Journal of Experimental Psychology:
Learning, Memory, and Cognition
2006, Vol. 32, No. 5, 966–982

0278-7393/06/$12.00 DOI: 10.1037/0278-7393.32.5.966

Simple Predictions Fueled by Capacity Limitations: When Are They Successful?

Wolfgang Gaissmaier, Lael J. Schooler, and Jörg Rieskamp
Max Planck Institute for Human Development

Counterintuitively, Y. Kareev, I. Lieberman, and M. Lev (1997) found that a lower short-term memory capacity benefits performance on a correlation detection task. They assumed that people with low short-term memory capacity (low spans) perceived the correlations as more extreme because they relied on smaller samples, which are known to exaggerate correlations. The authors consider, as an alternative hypothesis, that low spans do not perceive exaggerated correlations but make simpler predictions. Modeling both hypotheses in ACT-R demonstrates that simpler predictions impair performance if the environment changes, whereas a more exaggerated perception of correlation is advantageous to detect a change. Congruent with differences in the way participants make predictions, 2 experiments revealed a low capacity advantage before the environment changes but a high capacity advantage afterward, although this pattern of results surprisingly only existed for men.

Keywords: short-term memory capacity, correlation detection, probability learning, ACT-R, sex differences

Nearly 50 years ago, Miller (1956) concluded that people can consider about seven items or categories simultaneously (plus or minus two). The premise of limited cognitive capacities is often directly linked to its supposed negative consequences, such as reasoning errors or poor cognitive performance (e.g., Johnson-Laird, 1983; Kahneman, Slovic, & Tversky, 1982). But are limited capacities merely a liability? There is growing evidence that they can also be beneficial (for an overview, see Hertwig & Todd, 2003).

In the present article, we discuss the benefits of cognitive limits for the detection of correlations, which were shown by Kareev and colleagues (Kareev, 1995a, 1995b, 2000, 2004; Kareev, Lieberman, & Lev, 1997). They made the counterintuitive prediction that limited capacities are beneficial in correlation detection because they force people to rely on small samples. This prediction was derived from the statistical fact that correlations tend to be overestimated in small samples, which was initially supported by behavioral data. However, their theoretical account has been challenged recently because small samples also yield a higher risk of false alarms (R. B. Anderson, Doherty, Berg, & Friedrich, 2005; Juslin & Olsson, 2005). Furthermore, we review empirical evidence that is in conflict with Kareev's theoretical account. Because of these challenges, our goal is to present an alternative explanation for the findings by Kareev and colleagues that they interpreted as supporting their theory. Our alternative explanation is drawn from the probability learning literature, which is tested against Kareev's hypothesis. Before doing this, we describe the domain of correlation detection and explain Kareev and colleagues' arguments and their challenges in more detail.

Limited Capacities and Correlation Detection: The Small-Sample Hypothesis

Correlation detection (or, more generally, contingency assessment) is considered to be an important component of adaptive behavior and has been studied in a variety of domains and with a variety of tasks (for reviews, see Alloy & Tabachnik, 1984; De Houwer & Beckers, 2002). Most studies of contingency assessment are concerned with contingencies between binary variables. They can be described by a 2×2 contingency table (see Figure 1) that shows the frequencies (or probabilities) of the presence or absence of one variable (outcome, e.g., a disease), given the presence or absence of another variable (input, e.g., a symptom).

The phi coefficient,[1] a common measure to compute contingencies between binary variables, is defined as

$$\Phi = (ad - bc)/\sqrt{(a + b)(c + d)(a + c)(b + d)}. \quad (1)$$

Kareev (1995b) argued that people rely on samples from the environment to assess correlations between, for example, two dimensions of a set of objects. The size of these samples is supposed to be bounded by short-term memory capacity. In a

Wolfgang Gaissmaier, Lael J. Schooler, and Jörg Rieskamp, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Berlin, Germany.

