variability in drift rate and start point, respectively; b parameters represent response thresholds; v_i and u_i represent drift rates for a positive and negative match, respectively, between an item at lag i and the probe; P represents a primacy parameter; and ΔS represents the change in drift rate with set size

discuss more fully in our General Discussion section, this result has interesting implications for the interpretation of *limited-capacity* parallel-processing accounts of short-term memory scanning. Finally, for the global-familiarity and parallel architectures, thresholds for positive-match responses increase with memory set size ($\Delta b_{POS} > 0$), with negative-match response thresholds tending to show less increase. Apparently, as memory set size grows, observers choose to accumulate more evidence before making their recognition decision. Armed with this information regarding the pattern of best-fitting parameters, we now discuss the models' account of the main qualitative effects in the data.

Qualitative model predictions

Mean and standard deviation of response times The predicted means and standard deviations of RTs from the global-familiarity, parallel self-terminating, and serial-exhaustive architectures are shown in the second through fourth rows of Fig. 2. For all architectures, the specific model versions are the ones with free parameters reported in Table 2.

As can be seen in Fig. 2, both the global-familiarity and parallel models generally account very well for the mean RTs, regardless of whether they are plotted as a function of set size only (column 1) or as a function of set size and lag (column 2). (The main misfit is that, for participant 2, the two models underpredict the influence of primacy—i.e., the final lag position in each set size.) The models' account of the lag effects is straightforward. As lag increases, drift rates decrease (see Table 2), so the time for a match decision to complete gets longer. The main exception is a systematic primacy effect: Within each set size, the item with the greatest lag (i.e., the item in the first serial position) receives a boost to its drift rate. The set size effects for positive probes (column 1) arise for two reasons. First, as set size increases, there is an increased proportion of cases in which the matching item has high lag. Second, observers set higher response thresholds as set size increases (Table 2), which slows down the completion of the LBA processes. Note that the increased response thresholds tend to cause (slightly) longer RTs with increased set size even when lag is held fixed (column 2).

The models predict pronounced set size effects for the negative probes for several reasons. Regarding the parallel model, note that a "no" response is made only when the LBA processes for all of the individual memory set items reach their negative response thresholds. In general, the larger the set size, the longer it takes for the slowest such process to complete. Regarding the global-familiarity model, the larger the set size, the slower is the global rate of drift toward the negative response threshold (Table 2). Such a result is expected because, as set size increases, the summed similarity of a negative probe to the memory set items will tend to increase (Nosofsky et al., 2011). Finally, set size effects for the negative probes also arise because, at least in some cases, the larger the set size, the greater is the setting of the response threshold parameters (Table 2).

Unlike the parallel and global-familiarity models, the serial-exhaustive model has difficulty accounting for the serial-position effects on the mean RTs. As is well known, the baseline version of the serial-exhaustive model considered by Sternberg (1966) predicts flat serial-position curves. We should emphasize that our more fully parameterized version of the serial-exhaustive model is not subject to this constraint (see the fourth rows of Fig. 2). The total decision time from the serial-exhaustive model arises as a sum of the individual-item comparison times across all items in the memory set. As lag decreases, the individual-item comparison time for a positive probe to match its representation in memory decreases, so the overall sum decreases. Nevertheless, the present version of the serial-exhaustive model underpredicts the magnitude of the serial-position effects. In addition, its main failing is that, once one conditionalizes on study–probe lag, it predicts too much separation between the different set size functions (column 2). The reason is that, the greater the set size, the greater will be the sum of the individual-item comparison times associated with the mismatching items from the memory set.

Regarding the standard deviation of RTs, the global-familiarity and parallel self-terminating models do a reasonable job of capturing the main qualitative trends in the observed data (see Fig. 2). However, their quantitative fit is not as good as it was for the mean RTs. Both models tend to underpredict the standard deviation of RTs on lure trials, particularly for the larger set sizes. The serial-exhaustive model shows a similar limitation.

Predicted quantiles The open circles in Fig. 3a, b represent the predicted quantiles from the global-familiarity and serial-exhaustive architectures, respectively.[5] The predictions from the parallel self-terminating model were almost identical to those of the global-familiarity model, so we have placed them in an online supplement. The vertical displacement between predicted and observed quantiles indicates how closely the models account for the shape and location of the observed RT distributions. We use horizontal displacement to provide an indication of how closely the models predict the proportion of correct responses. Predicted quantiles that sit close, both vertically and horizontally, to the black observed quantiles are models that fit the data well.

[5] It should be noted that the models were fit not to empirical quantiles, but to the entire RT distributions. We use quantiles to summarize the model fits due to the savings in space that they afford, but note that better agreement between predicted and observed quantiles would likely be possible using a quantile-based objective function.

It is evident from comparison between Fig. 3a and b that, consistent with the patterns in BIC values reported earlier, the serial model fits the data worse than do the parallel and global-familiarity models. Because the fit routine settled on different patterns of parameter estimates in trying to match the serial model's predictions to the data, the nature of the discrepancies differs across the participants. In general, however, for most participants, the serial model predicted a leading edge of the target RT distributions that was too fast for the small set sizes. In addition, excluding the very longest lag (i.e., the primacy serial position), the serial model tended to underpredict the degree of positive skewing at long lags for large set sizes. Overall, the global-familiarity and parallel models did a good job of predicting both the proportion of correct responses and the distributions of RTs for all combinations of set size and lag. The models capture the major trends in the quantiles—namely, that accuracy decreases and RT distributions become shifted and more skewed as study–probe lag increases. All of the architectures predict the shift in RT distributions as memory set size increases. The main failing for the global-familiarity and parallel models is that they predict too little skew for lure items, underpredicting the .9 quantiles, particularly for larger memory set sizes. This problem is magnified for the serial model.

To further characterize the RT distributions, we fit them with an ex-Gaussian distribution (cf. Heathcote, Popiel, & Mewhort, 1991; Hockley, 1984; Ratcliff & Murdock, 1976). The parameters of an ex-Gaussian distribution can be useful for describing the shape of RT distributions, including their leading edge and degree of positive skew. We report the detailed results of these analyses in the online supplement. The main issue that we investigated was the extent to which the models predicted accurately how the leading edge and positive skew of the RT distributions changed with increases in set size. Overall, the global-familiarity and parallel self-terminating models did a good job in these respects. By contrast, in various cases, the serial model predicted changes in the leading edge that were too large and tended to underpredict the degree of positive skew for larger set sizes.

Encoding-time hypothesis

In discussing challenges to the serial-exhaustive model, Sternberg (1975, p. 12) suggested that recency effects (i.e., effects of study–probe lag) may reside in other stages of information processing than the memory comparison stage. For example, he hypothesized that "the time needed to form an internal representation of the test stimulus in the encoding stage might depend on how . . . recently that stimulus had been presented." To test this hypothesis on the present data, we fitted an elaborated version of the serial-exhaustive model in which the location parameter of the log-normal distribution of base times (T_{POS}) was allowed to vary freely with lag. Note that because the log-normal shifts to the right and grows more positively skewed as T_{POS} increases, this mechanism could potentially allow the elaborated serial-exhaustive model to account for both effects of lag on the RT distributions. Although this added flexibility indeed improved the BIC fit of the serial-exhaustive model (see Table 1, row 4), it still performed far worse than did the parallel self-terminating and global-familiarity models. Its main failing is highly instructive and is the same as the one that we described earlier in this article. For any *fixed* lag, the serial-exhaustive model predicts a big effect of set size on RT (Fig. 2, row 4, column 2). That is, even if a recently presented test probe is encoded and compared very rapidly, the system must still scan exhaustively through all of the remaining mismatching items in the memory set. As can be seen from the observed data in Fig. 2 (column 2), however, although there is some separation between the different set size functions once one conditionalizes on lag, the separation is not nearly as great as predicted by the serial-exhaustive model.

Discussion

Our model-based analysis of the RT distribution data collected by Nosofsky et al. (2011) revealed an inadequacy of a serial-exhaustive architecture, whereas both the global-familiarity and parallel self-terminating architectures appear to be viable candidates. In addition, our use of multiple parameterizations led us to the firm conclusion that participants set different thresholds for responding on the basis of two factors: (1) the size of the memory set and (2) whether the evidence accumulated is for a positive or a negative match between study items and the probe. We also found that the rate at which evidence accumulates is driven primarily by the lag between study items and the probe, and perhaps also according to an overall influence of set size. Our model-based analysis failed, however, to distinguish between the global-familiarity and parallel self-terminating architectures of short-term memory scanning.

As was expected, the serial-exhaustive model predicted well the overall effect of set size on mean RTs. More interesting, unlike the baseline model considered by Sternberg (1966), the version that we tested was not constrained to predict flat serial-position curves. Nevertheless, one of its main failings is that it predicts big effects of set size even when one conditionalizes on lag of a positive probe, because the system must scan exhaustively through all the nonmatching items from the memory set. These predicted effects were much larger in magnitude than those observed in the data. One could account for the positive-probe data by assuming that LBA drift rates associated with (correct) matches are strongly influenced by lag, while the processes that lead to rejection of nonmatching items occur very fast. But then the model would be incapable of predicting the steeply increasing set size

functions for the lures (for a related analysis, see van Zandt & Townsend, 1993).