Additional materials are on the Web at http://dx.doi.org/10.1037/0278-7393.32.5.966.supp

Our thanks go to Yakoov Kareev for providing his data; Dan Bothell and Niels Taatgen for helping with modeling ACT-R; Michael J. Kane and Richard P. Heitz for providing task materials; and Richard Anderson, Tom Beckers, Rainer M. Bösel, Arndt Bröder, Michael Doherty, and the ABC Research Group for many constructive comments on an earlier version of this article.

Correspondence concerning this article should be addressed to Wolfgang Gaissmaier, Center for Adaptive Behavior and Cognition, Max Planck Institute for Human Development, Lentzeallee 94, Berlin 14195, Germany. E-mail: gaissmaier@mpib-berlin.mpg.de

[1] If correlations are symmetrical (i.e., $a - b = d - c$) and marginal distributions are equal (i.e., $a + b = c + d$), the phi coefficient leads to the same nominal value as ΔP, defined as $\Delta P = a/(a + b) - c/(c + d)$.

	Outcome1	Outcome2	
Input1	a	b	a + b
Input2	c	d	c + d
	a + c	b + d	a + b + c + d
			= N

Figure 1. Prototypical contingency table.

theoretical analysis, Kareev concluded that the use of small sample sizes facilitates the early detection of correlations by amplifying them. Specifically, both the median and the mode of the sampling distribution exceed the population correlation, and the smaller the sample, the more so. Building on the assumption that people's perception of correlation is the result of calculating the correlation on the basis of a sample, Kareev assumed that consideration of a small sample is more likely to result in a more extreme perception of correlation. Because the samples people consider are smaller for people with a lower short-term memory capacity (*low spans*) than for those with a higher short-term memory capacity (*high spans*), the argument goes, low spans should be more likely to perceive the correlation as more extreme, and thereby detect it earlier.

Kareev and his colleagues provided experimental support for this theoretical argument because low spans indeed performed better on a correlation detection task (Kareev et al., 1997). The task consisted of predicting, trial by trial, which of two possible symbols (X or O) an envelope (which could be either red or green) contained. The number of Xs and Os within the envelopes was varied to yield correlations ranging from $\Phi = -.60$ to $\Phi = .60$. A correlation here means that, for example, there are more Xs in red envelopes and more Os in green envelopes. Detecting this correlation helps people to increase their predictive performance. We refer to this task as the *envelope task.* Kareev et al. (1997) concluded that people with a lower short-term memory capacity, and hence a smaller sample size to consider, "perceived the correlation as more extreme and were more accurate in their predictions" (p. 278). We call this Kareev's *small-sample hypothesis* of correlation detection in the remainder of this article.

The phenomenon of a low capacity advantage in correlation detection is particularly surprising, considering that short-term memory capacity has generally been found to be positively correlated with a variety of cognitive abilities, for example, executive functioning (Miyake, Friedman, Rettinger, Shah, & Hegarty, 2001) or performance on the Scholastic Aptitude Test (SAT; Engle, Tuholsky, Laughlin, & Conway, 1999). The correlation between the related construct of working memory capacity and reasoning ability is even more pronounced (Kyllonen & Christal, 1990). Moreover, the theoretical explanation of this low capacity advantage, the small-sample hypothesis, has been criticized on theoretical grounds, and there is also conflicting empirical evidence, both of which are reviewed in the following.

Theoretical Limitations of the Small-Sample Hypothesis

Juslin and Olsson (2005) criticized Kareev (2000) for only taking into account the hit rate when discussing the small-sample advantage, that is, detecting a sample correlation (Φ) given that there is a population correlation (ρ), $p(\Phi|\rho)$. In contrast, Juslin and Olsson stressed the importance of the posterior probability of a hit, $p(\rho|\Phi)$. That is, it is important to consider how likely it is that one correctly infers that there is a population correlation ρ, based on a sample correlation Φ. Applying this method takes false alarms (e.g., believing that there is a positive correlation when it is in fact zero or negative) into account and leads to the conclusion that the alleged benefits of small samples do not occur.

R. B. Anderson et al. (2005), using a signal detection approach, specified some conditions under which a small-sample advantage could hold, even if one takes false alarms into account. Their simulations demonstrated that a small-sample advantage can exist if one makes the additional assumption that people only decide that a correlation is present in the population when the correlation they observe in the sample exceeds a decision threshold. Otherwise, the observed correlation is ignored. If the decision threshold is above or equal to the correlation in the population, a small-sample advantage exists. For more liberal correlation thresholds (i.e., between zero and the population parameter), however, there is a large-sample advantage.

In response to these criticisms, Kareev refined the small-sample hypothesis, arguing that a small-sample advantage is only possible for large correlations (Kareev, 2005; see also Kareev, 2000). However, this restriction makes it problematic to explain, with the small-sample hypothesis, the low capacity advantage observed in Kareev et al. (1997, Experiment 1) because a low capacity advantage was also observed for small correlations. Moreover, empirical evidence conflicting with the small-sample hypothesis also exists, which is reviewed in the following.

Conflicting Empirical Evidence

Kareev et al. (1997) assumed that people who consider smaller samples are likely to perceive correlations as more extreme than they actually are in the population. From this assumption, it follows that people should also estimate correlations as being higher when they base their estimate on a small, compared with a large, sample. However, in experiments in which participants repeatedly explicitly estimate correlations, participants do not estimate higher correlations based on smaller samples, but rather the tendency is that those estimates increase with increasing sample size (e.g., Clément, Mercier, & Pasto, 2002; Shanks, 1985, 1987). Moreover, studies with measures related to short-term memory capacity suggest that people with lower capacities are less accurate in correlation assessment. For instance, such people include those with lower general cognitive ability as measured by SAT scores (Stanovich & West, 1998), those who are elderly (e.g., Mutter & Williams, 2004; Parr & Mercier, 1998), and those who are performing under increased memory demands (Shaklee & Mims, 1982). That is, in correlation assessment, neither a small-sample advantage nor a low capacity advantage has been reported, but rather the opposite. But then, the empirical finding of the low capacity advantage on correlation detection reported by Kareev et al. (1997) has to be reconciled with these other results.

An Alternative Explanation: Differences in Predictive Behavior

Juslin and Olsson's (2005) arguments imply that Kareev et al.'s (1997) task is not really about the detection of correlation. Partic-

ipants did not have to detect a correlation among trials with a correlation present (signal trials) and trials without a correlation (noise trials), but they were separately tested on either signal or noise trials, thereby not encountering the risk of false alarms. Because the task therefore does not really pose a detection problem, one cannot conclusively argue for a low capacity advantage in the detection of correlation. It has only been shown that low spans are more successful given that there is a correlation. The theoretical limitations of the small-sample hypothesis suggest that a different cognitive mechanism could underlie this low capacity advantage. In the following, we illustrate an alternative explanation that builds on a reinterpretation of the task as simple probability learning.

Kareev et al. (1997) assumed that a low capacity advantage in correlation detection stems from a more exaggerated perception of correlation. However, the envelope task (Kareev et al., Experiment 1) did not assess differences in the perception of correlation. Kareev et al. refrained from asking their participants about their perception but rather inferred their perception from their *predictive behavior* (a term used by Estes, 1976, for example). That is, they counted how often a participant predicted an event, given the color of the envelope, for example, how often he or she predicted X, given a red envelope. These frequencies were used to compute what Kareev et al. called the *perceived correlation* by entering them into a contingency table, such as Figure 1, and determining the phi correlation from this table. Inferring perception from behavior requires the strong assumption that people predict events exactly with the relative frequency with which they perceive them.

In our view, it is necessary to disentangle perception and predictive behavior because predictive behavior can differ between people who perceive the same correlation. Thus, it is possible that differences in predictive behavior alone could be sufficient to explain the low capacity advantage. To understand the difference between perception and predictive behavior, and to understand how differences in predictive behavior could be related to capacity limitations, we next draw a connection to the *probability learning* literature that goes back to Brunswik (1939) and Humphreys (1939), and which has been extensively studied since the 1950s (e.g., Estes & Straughan, 1954; for reviews, see Myers, 1976; Vulkan, 2000).