These limitations of the serial-exhaustive model arose even when we elaborated the model by allowing separate encoding-time distributions for each different lag of a positive probe. A related possibility to consider is that there are encoding difficulties during initial study, such that some proportion of memory set items are not encoded into memory at all. Such encoding failures may have occurred due to the fast presentation rates of the memory set items. In this case, participants may resort to guessing strategies on some proportion of trials, and a complete model must include this guessing component. Although we cannot rule out such a possibility, in our view this approach to saving the serial-exhaustive model seems strained. First, the hypothesis that participants often resorted to guessing suggests that error rates should have been quite high. Although there are occasional examples of high error rates for specific study–probe lags for participants 1 and 3, error rates were uniformly low for participant 2 and, especially, participant 4 (see Fig. 3). Second, the parallel self-terminating and global-familiarity models already do an excellent job of accounting for the data without the need to posit a guessing process.

As was noted in our introduction, Sternberg (1975) raised the possibility that the methodology used, especially the presentation rates of the memory set items and test probes, may lead to the use of alternative processes for memory scanning. Nosofsky et al.'s (2011) Experiment 2 used rapid presentation rates (500 ms) for study items and very little time between the final study item and the presentation of the probe (400 ms). In the experiment conducted by Sternberg (1966), for which the serial-exhaustive model was proposed, study items were presented for longer durations (1.2 s), and there was a 2-s break between the final study item and the probe. It is possible that participants may utilize global familiarity or parallel memory scanning when presentation rates are rapid but may use serial-exhaustive strategies when presentation rates are slower.

Therefore, we now present the results of a new experiment in which we attempt, as closely as possible, to replicate the conditions used by Sternberg (1966), but with enough trials to look at individual-participant RT distributions. If memory scanning is indeed accomplished via a serial-exhaustive process when presentation rates are slowed, we would expect that the architecture might perform as well as or better than the global-familiarity and parallel models.

Sternberg (1966) replication experiment

Method

Participants Three participants completed ten 1-h sessions on separate days. For each session, participants were reimbursed $9, with a $3 bonus for overall accuracy greater than 95%.

Stimuli The stimuli were the ten digits 0–9 (cf. Sternberg, 1966). Each digit was presented in the center of a computer monitor at a visual angle of approximately 3°.

Procedure Memory set size varied from one to five.[6] Each trial began with a fixation cross for 500 ms, followed by the individual presentation of each item in the study list for 1,000 ms, with a 200-ms break between successive study items. After the final study item, an asterisk was presented for 2,000 ms to signal that the next digit presented was the test probe. The probe then remained on screen until a response was made. Participants were asked to indicate whether the probe was a member of the study list (an old item, using the "F" key) or was not a member of the study list (a new item, using the "J" key). Feedback as to the accuracy of the response was then presented for 1,000 ms. After feedback, participants were asked to recall the entire study list, in order, by entering the corresponding digits using the keyboard (cf. Sternberg, 1966). Participants were then forced to take a 1,500-ms break, during which time they were to return their index fingers to the "F" and "J" response keys. Participants then pressed the space bar with either thumb to indicate that their index fingers were in the appropriate location, which prompted the start of the next trial.

The composition of the study list for each trial was made up of digits randomly sampled from the entire set of ten stimuli. The number of target and lure trials was equal, and their order was random. If the probe on a given trial was a target, the serial position of that item within the study list was chosen at random. In each block of 50 trials, there were 10 trials of each of the five study set sizes, and each session was composed of six blocks of trials. Each participant, therefore, completed a total of 3,000 trials, fewer than the approximately 5,000 trials that participants in the rapid-presentation experiment completed, but enough to consider full RT distributions (with an average of at least 60 observations in each serial-position—set-size combination).

Results

Mean and standard deviation of response times In Fig. 5, we show the mean and standard deviation of RTs as a function of set size and study–probe lag. Again, in general, mean RTs increase roughly linearly as a function of set size, and the set size functions for positive and negative probes

[6] Sternberg (1966) also tested six-item lists. To obtain suitable sample sizes for fitting our individual-condition RT distributions, we decided not to include the six-item lists.

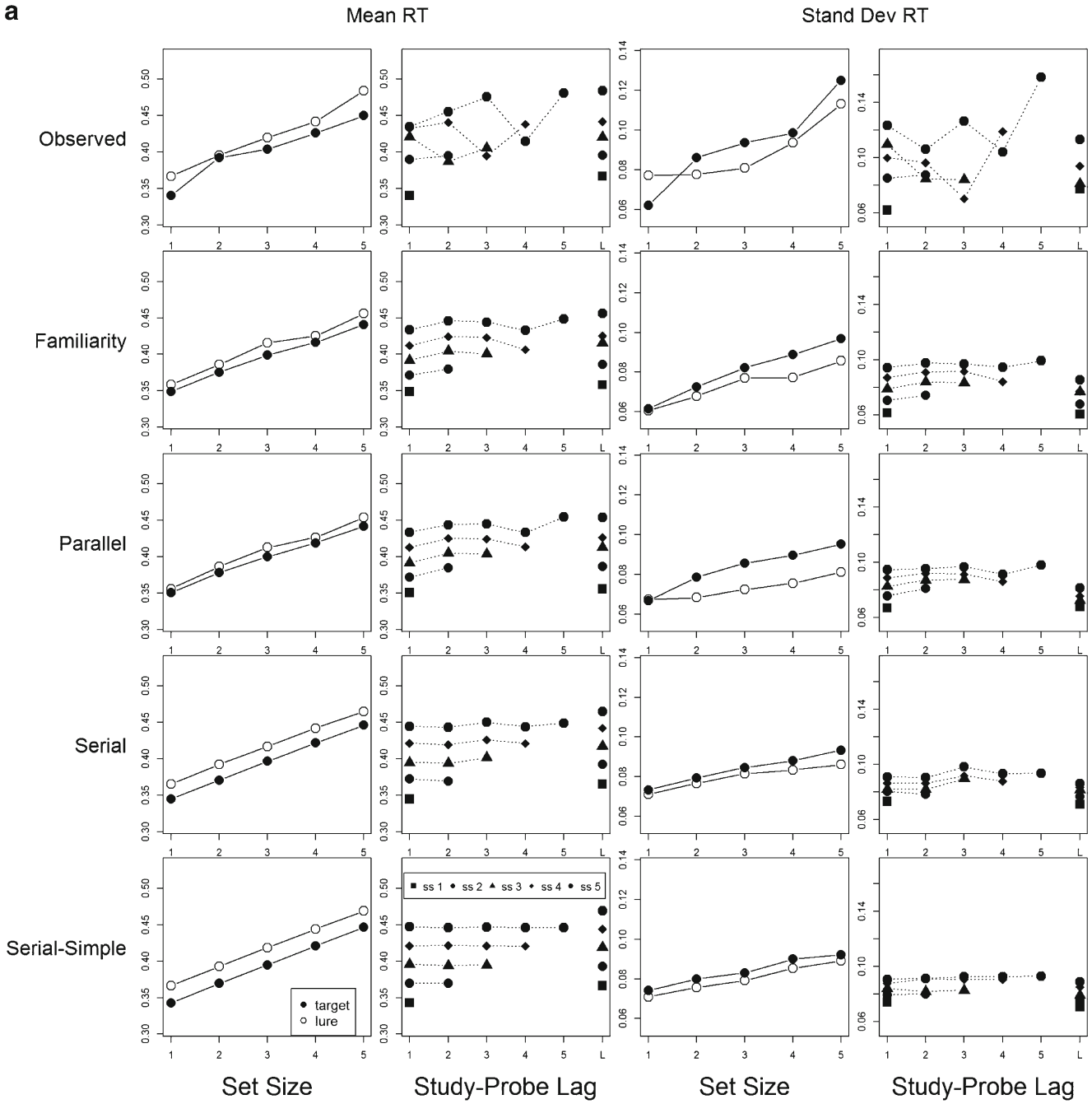

Fig. 5 Observed and predicted means and standard deviations of response times (RTs) for the Sternberg (1966) replication experiment. The format of the figure is almost identical to Fig. 2, with the exception that model predictions in row 5 come from a simple version of the serial -exhaustive model in which drift rates are fixed across study–probe lag and set size and response thresholds are fixed across set size

are roughly parallel to one another. The notable exception is for participant 7, whose lure RTs were flat for the first four set sizes and then slowed for the final set size.

The second column of Fig. 5 contains a plot of mean RTs for correct responses as a joint function of lag and set size. Most notable is that, for all 3 participants, the serial-position effects are much smaller than when presentation rates were faster. Indeed, in most cases, the functions are nearly flat. This result is particularly interesting given that there is still a large influence of memory set size on RTs. Whereas the set size effects in Fig. 2 were driven primarily by the presence of trials with larger lags, such is not the case for Fig. 5. Interestingly, we observe a similar pattern in standard deviations of RTs, such that they increase with set size (for participants 5 and 6) but remain relatively constant for different study–probe lags.