Correlation Detection as Probability Learning

The typical probability learning task consists of repeatedly predicting which of two events will occur next, with one event usually having a higher probability of occurrence. The correlation detection task used by Kareev et al. (1997) is similar because it also requires predicting one out of two events (the symbols X and O, given the color of the envelope). Bauer (1972), for example, used a task that is almost identical to the one used by Kareev et al. However, she did not cast it as a correlation detection task but rather as a probability learning task with two cues (the colors) and criterion events (the symbols).

A very simple predictive behavior that performs well is to always predict the event that, so far, has been observed most frequently. For example, if one event occurs with a probability of 70%, always predicting this event will result in an accuracy of 70%, on average. This behavior is called *maximizing.* Most often, it has been found, however, that the majority of people do not maximize. Instead, what is often found is *probability matching* (Vulkan, 2000), which consists of predicting an event in proportion to its probability of occurrence (i.e., an event that occurs with a probability of 70% is predicted to occur in 70% of the trials). Probability matching, on average, leads to lower accuracy (i.e., expected accuracy of .7*.7+.3*.3 = .58).

The distinction between maximizing and probability matching is relevant for the correlation detection task used by Kareev et al. (1997). Consider the two types of envelopes and the conditional probabilities of the events, given the color of the envelope. Maximizing implies always predicting X when, for example, a red envelope is shown, if X has been observed more frequently in the past when opening red envelopes. Given a correlation between the envelopes' color and the symbols, this would then imply always predicting O when a green envelope is encountered.[2]

The important point is that a person might have perfect perception of the probability of the events (or of the correlation) but behave differently, for instance, by probability matching or maximizing. In contrast, Kareev et al.'s (1997) assumption that it is possible to deduce perception from behavior presupposes that everyone's behavior matches their perception of the conditional probabilities, but that low spans have a distorted perception of this correlation.

Moreover, Kareev et al.'s (1997) explanation requires that people actually think about the task in terms of the correlation between the color of the envelopes and the frequencies of the different symbols within them. But just because this task can be described as a correlation detection task does not mean that the participants view it this way. From the probability learning perspective (e.g., Bauer, 1972), one could assume that participants learn the conditional probabilities of a symbol given a color independently for each color. Thus, the perception of correlation argument would not be applicable. But then, we need to explain how short-term memory limitations could be beneficial from the probability learning perspective, which we do next.

Maximizing Is Fostered by Limited Memory Capacities

The probability learning literature has struggled with the phenomenon of probability matching because it is inconsistent with a person's goal to maximize his or her payoff. West and Stanovich (2003) argued that this inconsistency results from insufficient cognitive capabilities, and it has been shown that this inconsistency can be reduced with extensive training and high monetary payoffs (e.g., Shanks, Tunney, & McCarthy, 2002). At odds with this perspective that people are not smart enough to maximize is evidence that reduced or limited memory capacities are associated with a higher prevalence of maximizing.

On the one hand, there are studies demonstrating that people with lower memory capacities maximize more frequently. Maximizing was shown to be more prevalent for people with lower intellectual abilities (Singer, 1967), for children (Derks & Pa-

[2] However, it is interesting to note that in the case of asymmetric marginal distributions, it can occur that even when observing a positive correlation between two variables, maximizing implies always predicting the same event. For example, consider a sample of 20 red and 20 green envelopes. Imagine that 19 Xs had been observed in red envelopes and 11 Xs had been observed in green envelopes, which leads to a substantial phi coefficient of .46. Nevertheless, because X is the most frequent event for both kinds of envelopes, maximizing implies predicting X every time.

clisanu, 1967; Weir, 1964), and for different kinds of animals, such as pigeons (Herrnstein & Loveland, 1975; Hinson & Staddon, 1983), rats (Bitterman, Wodinsky, & Candland, 1958), and monkeys (Wilson & Rollin, 1959). On the other hand, the likelihood of maximizing is higher for people under the cognitive load of a secondary task, which was shown with a concurrent estimation task (Bauer, 1972; Neimark & Shuford, 1959) and with a verbal working memory task (Wolford, Newman, Miller, & Wig, 2004).