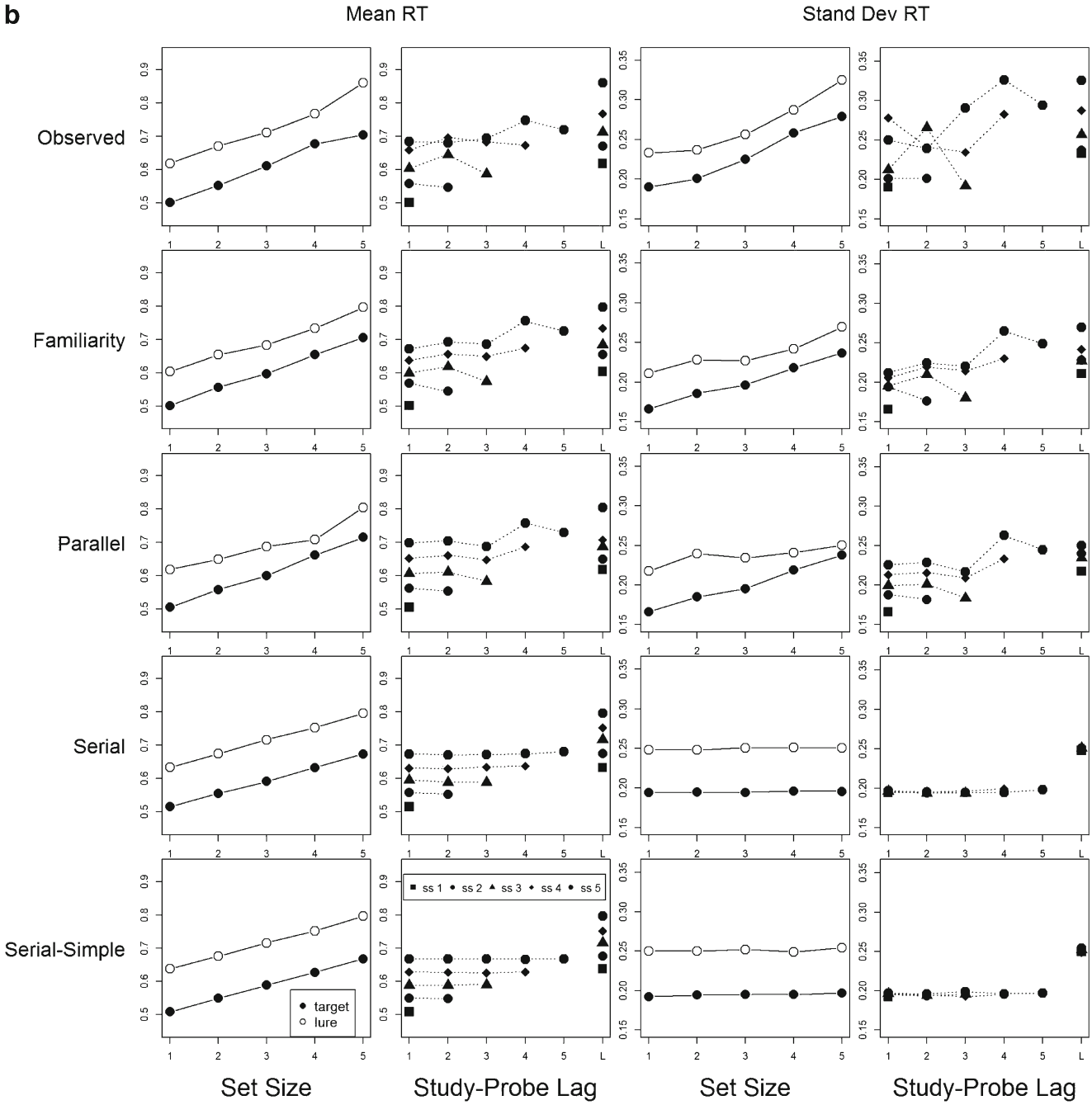

Fig. 5 (continued)

Response time distributions Figure 6 contains plots of observed quantiles for each combination of study set size and study–probe lag for targets and lure items for each of the 3 participants. Despite differences in the pattern of mean RTs, the shapes of the distributions in this slow presentation rate experiment look similar to those in the fast presentation rate experiment. RT distributions are positively skewed (as represented by the gradual increase in separation between increasing quantiles). In addition, as set size increases, the lengthening of RTs is associated with a shift in the entire distribution and with moderate increases in skew. Consistent with the lack of mean serial-position effects, however, the lag between study and probe has much less influence on the entire RT distributions under these slow presentation rate conditions. The proportion of correct responses was generally much higher in the present experiment, although we still generally observed slight decreases in accuracy with increasing lag between study item and probe.

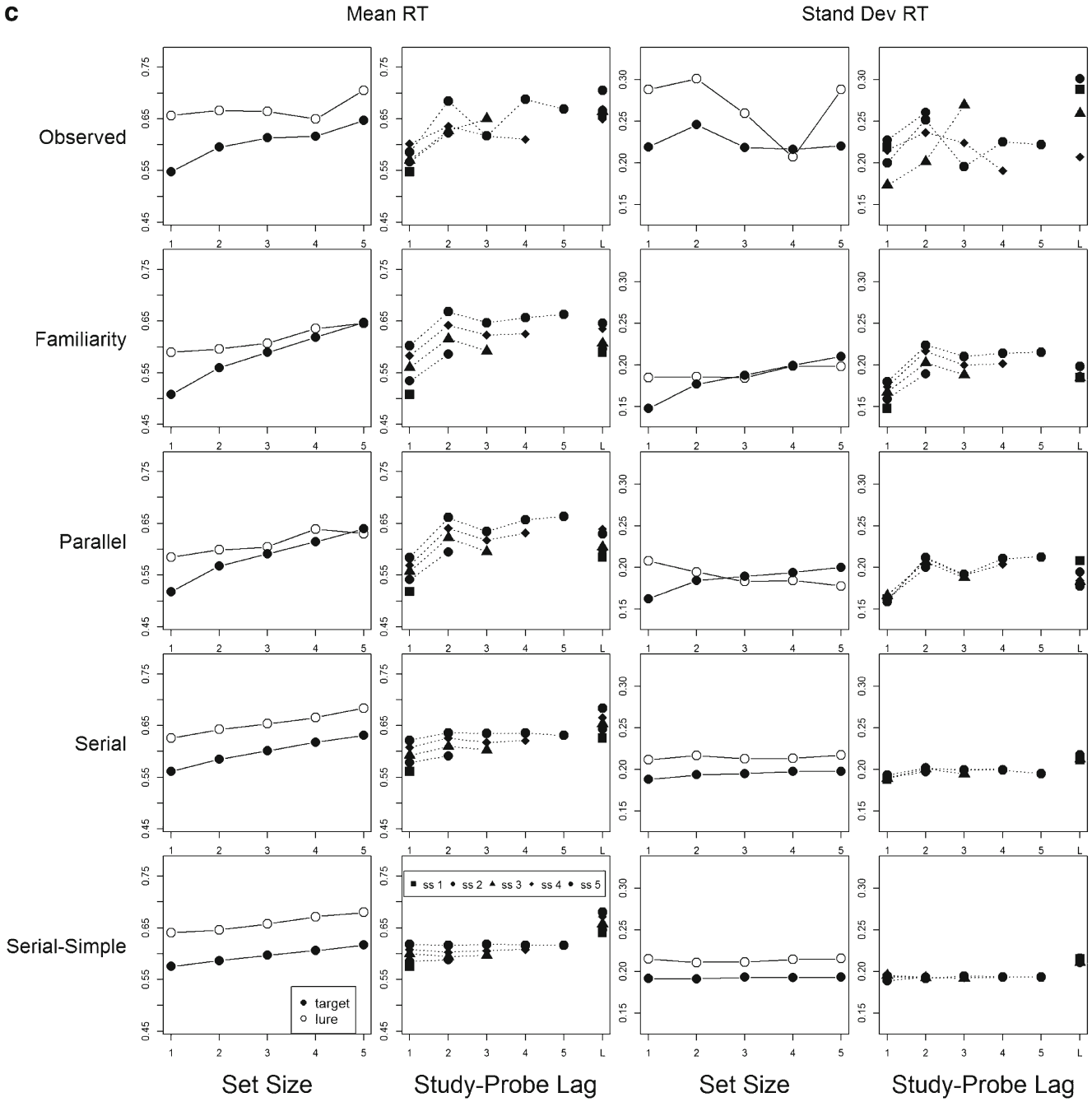

Fig. 5 (continued)

Preliminary discussion Many of the qualitative patterns in the data are consistent with the predictions from a serial-exhaustive model. For example, the model predicts the linearly increasing set size functions, while predicting little or no effect of study–probe lag on the mean and standard deviation of RTs. Thus, the different presentation rates may indeed influence the way that short-term memory is accessed. To test this hypothesis in greater depth, we now turn to the model-based analyses of the RT distribution data.

Model selection and best-fitting parameters

We again fit to the data all 36 of the previously outlined models. The fitting method and calculations of BIC were done in the same manner as for the fast presentation experiment. Perhaps most interesting is that the serial-exhaustive model is now very competitive with the other architectures. To compare the serial-exhaustive model with the other architectures, we again took the best-fitting models (according to BIC values calculated from analytic fits) from the

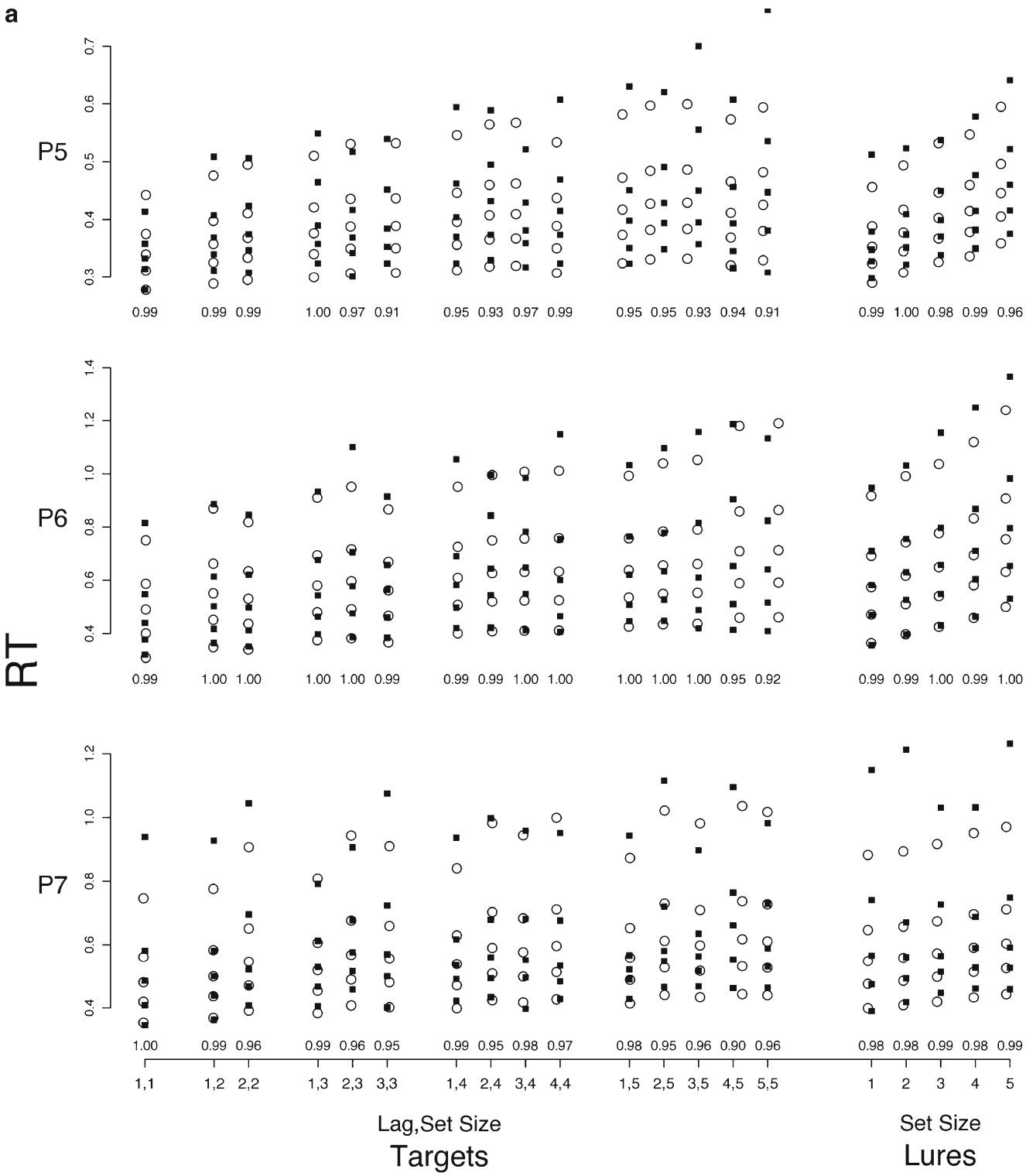

Fig. 6 Observed and predicted quantiles for the Sternberg (1966) replication experiment. The format of the figure is identical to Fig. 3

global-familiarity and parallel self-terminating architectures and fit them to binned RT data, using simulation. As is reported in Table 3, the best representative from the serial-exhaustive architecture had BIC values similar to those for the global-familiarity and parallel self-terminating models (i.e., ΔBIC values are as small for the serial-exhaustive model as they are for the other architectures).

ΔBIC values for the full set of parallel and global-familiarity models are plotted for each participant in Fig. 7. There is again relatively little difference between

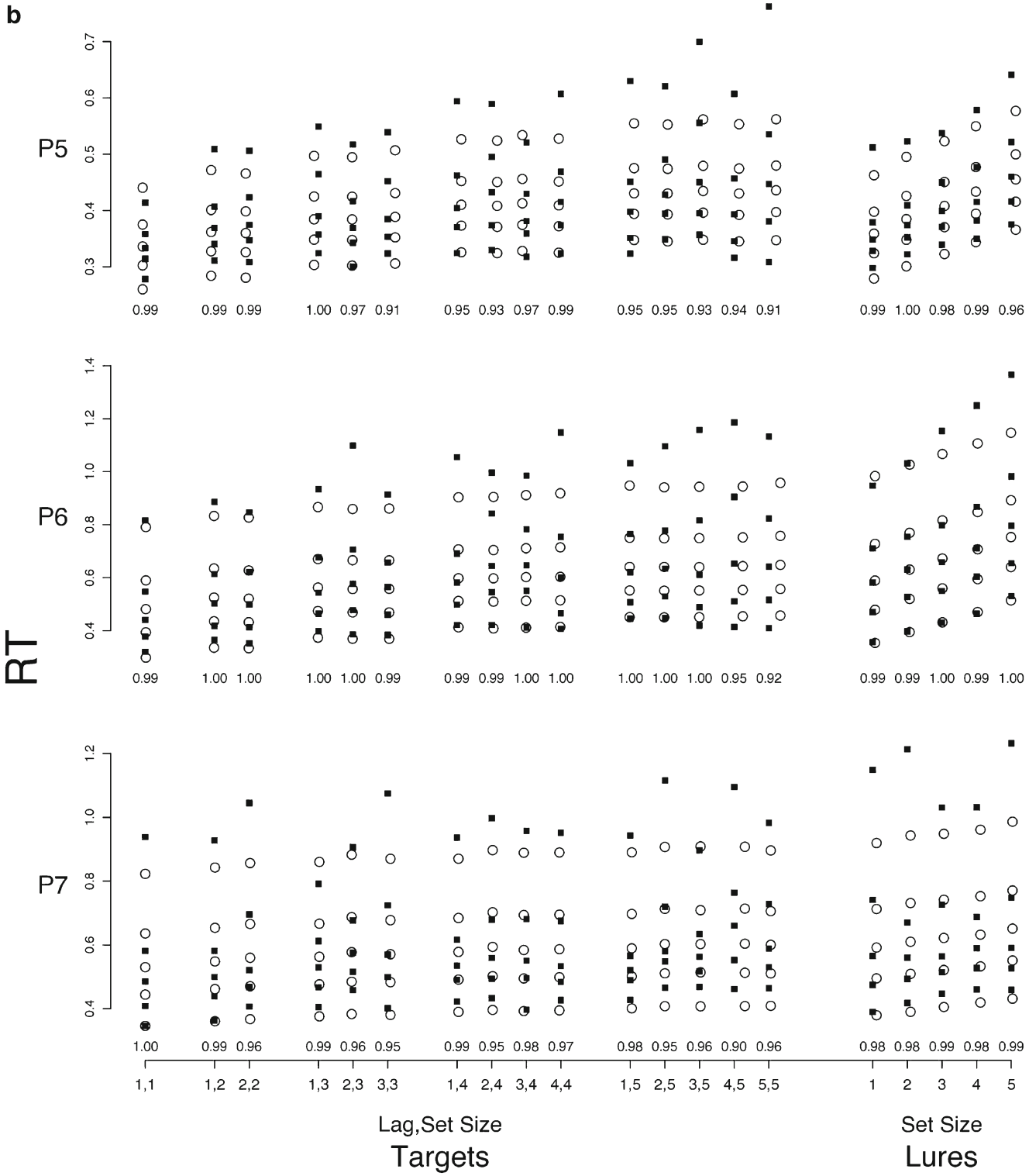

Fig. 6 (continued)

the best-fitting members of these two classes of models. Regarding the alternative parameterizations, the pattern is very similar to the one shown in Fig. 4. The best versions are those in which response thresholds change over evidence valence and study set size. Also, we again observe that drift rate need not vary freely across all unique combinations of serial position and set size. Instead, drift rate varies as a systematic function of these variables.

The best-fitting parameters for each of the architectures are reported in Table 4. For the global-familiarity and

Table 3 ΔBIC values based on simulation for each of the global-familiarity (GF), parallel self-terminating (PST), and serial-exhaustive (SE) architectures for each of the 3 participants in the slow presentation rate experiment

	ΔBIC			*k*
	P5	P6	P7	
GF	11	64	46	19
PST	0	88	0	19
SE	10	0	80	9[a]
SE w/enc	35	33	120	24

Note. For all architectures, we report the best-fitting version of the model and the number of parameters. We also include the ΔBIC values for the serial-exhaustive model that includes encoding time that varies with study–probe lag

[a] Note that this best-fitting version of the serial-exhaustive model has so few parameters because it assumes that response thresholds are fixed across set size and drift rate is fixed across lag and set size. Free parameters for this model are as follows: s, T_{POS}, T_{NEG}, S_T, A, b_{POS}, b_{NEG}, v, and u

parallel models, response thresholds again increase in magnitude with increases in set size. By contrast, according to the serial-exhaustive model, there is zero increase in the magnitude of the response thresholds. The main change in the pattern of parameter estimates, as compared with the fast presentation rate experiment, is in the estimated drift rates. With the possible exception of lag 1 for participant 7, the estimated drift rates are roughly flat across the different study–probe lags according to all of the models.