An explanation for this could be that maximizing is very simple—a feature that is often overlooked (Bauer, 1972). In contrast, probability matching could be the remnant of more involved cognitive processes, such as searching for patterns in the sequence of events, which has been nicely demonstrated in a probability learning study by Yellott (1969). In the last block of his experiment, participants always received feedback indicating that their predictions were correct, irrespective of what they predicted. They continued to match probabilities as they did previously, and when they were asked for their impressions afterward, most responded that they finally found the pattern in the sequence. Wolford, Miller, and Gazzaniga (2000) hypothesized that the search for such a pattern necessarily results in behavior that appears to be probability matching because every reasonable pattern will have to match the probabilities.

Preventing complex hypothesis testing, such as searching for patterns by means of instruction, for example, by telling people that the best they could do is reach an accuracy of 75% (Fantino & Esfandiari, 2002), or by making the task look like a gambling task and not a problem-solving task (Goodnow, 1955), increased the prevalence of maximizing. Because working memory capacity is related to hypothesis generation (Dougherty & Hunter, 2003), lower memory capacities could foster maximizing by making complex hypothesis testing, and thereby complex predictive behavior, less likely because it is more memory demanding.

Summary: Differences in Perception Versus Differences in Predictive Behavior

The findings that people with lower or reduced memory capacities show a higher prevalence of maximizing could present a plausible alternative explanation for the low capacity advantage found by Kareev et al. (1997). This implies that low spans are more likely to maximize because they are less likely to test complex hypotheses, and are thereby more likely to settle on simple maximizing. The reasoning behind this explanation and the explanation given by Kareev et al. are strikingly different. Kareev et al. stressed the influence of short-term memory capacity on the perception of correlation, which implies that the behavioral response to the perception is always identical, whereas our alternative explanation builds on the idea that people could very well share the same accurate perception but still differ in how they respond to their perception. That is, we have here two competing hypotheses. In the remainder of this article, we call Kareev et al.'s explanation the *small-sample hypothesis,* and we call our alternative explanation the *predictive behavior hypothesis.*

Modeling the Competing Hypotheses in ACT-R

The central goal of this article consists of testing these hypotheses for a low capacity advantage on the correlation detection task, used by Kareev et al. (1997), against each other. An important step in doing this, which Kareev has not yet carried out, is to specify a precise computational model of the cognitive process of correlation detection. We think it is important that the model specifies the learning process, resulting in a certain perception of correlation and the behavioral response, so that both processes can be disentangled. To model the processes, we use Atomic Components of Thought Rational (ACT-R), which has been developed by Anderson and his colleagues (e.g., J. R. Anderson et al., 2004; J. R. Anderson & Lebiere, 1998). ACT-R is able to account for a wide variety of phenomena including, for example, practice and retention (J. R. Anderson, Fincham, & Douglass, 1999), decision making (Gonzalez, Lerch, & Lebiere, 2003), language learning (Taatgen & Anderson, 2002), and, important for us, probability learning (Lovett, 1998). Implementing the correlation detection task in ACT-R allows us to model the explanation for a low capacity advantage on the basis of differences in perception, as provided by Kareev et al. (1997), versus the explanation based on differences in predictive behavior. Thereby, these models allow us to make divergent predictions for people who differ in their short-term memory capacity.

Implementing the Correlation Detection Task in ACT-R

The core of ACT-R is constituted by the declarative memory system for facts and the procedural system for rules. Here, we focus on the declarative memory system to model the correlation detection task that results in an instance-based model, building on Logan's (1988) idea that previous solutions to a problem are stored in memory as examples that can be retrieved to solve future problems (for a more detailed description of instance learning in ACT-R, see Taatgen, Lebiere, & Anderson, 2006). The declarative memory system consists of chunks that represent information (e.g., about the outside world, about oneself, about possible actions, etc.). These chunks take on activations that determine their accessibility; that is, whether they can be retrieved. When applied to the correlation detection task, chunks represent instances of possible responses to the envelopes encountered in each trial. Altogether, there are four chunks to represent all possible combinations of the envelopes' two colors and the two possible events connected with the envelopes (i.e., "red envelope: X," "red envelope: O," "green envelope: X," and "green envelope: O"). As a consequence of following ACT-R's standard rule for reinforcing chunks, the history of how often and when chunks have been used in the past determines their activation (see below). Because activation is a combination of frequency and recency, different histories can lead to the same activation in any given moment of time.