Qualitative model predictions

Mean and variance of response times Rows 2–5 of Fig. 5 contain the predictions for means and standard deviations of RT from the global-familiarity, parallel self-terminating, and the two versions of the serial-exhaustive models. The second serial model is a constrained version in which drift rates are held constant across serial position and set size and response thresholds are held constant across set size. The global-familiarity and parallel models do a good job of accounting for the qualitative patterns in mean and standard deviations in RTs. Most important, the serial model now also does a good job of accounting for these data.

Predicted quantiles Figure 6a contains the predicted quantiles from the global-familiarity model, and Fig. 6b contains predictions from the serial-exhaustive model. Again, the predictions from the parallel self-terminating model were very similar to those of the global-familiarity model and, so, appear in the online supplement. The parallel and global-familiarity models continue to do a good job of accounting for both the speed and accuracy of responses for both target and lure items. Now, however, the serial-exhaustive model also appears to do a relatively good job of explaining the shapes of the RT distributions.

The procedure for our Sternberg replication experiment was similar to one used by Hockley (1984), who also used slow presentation rates. A main theme of his investigation was to analyze his RT distribution data by fitting the ex-Gaussian distribution to them and examining how the ex-Gaussian parameters changed as a function of set size. Among the differences in procedure was that our participants participated in far more sessions than did Hockley's. In addition, Hockley did not require participants to serially recall the memory set items following their probe recognition judgments. We report ex-Gaussian analyses for our data in the online supplement. In brief, we observed an increase in τ (i.e., positive skew of the RT distributions) with set size that was consistent with Hockley's results. However, we also observed an increase in μ (i.e., the leading edge of the distributions) with set size that was much larger than what Hockley observed. Possibly, participants engaged in serial search under our experimental conditions but did not do so under Hockley's conditions. Alternatively, due to their greater experience in the task, participants in our experiment may have learned to adjust response thresholds for trials involving differing memory set sizes.

Discussion

Participants appeared to behave differently in our replication of Sternberg (1966) than in Nosofsky et al.'s (2011) Experiment 2. One of the main differences was that, under these slow presentation rate conditions, there was essentially no effect of study–probe lag. This difference had a dramatic influence on the results of our model-based analysis, since the serial-exhaustive architecture was now able to account for the data as well as the global-familiarity and parallel self-terminating architectures. Indeed, even a highly simplified version of the serial-exhaustive model, in which drift rate for targets and lures was held fixed across all set size and lag conditions and in which response thresholds were held fixed as well, provided a good account of the data. We consider the comparisons among the models in greater depth in our General Discussion section.

General discussion

In this research, we formulated alternative information-processing architectures designed to account for detailed RT distribution data and error rates observed in the classic Sternberg (1966) memory-scanning paradigm. Our approach was to assume that the individual-item comparisons (or single global-familiarity comparison) that take place in each model architecture are governed by LBA processes.

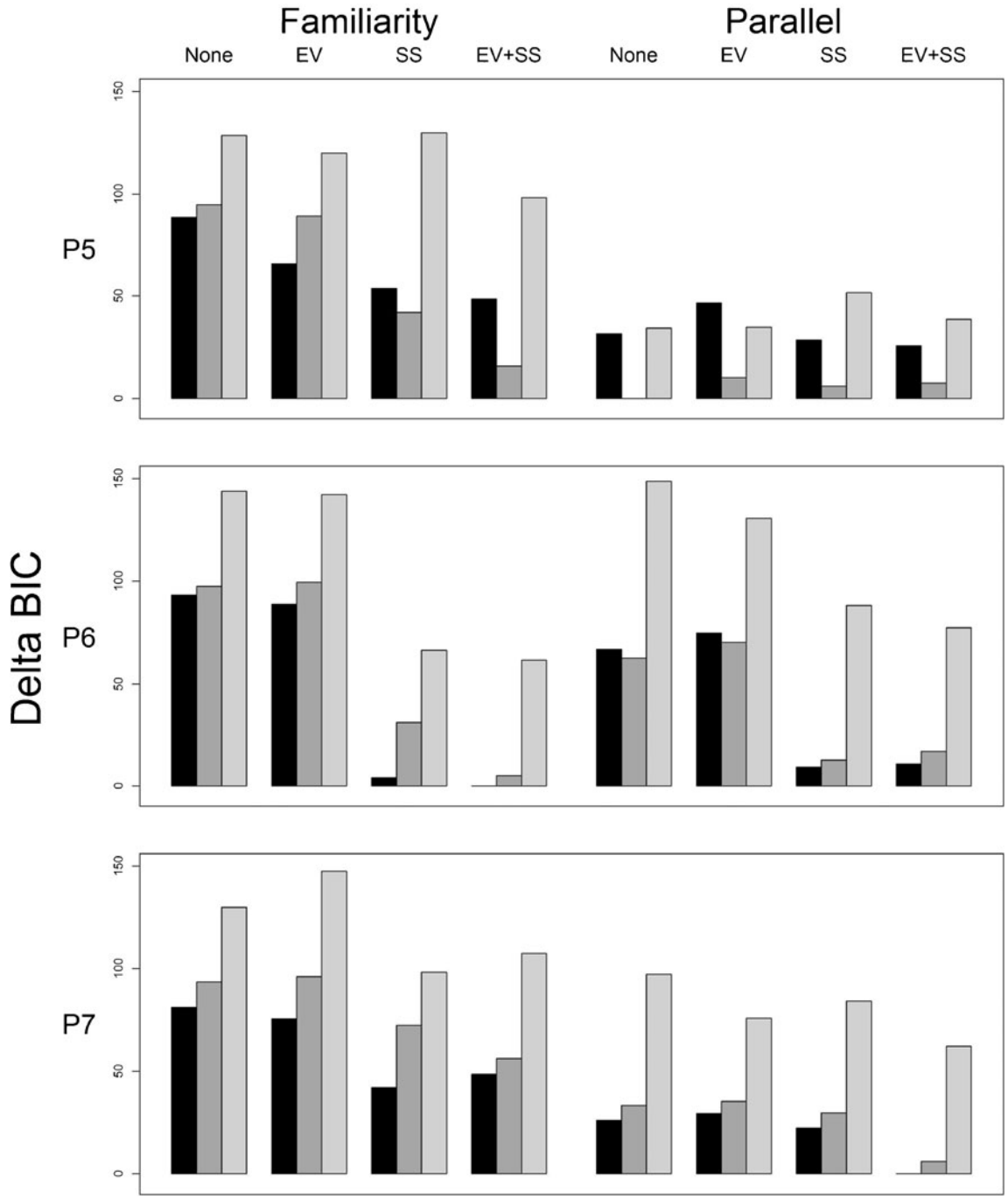

Fig. 7 ΔBIC values are plotted for each of the global-familiarity and parallel self-terminating models fit to each of the 3 participants in the Sternberg (1966) replication experiment. As in Fig. 4, EV and SS refer to evidence valence and set size. Recall that smaller values of ΔBIC indicate that a model is more parsimonious

Because the LBA approach has fared successfully in the modeling of RT distribution data associated with simple, elementary decisions, it is a reasonable building block for the more complex model architectures investigated here. By placing reasonable constraints on the form of the elementary processes that govern individual-item comparisons, we were able to test the alternative architectures on their ability to predict the composite RT distributions associated with search of the entire memory sets. This approach also provided principled predictions of error rates and speed–accuracy trade-offs observed in the memory-scanning task. Finally, we used a dual-route attack, in which we examined a variety of qualitative predictions from the competing models, along with their quantitative fits to the complete RT distributions.

Motivated by an old hypothesis advanced by Sternberg (1975), we tested the models in two main experiments that

Table 4 Parameter estimates for each of the global-familiarity (GF), parallel self-terminating (PST), and serial-exhaustive (SE) architectures for each of the 3 participants in the slow presentation rate experiment

	P5			P6			P7		
	GF	PST	SE	GF	PST	SE	GF	PST	SE
s	0.17	0.22	0.86	0.32	0.36	0.31	0.34	0.39	1.15
T_{POS}	0.1	0.12	0.30	0.15	0.19	0.43	0.16	0.19	0.50
T_{NEG}			0.34			0.55			0.59
S_T			0.001			0.002			0.001
A	0.08	0.12	0.05	0.55	0.70	0.05	0.26	0.37	0
b_{POS}	0.24	0.26	0.02	0.69	0.84	0.03	0.47	0.55	0.07
b_{NEG}	0.26	0.27	0.02	0.75	0.81	0.02	0.51	0.55	0.04
Δb_{POS}	0.001	0.008	0	0.04	0.04	0	0.02	0.03	0
Δb_{NEG}	0.01	0	0	0.04	0.004	0	0.02	0	0
V_1	0.80	0.92	1.70	1.15	1.42	1.25	1.04	1.17	2.85
V_2	0.77	0.87	1.62	1.12	1.39	1.24	0.91	0.99	2.15
V_3	0.77	0.87	1.51	1.13	1.43	1.31	0.95	1.04	2.30
V_4	0.81	0.92	1.69	1.00	1.27	1.24	0.93	1.00	2.24
V_5	0.75	0.83	1.52	0.98	1.30	1.11	0.91	0.96	2.38
U_1	0.85	0.93	2.37	1.09	1.21	1.33	0.96	0.99	3.86
U_2	0.81	1.10	2.15	1.07	1.35	1.23	0.98	1.63	3.72
U_3	0.77	1.11	2.29	1.09	1.87	1.36	0.99	1.56	3.66
U_4	0.78	1.09	2.29	1.07	2.11	1.35	0.97	1.17	3.66
U_5	0.73	1.15	2.43	1.04	1.41	1.28	0.99	1.10	3.82
P	1.01	1	1	1.08	1.04	1	1.01	1.02	1
ΔS	−0.05	−0.03	−0.003	0	−0.02	0	0	0.02	−0.01

Note. The parameter estimates are the maximum likelihood estimates based on fits of the model in which response thresholds changed across study set size and evidence valence and drift rate changes systematically across study–probe lag and set size

varied the presentation rates of the memory set items and the test probes. In line with our expectations, we found support for familiarity-based or parallel-access models in an experiment in which rapid presentation rates were used. By contrast, in an experiment in which presentation rates were slow, a serial-exhaustive architecture yielded accounts of the data that were as good as those of the global-familiarity and parallel-access models. As is discussed in more detail below, one possibility is that a single architecture governs performance across the fast presentation rate and slow presentation rate conditions, with systematic changes in parameter settings occurring across these different conditions. An alternative possibility is that the memory-scanning architecture changes depending on the presentation rate conditions, becoming serial exhaustive in character when presentation rates are slowed. In our view, each alternative interpretation has certain advantages and disadvantages.

Fast presentation rate experiment

In the fast presentation rate experiment, strong evidence was found favoring either the parallel self-terminating or global-familiarity architectures, as compared with the serial-exhaustive one. The serial-exhaustive architecture, even when elaborated to allow encoding-time differences due to recency, was unable to capture the joint finding of strong serial position effects in the mean RTs combined with the roughly parallel and steeply linearly increasing set size functions observed for positive and negative probes. Also, as memory set size increased, the serial model often predicted changes in the leading edge of the RT distributions that were too large, while underpredicting the increases in positive skew of the RT distributions (see the online supplement). Both the parallel self-terminating and global-familiarity architectures did a reasonably good job of capturing all of these aspects of the mean RT and RT distribution data.

Within the framework of the parallel and global-familiarity architectures, our detailed parametric investigations also suggested two main effects of the variables of set size and serial position. First, as the lag between a positive probe and a matching memory set item increased, the drift rate associated with the LBA match process decreased. Such an effect is consistent with the idea that the memory strength of a study item decreases with its lag of presentation. Interestingly, once one took lag into account, there was no evidence that increases in set size per se led to any further decrease in drift rates. A common metaphor from the information-processing literature is that memory search may involve a limited-capacity parallel process in which total capacity is "shared" among the different memory set items (e.g., Townsend & Ashby, 1983, p. 14). As set size

increases, less capacity can be devoted to any individual item. Our detailed modeling of our RT distribution data, however, lent no support to that hypothesis. Instead, the slowdowns in RT with set size were due to increasing lag and to changes in response thresholds, which we discuss next.

The second major effect was that, as memory set size increased, participants increased their response thresholds for making match/mismatch decisions. The basis for this effect remains as a topic for future investigation. One possibility is that it reflects the cognitive system's manner of dealing with increases in noise associated with larger set sizes. For example, as memory set size increases, there are more opportunities for an observer to false alarm to lures. Within the parallel self-terminating architecture, if any individual-item comparison yields a false alarm, the final response will be a false alarm; therefore, as set size grows, overall false alarm probability will increase. Likewise, as set size grows, global familiarity on lure trials would also be expected to increase (Kahana & Sekuler, 2002; Nosofsky et al., 2011). Thus, increasing the magnitude of the response thresholds may be the system's way of holding down the false alarm rate, because it requires that more evidence be gathered before decisions are made.

Slow presentation rate experiment

In the case in which presentation rates were slowed, the serial-exhaustive model yielded quantitative fits to the detailed RT distribution data that were as good as those of the parallel and global-familiarity models. One of the major qualitative changes across the fast presentation rate and slow presentation rate experiments was that the serial position effects largely disappeared under the latter conditions. In the parallel and global-familiarity models, this change was modeled in terms of a different pattern of drift rate parameters. Whereas drift rate decreased with increasing lag in the fast rate experiment, the measured drift rates were essentially flat in the slow rate experiment. There are a couple of alternative explanations for such a result. First, if memory strength decreases, at least in part, with the simple passage of time (but see Lewandowsky & Oberaurer, 2009) and if the strengths asymptote at some low level, one would expect to see much reduced lag-based differences in drift rates under slow presentation conditions than under fast presentation conditions. Alternatively, items may have entered a longer-term store outside of short-term memory, where they become less susceptible to the effect of lag. Second, in cases in which presentation rates are slowed, it is much more likely that observers engage in a variety of rehearsal strategies. If so, then the memory strengths and associated drift rates will no longer be a simple function of lag. Instead, they will depend on the specific rehearsal strategy that is used. Furthermore, if different rehearsal strategies are used on different trials, the patterning of memory strengths may be quite complex. A set of roughly flat or unsystematic drift rates may be a reflection of a confluence of different rehearsal strategies. To test this idea, future research might bring rehearsal strategies into control by providing explicit instructions on which strategy to use or by asking participants to rehearse overtly. Alternatively, it may be possible to fit mixture models to the data that make explicit hypotheses about the alternative rehearsal strategies that participants use across trials.

An alternative interpretation of our results is that a mixture of mental architectures may underlie short-term memory scanning, and there was a major shift toward greater use of serial-exhaustive processing in the slow presentation rate condition. Sternberg (1975, pp. 13–14) suggested various factors that may lead to mixed strategies. In the present case, perhaps fast presentation rate conditions give rise to large differences in familiarity between positive and negative probes, so relying on global familiarity becomes a more efficient strategy than serial-exhaustive search. But under slow presentation rate conditions, those familiarity differences may begin to disappear.

Perhaps the main argument in favor of this interpretation is that most of the fundamental qualitative effects in the data were those predicted a priori by a simple, baseline version of the serial-exhaustive model. These effects include the roughly parallel and linearly increasing set size functions observed for positive and negative probes, the nearly flat serial position curves observed at each set size, the increase in the leading edge of the RT distributions as set size increased, and the fairly moderate increases in positive skew with increases in set size (as compared with Experiment 1). Although these effects can be captured by versions of the parallel self-terminating and global-familiarity models, they hold only under certain types of parametric assumptions. Future work is therefore needed to understand the overall flexibility of the parallel self-terminating and global-familiarity models. In addition, future work is needed to determine whether the parametric assumptions that yield good fits of those models to the slow presentation rate data may have some deeper theoretical basis. If the models capture the qualitative effects in the data only with a restricted set of unmotivated parameter settings, their account of those data must be brought into question.[7]

Is it plausible that the memory-scanning architecture itself might change across these experimental conditions?

[7] On the other hand, in order for the serial-exhaustive model to account for the full set of RT distributions in the slow presentation rate experiment, we found that we had to make two assumptions that were not necessary for the global-familiarity and parallel self-terminating architectures. Both were related to nondecision time. The first was that nondecision time was different depending on whether the probe item was a target or a lure. This assumption was fundamental in Sternberg's (1966) pioneering work as well. The second additional assumption was that nondecision time varied from trial to trial according to a log-normal distribution (whereas a constant nondecision time was assumed in the parallel and global-familiarity models). Although this aspect of the serial-exhaustive model is more complex than for the other models, in our view the assumption that there is variability associated with nondecision time is extremely plausible and hardly requires justification.

Some precedence for such a possibility can be found, for example, from earlier investigations conducted by McElree and Dosher (1989, 1993). As was noted earlier, in their investigation of short-term recognition (using fast presentation rates), McElree and Dosher (1989) obtained evidence consistent with predictions from parallel or direct-access models. However, in a closely related study in which participants were required to make judgments of recency, McElree and Dosher (1993) found evidence that order information is retrieved by a serial-retrieval mechanism. Admittedly, the task goals differed across their two studies, whereas in our investigation the primary task goal always involved recognition. [8] Nevertheless, in our view, McElree and Dosher's (1989, 1993) findings lend support to the possibility that the mental architectures that govern access to short-term memory are not fixed but might change across experimental conditions.

Other future directions

Finally, future work should also aim to distinguish between the global-familiarity and parallel self-terminating models of memory scanning. We were unable to derive any focused qualitative contrasts for distinguishing those models. Moreover, our failure to discriminate between the two architectures on quantitative grounds is reflective of how well these two architectures are able to mimic each other, even at the level of full RT distributions. This mimicry is especially likely to happen when the models are given the parametric freedom required to explain the 40 different RT distributions and response proportions that served as our test bed.