The model represents the cognitive processes of one single individual solving the envelope task. Each time an envelope is presented, the model attempts to retrieve one of the two responses associated with the envelope's color. For example, if there is a red envelope, the model attempts to retrieve the chunks "red X" and "red O." These two chunks enter a retrieval competition because only one of them can be retrieved at a time. The likelihood of each chunk winning this competition depends on its activation. The activation of a chunk is higher the more frequently and the more recently it has been used. Depending on the activation level, a chunk is probabilistically selected and determines the model's predicted response. After the response, the model receives feedback whether it was right or wrong, which leads to reinforcing the

chunk representing the correct answer. Thus, the chunk that was retrieved and triggered the response, and the correct chunk, are reinforced, which thereby strengthens their activation. This also implies that a correct answer will be reinforced twice, while an incorrect answer results in reinforcing both the chosen and the correct response once.

Formal definitions. Formally, the activation of a certain chunk i is defined as

$$A_i = B_i + \sum_j w_j S_{ji}, \tag{2}$$

where B_i is the base-level activation of chunk i that reflects its learning history, the W_js reflect the attentional weighting of the elements that are part of the current goal, and the S_{ji}s are the strengths of association from the elements j of the current context to chunk i. For our purpose, only the base-level activation is relevant. The base-level activation of a chunk is defined by

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right), \tag{3}$$

where t_j is the time since the jth practice of an item and d is a decay parameter for which .5 has emerged as a default value across a variety of studies (J. R. Anderson et al., 2004). A chunk can only be retrieved if its activation A_i is above a retrieval threshold τ; accordingly, the probability that a chunk is retrieved is

$$P_i = \frac{1}{1 + e^{-(A_i - \tau)/\sqrt{2}s}}, \tag{4}$$

where s controls the noise of the retrieval process. If there is more than one chunk that matches a retrieval request, the probability that a particular chunk is retrieved is

$$P_i = \frac{e^{A_i/\sqrt{2}s}}{\sum_k e^{A_k/\sqrt{2}s}}. \tag{5}$$

If a chunk has been retrieved, the retrieval time is defined as

$$T_i = Fe^{-A_i}, \tag{6}$$

where F is a latency factor.

Parameters that relate to the competing hypotheses. There are two parameters that are of interest to us because they can be related to the two hypotheses (small-sample hypothesis vs. predictive behavior hypothesis), the decay parameter d in the base-level learning equation, and the noise parameter s in the equation specifying the probability of winning the retrieval competition. The decay parameter d affects the impact of recency on the activation of chunks. Note that there is no differentiation between short- and long-term memory in ACT-R. The base-level learning equation that produces rapid initial decay and slower later decay is key to accounts of both short-term memory tasks, such as memory span, and long-term memory tasks, such as free recall (J. R. Anderson, Bothell, Lebiere, & Matessa, 1998). Without decay, each outcome would be weighed equally, irrespective of how long ago it has been observed. A model with high decay puts more weight on recent information and tends to disregard old information. We believe that this parameter offers a precise way to relate the small-sample hypothesis proposed by Kareev (1995b; Kareev et al., 1997) to processes in ACT-R. The higher the impact of recency, the fewer items are important for a decision, which leads to paying attention to a small sample.

The noise parameter s affects how likely it is that the more activated chunk will win the competition. Without noise (i.e., $s = 0$), the most activated chunk will always be retrieved (given that it is above the retrieval threshold τ), resulting in perfect maximizing in the limit. Higher noise allows less activated chunks to be retrieved from time to time. Although such noise results in suboptimal behavior under some conditions, it is also used to model exploration (Taatgen et al., 2006). Thus, the noise parameter provides a simple way to model facets of predictive behavior, without developing a precise model of how people go about searching for patterns. In this regard, it is important not to interpret noise solely as error. Rather, higher levels of noise capture a proliferation of hypotheses that a participant may entertain, yielding behavior that looks like the model is searching for patterns in the data. This searching results in probability matching (the precise value for s leading to probability matching behavior depends on the task), whereas low levels of noise result in deterministic maximizing behavior. We think that the higher complexity of this behavior makes the relation to short-term memory plausible. Therefore, we believe that variation in this parameter nicely captures the predictive behavior hypothesis.