Regarding this mimicry, we should note that in the present investigation, our focus was on the classic version of the Sternberg paradigm in which highly discriminable memory set items were used and accuracy was at very high levels. The versions of the models that we formulated had these types of testing conditions in mind. However, in recent years, an interesting theme has been to examine performance under conditions in which similarity relations among memory set items and test probes are manipulated. Pursuing this avenue in combination with the collection of detailed RT distribution data might provide more highly diagnostic information for telling apart the alternative architectures.

We should emphasize that one recently formulated member of the class of global-familiarity models—namely, the EBRW model (Nosofsky et al., 2011)—has shown initial success in accounting for performance in such experiments. For example, following a paradigm introduced by Kahana and Sekuler (2002), Nosofsky et al. conducted an extended version of the Sternberg task in which the memory set items and test probes were embedded in a continuous multidimensional similarity space (i.e., colors varying in hue, brightness, and saturation). In addition to varying memory set size and lag of positive probes, Nosofsky et al.'s design sampled broadly from the similarity space, creating 360 unique lists with varying set size, lag, and similarity structure. The goal was to use the EBRW model to predict jointly both the accuracies for the individual lists (which varied considerably due to the confusability of the stimuli) and the individual-list mean RTs. Furthermore, rather than allowing the random-walk drift rates for the individual lists to vary as free parameters, the drift rates were instead computed from the summed similarity of the test probes to the specific memory set items (for details, see Nosofsky et al., 2011, Experiment 1). Despite using relatively few free parameters, the model provided excellent overall quantitative accounts of the choice probabilities and mean RTs associated with the individual lists and captured fundamental qualitative effects involving the roles of set size, lag, and similarity.

Furthermore, the EBRW modeling conducted by Nosofsky et al. (2011) seems broadly consistent with recent results involving similarity effects for lures in short-term recognition. For example, Johns and Mewhort (2011) conducted experiments in which the lag of target items was manipulated, but also in which the lag of lures was manipulated. Specifically, lures were presented that were highly similar to memory set items from specific serial positions. With regard to target items, Johns and Mewhort observed the common result that more recently presented targets had shorter RTs than did less recently presented targets (except for a small primacy effect). This finding is consistent with the idea that more recently presented targets have greater memory strengths (cf. Donkin & Nosofsky, in press; Nosofsky et al., 2011). The novel and more interesting finding was that the serial position function for the lures showed the same pattern: Lures that were highly similar to more recently presented memory set items had shorter correct rejection RTs than did lures that were highly similar to less recent memory set items. Although some versions of global-familiarity models might predict the reverse result (see Johns & Mewhort, 2011), such is not the case for the EBRW model. In particular, a fundamental component of Nosofsky et al.'s EBRW modeling involved the assumption that whereas memory strength of study list items decreased with lag, so did memorial *sensitivity*—that is, the ability to discriminate between memory set items and high-similarity lures. (Intuitively, it is easier to discriminate between two close shades of the same color if the first was just recently presented, rather than presented in the distant past.) The precise predictions from the EBRW model would depend on detailed parameter settings from the model and the levels

[8] As part of our direct replication of Sternberg (1966), we did require participants to recall the entire sequence of items presented. A future experiment, in which slow presentation rates are used without the necessity to recall the entire set of items (or fast presentation rates that do require full recall) may help disentangle this issue.

of manipulated similarity in the experiment, but the general result observed by Johns and Mewhort seems within the model's scope. In the same way that the EBRW model has been formalized to account for similarity effects in short-term recognition, it seems likely that analogous extensions could be made to members of the class of parallel self-terminating and serial-exhaustive models. It is an open question whether such models could also capture the types of similarity effects reviewed briefly above.

Still another approach to contrasting the models may be to make use of the *systems factorial technology*, a still-developing set of interrelated RT methods for distinguishing among alternative mental architectures (e.g., Schweickert, 1992; Townsend & Nozawa, 1995; Townsend & Wenger, 2004). These methods involve the study of RT distribution data in conditions in which similarities and factorial structure among a set of stimuli are very carefully calibrated. Such techniques have been used recently, for example, to distinguish among a variety of different mental architectures of rule-based classification performance (Fific et al., 2010; Little, Nosofsky, & Denton, 2011), and preliminary attempts have been made to use the methods in the domain of memory search with highly specialized types of stimuli (Townsend & Fific, 2004). Combining these techniques with the present RT distribution approaches may yield still greater power in allowing one to distinguish among the alternative models of short-term memory search.

Appendix

Model parameterizations

Response threshold parameterizations

We consider four different response threshold parameterizations. In the first, we used just one response threshold for all responses and set sizes and, so, required just one response threshold parameter, b. In the second version, we estimated separate response thresholds for accumulators collecting evidence for either a positive or a negative match between probe and study items. This parameterization required two threshold parameters, b_{POS} and b_{NEG}, for positive and negative matches between study items and probes, respectively. We also fit a parameterization of the model in which response threshold changed with study set size and, so, required an additional slope parameter Δb for the linear relationship between study set size and threshold, such that response threshold for both positive and negative matches in set size i was determined by $b + \Delta b(i - 1)$. In the fourth and final version, we allowed response thresholds to differ both for positive and negative match accumulators and also across set sizes. This version required four parameters in total, b_{POS} and b_{NEG}, the response thresholds for set size one, and Δb_{POS} and Δb_{NEG}, the linear slope parameters for each of the response thresholds.

Drift rate parameterizations

We first consider a parameterization in which the rate of evidence accumulation depends only on the lag between when the item is presented during study and when the item is later probed. For parallel and serial architectures, in which the match between each item in the study set and the probe item is assessed separately, the rate of evidence accumulation for each item is determined by the lag between the study item and probe. In particular, we assume that the evidence accumulation rate for a positive evidence match between probe and study items depends on the lag between when that item was studied and when it was subsequently probed. Similarly, we assume that the rate of negative evidence accumulation when a probe item is compared with a study item that it does not match will also depend on study–probe lag. That is, we expect that more recently presented items will be easier to identify as not matching the probe (i.e., more negative evidence). In other words, we expect that the lag between study and probe will facilitate the time taken to identify both a positive and a negative match between study and probe items. We estimate, therefore, separately for each study–probe lag, i, a rate of accumulation parameter for both a positive (v_i) and a negative (u_i) match between study and probe items, yielding a total of ten drift rate parameters.

An example may help make this clearer. Imagine a trial on which the study list was made up of the items "7," "4," and "6," presented in that order (which we will represent as [7, 4, 6]) and a probe item "4" is presented. There are three accumulators collecting positive evidence, one for each study item, and three accumulators accumulating negative evidence for a match between each study item and the probe. Consider first just the accumulators associated with the correct response—that is, the accumulator collecting positive evidence for the study item "4" and those collecting negative evidence for the study items "7" and "6." The rate at which positive evidence is accumulated for the match between the memory for "4" and the probe item is estimated as v_2, because the lag between "4" and the probe item is two. The negative evidence for the match between "6" and the probe would accumulate at rate u_1, and the negative evidence for the match between "7" and the probe would accumulate at rate u_3. The rate at which evidence accumulates for the incorrect responses will be discussed later.

Evidence accumulation rates are estimated differently in global-familiarity architectures, because they are based on the combined influence of all items in memory. We assume that when there is a match between the probe and any of the study items, the lag between the study and probe item, i,

determines the rate of evidence accumulation for a positive match between the entire contents of memory and the probe, v_i. When the probe item does not match any item in memory, we estimate the rate of evidence accumulation, u_j, based on the number of items in memory, j. This parameterization may seem unintuitive at first but is a result of our reluctance to place a constraint on the way that information from the items in memory is combined. For example, consider a trial on which only one item is presented and the probe does not match this item. We would estimate the rate of negative evidence accumulation to be u_1. If we now have a trial on which two items are present, neither of which is the probe item, the lack of match between two items is combined in some way to indicate that the probe is a lure. It might be possible that this information is combined in a functional way; for example, the rate for the most recent item, u_1, might be multiplied by the rate for the next most recent item, u_2, allowing us to estimate a drift rate for negative match that is directly related to study–probe lag. Because we do not wish to place such a constraint on how the negative match between the probe and each item in memory is combined, however, our only option is to estimate a separate rate of accumulation for the case in which two items are present, and more generally, depending on the number of items in memory. So, finally, consider again a trial on which the study list was [7, 4, 6]. If the probe was "6," the rate of positive evidence would be v_1; if the probe was "7," the rate of positive evidence would accumulate at rate v_3; and if the probe was a new item (i.e., not a member of [7, 4, 6]), the rate of negative evidence accumulation would be u_3.

Beyond study–probe lag, there is also considerable evidence for a primacy effect in short-term memory recognition; the first item in the list is often more easily recognized than other items presented early in the study list. We decided, therefore, to include in all of our parameterizations, a primacy parameter, P, by which we multiplied the lag-based drift rate parameter when it was also the first item in the study list. For example, say that the probe on a given trial was "7" and the study list was [7, 4, 6]. In all architectures, the rate of evidence for a positive response would be $v_3 \times P$. If, however, the probed item was "2," then in the parallel and serial architectures, the influence of primacy would be on the negative evidence for the item "7," turning it into $u_3 \times P$, rather than just u_3. For the sake of simplicity, because there are many possible ways in which primacy may influence the rate of negative evidence accumulation in a global-familiarity architecture, we assume that primacy has no influence on the rate of negative evidence accumulation (or on positive evidence accumulation when the probe does not match any study item). Note that the primacy parameter is not redundant, because the same lag is present in multiple set sizes; for example, a lag of 3 in a list of set size 3 would be subjected to the primacy multiplier, but would not in set sizes 4 and 5.

The second drift rate parameterization we considered assumed that in addition to lag and primacy, the evidence accumulation rate for a match between study and probe items is influenced by the number of items in the study list. For example, consider a study list [7, 4, 6] in which "4" is presented as the probe. In the parallel and serial architectures, the accumulation rate of positive evidence for a match between the second item and the probe is estimated as $v_2 \times S_3$, where S_3 refers to the additional influence on drift rate from being an item in a list of length three. The accumulation rate for a negative match between the other items will be estimated as $u_1 \times S_3$, and $u_3 \times S_3 \times P$ for "6" and "7," respectively. In the global-familiarity architecture, S_i will influence only the drift rates for positive matches (v_i), and not negative matches, because those parameters are already estimated separately for different set sizes. So, for the study list [7, 4, 6] with "4" as the probe, the drift rate for a positive match will be $v_2 \times S_3$. For modeling purposes, we assumed that the change in drift rate with study list length would be linear and set S_1 to 1 and estimated a single parameter ΔS to govern the change in the list length multiplier, S, as the study list length increased. So the list length multiplier for a study list of length i was $1 + \Delta S(i - 1)$.

In the third drift rate parameterization, we estimated a rate of accumulation of positive evidence for each of the possible study–probe lags within each set size. In other words, we freely estimated a drift rate parameter for all 15 combinations of set size and serial position. We also estimated six free parameters for the rate of negative evidence accumulation when the study item did not match the probe (five values of u and one ΔS). These parameters are identical to the parameters estimated in the second drift rate parameterization, and so for the parallel and serial architectures, the drift rate for negative evidence depends on both the study–probe lag and an overall influence of study set size. For the global-familiarity model, the drift rate for negative evidence depends on study set size. So, given the study list [7, 4, 6] and a probe of [4], in a parallel and serial architecture, the drift rates for a positive match between the probe and the representation for the study item "4" would be $v_{3.2}$, where the 3 refers to the study set size and the 2 refers to the lag between study and probe item, and the negative match between the probe and the other study items would be $u_3 \times S_3$ and $u_1 \times S_3$ for items "7" and "6," respectively. For the global-familiarity model, the positive evidence would simply accumulate at rate given by $v_{3.2}$ if the probe matches an item in the study list and at rate u_3 if the probe does not match an item in the study list.

So far, we have discussed estimation of drift rate parameters for only correct decisions—that is, when an accumulator is collecting positive evidence for a matching probe or negative evidence for a mismatching probe. We chose to fix the mean of the drift rate distribution for any accumulator

associated with what would be an incorrect decision to be one minus the mean drift rate for the correct decisions. This constraint is standard in multiple accumulator models of choice RTs (e.g., Brown & Heathcote, 2005, 2008; Usher & McClelland, 2001) and is used to satisfy a scaling property when fitting RT models (but see Donkin, Brown, & Heathcote, 2009, for a discussion of alternatives). So, for example, if the study list was [3, 8, 1] and the probe was an "8," then in the parallel and serial architectures, the negative evidence for a match between the memory for "8" and the probe accumulates at the rate $1 - v_2$, and the positive evidence for a match between "3" and "1" items and the probe accumulates at rate $1 - u_3 \times P$ and $1 - u_1$, respectively.

References

Ashby, F. G., Tein, J.-Y., & Balakrishnan, J. D. (1993). Response time distributions in memory scanning. *Journal of Mathematical Psychology, 37,* 526–555.

Brown, S., & Heathcote, A. (2005). A ballistic model of choice response time. *Psychological Review, 112,* 117–128.

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

Corballis, M. C. (1967). Serial order in recognition and recall. *Journal of Experimental Psychology, 74,* 99–105.

Donkin, C., Brown, S., & Heathcote, A. (2009). The over-constraint of response time models: Rethinking the scaling problem. *Psychonomic Bulletin & Review, 16,* 1129–1135.

Donkin, C., Brown, S., & Heathcote, A. (2011). Drawing conclusions from choice response time models: A tutorial using the linear ballistic accumulator. *Journal of Mathematical Psychology, 55,* 140–151.

Donkin, C., Brown, S., Heathcote, A., & Wagenmakers, E. J. (2011). Diffusion versus linear ballistic accumulation: Different models but the same conclusions about psychological processes. *Psychonomic Bulletin & Review, 18,* 61–69.

Donkin, C., & Nosofsky, R. M. (in press). A power-law model of psychological memory strength in short-term and long-term recognition. *Psychological Science*.

Eidels, A., Donkin, C., Brown, S. D., & Heathcote, A. (2010). Converging measures of workload capacity. *Psychonomic Bulletin & Review, 17*, 763–771.

Fific, M., Little, D. R., & Nosofsky, R. M. (2010). Logical-rule models of classification response times: A synthesis of mental-architecture, random-walk, and decision-bound approaches. *Psychological Review, 117,* 309–348.

Forrin, B., & Morrin, R. E. (1969). Recognition times for items in short- and long-term memory. *Acta Psychologica, 30,* 126–141.

Heathcote, A., Popiel, S. J., & Mewhort, D. J. K. (1991). Analysis of response time distributions: An example using the Stroop task. *Psychonomic Bulletin & Review, 109,* 340–347.

Hockley, W. E. (1984). Analysis of response time distributions in the study of cognitive processes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 598–615.

Johns, E. E., & Mewhort, D. J. K. (2011). Serial-position effects for lures in short-term recognition memory. *Psychonomic Bulletin & Review, 18,* 1126–1132.

Kahana, M. J., & Sekuler, R. (2002). Recognizing spatial patterns: A noisy exemplar approach. *Vision Research, 42,* 2177–2192.

Lewandowsky, S., & Oberauer, K. (2009). No evidence for temporal decay in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1545–1551.

Little, D. R., Nosofsky, R. M., & Denton, S. E. (2011). Response-time tests of logical-rule models of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1–27.

Marley, A. A. J., & Colonius, H. (1992). The "horse race" random utility model for choice probabilities and response times, and its competing risks interpretation. *Journal of Mathematical Psychology, 36,* 1–20.

McElree, B., & Dosher, B. A. (1989). Serial position and set size in short-term memory: Time course of recognition. *Journal of Experimental Psychology: General, 18,* 346–373.

McElree, B., & Dosher, B. A. (1993). Serial retrieval processes in the recovery of order information. *Journal of Experimental Psychology: General, 122,* 291–315.

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10,* 465–501.

Mewhort, D. J. K., & Johns, E. E. (2005). Sharpening the echo: An iterative-resonance model for short-term recognition memory. *Memory, 13,* 300–307.

Nelder, J. A., & Mead, R. (1965). A simplex algorithm for function minimization. *The Computer Journal, 7,* 308–313.

Nosofsky, R. M., Little, D. R., Donkin, C., & Fific, M. (2011). Short-term memory scanning viewed as exemplar-based categorization. *Psychological Review, 188,* 280–315.

Nosofsky, R. M., & Palmeri, T. J. (1997). An exemplar-based random walk model of speeded classification. *Psychological Review, 104,* 266–300.

Oberauer, K. (2008). How to say no: Single- and dual-process theories of short-term recognition tested on negative probes. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 439–459.

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

Ratcliff, R., & Murdock, B. B., Jr. (1976). Retrieval processes in recognition memory. *Psychological Review, 83,* 190–214.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics, 6,* 461–464.

Schweickert, R. (1992). Information, time, and the structure of mental events: A 25 year review. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 535–566). Cambridge: MIT.

Sternberg, S. (1966). High speed scanning in human memory. *Science, 153,* 652–654.

Sternberg, S. (1969). Memory scanning: Mental processes revealed by reaction-time experiments. *American Scientist, 4,* 421–457.

Sternberg, S. (1975). Memory-scanning: New findings and current controversies. *Quarterly Journal of Experimental Psychology, 27,* 1–32.

Townsend, J. T., & Ashby, F. G. (1983). *The stochastic modeling of elementary psychological processes*. Cambridge: Cambridge University Press.

Townsend, J. T., & Fific, M. (2004). Parallel versus serial processing and individual differences in high-speed search in human memory. *Perception & Psychophysics, 66,* 953–962.

Townsend, J. T., & Nozawa, G. (1995). Spatio-temporal properties of elementary perception: An investigation of parallel, serial, and global familiarity theories. *Journal of Mathematical Psychology, 39,* 321–359.

Townsend, J. T., & Wenger, M. J. (2004). A theory of interactive parallel processing: New capacity measures and predictions for a response time inequality series. *Psychological Review, 111,* 1003–1035.

Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: The leaky competing accumulator model. *Psychological Review, 108,* 550–592.

Van Zandt, T., & Townsend, J. T. (1993). Self-terminating versus exhaustive processes in rapid visual and memory search: An evaluative review. *Perception & Psychophysics, 53,* 563–580.