Method

We used two variants of the model to instantiate the two hypotheses for explaining the low capacity advantage. With the first *decay* variant of the model, we represent Kareev's small-sample hypothesis, with fast decay resulting in focusing on a small sample of recent events. With the second *noise* variant of the model, we represent the predictive behavior hypothesis, with low noise resulting in deterministic maximizing behavior.

Kareev kindly provided us with the data from Kareev et al.'s (1997) Experiment 1, with which we constrained the models used in our simulations. We chose the 128 trials from the conditions with $\Phi = |.375|$ with symmetric distributions of Xs and Os contained in the envelopes (i.e., there were 44 Xs [68.75%] and 20 Os [31.25%] contained in envelopes of one color, while this was exactly reversed for the other color). This is the condition that we also used in our experiments (see below). Note that the qualitative modeling results did not depend on the actual correlation; that is, modeling other conditions yielded the same qualitative results. The model was fit to the relative frequency of maximizing responses; that is, the average proportion choosing the maximizing answer on a particular trial that was further averaged within four blocks consisting of 32 trials each. This was done separately for high and low spans as defined by Kareev et al.

While it is, in principle, possible to differentiate between the two model variants quantitatively on the trials that were fitted, it is not possible to disentangle the two hypotheses qualitatively on those trials. Therefore, we considered a manipulation that distinguishes between the two model variants, and thereby the two hypotheses. A change in the correlational structure of the environment (simply referred to as *shift* in the following) allows for such a differentiation (see below). That is, after the initial 128 trials with a correlation of $\Phi = +.375$, we added 128 trials in which the correlation (i.e., the probability of each event given one or the other color) was exactly reversed, that is, $\Phi = -.375$. If, for example, red was predictive for X in the 128 fitting trials, it was predictive for O in the trials after the shift. Thus, we made predictions for how high and low spans would adapt their behavior to this shift,

depending on the variant of the model, and thereby the hypothesis.[3] However, note that this shift was not implemented in Kareev et al.'s (1997) experiment. Thus, we first fitted the two model variants to Kareev et al.'s data, and second, the fitted models were used to predict behavior for a hypothetical shift not conducted by Kareev et al.

To fit the models to Kareev et al.'s (1997) data, we only varied the one parameter representing either of the hypotheses in each of the model variants. That is, in the decay variant of the model, only decay *d* was varied to fit the curves of both low and high spans separately, while noise *s* was held constant. In the noise variant, only noise *s* was varied to fit the curves of both low and high spans separately, while decay *d* was held constant. All other parameters were set to identical values for both model variants. Our parameter search was informal, and there is no guarantee that they produce optimal fits on Kareev et al.'s data. But we were mostly interested in the predictions made by the two model variants after the hypothetical shifts, and there the qualitative results of the model did not change within a wide range of parameter values. Each simulation was run 10,000 times to obtain reliable results.

Results

Given the simplicity of the task, we think it is unrealistic that people fail to retrieve an answer at any point in time. Therefore, the retrieval threshold τ was set to –10 to ensure that the model never fails to retrieve a chunk in both model variants. The latency factor *F* was set to .1. These parameter values are well within the range of parameter values commonly used (see J. R. Anderson & Lebiere, 1998). In the decay variant, we found the best fit for low spans by setting *d* either to be fast (1, representing low spans, $R^2 = .70$) or absent (0, representing high spans, $R^2 = .93$), while keeping the noise *s* constant at .5. In the noise variant, we obtained a good fit by setting the noise *s* to either .45 for low spans ($R^2 = .74$) or .6 for high spans ($R^2 = .95$), while keeping the decay *d* constant at its default value of .5. Overall, the predictions of both model variants are quite good because both models appropriately describe the increasing frequency of maximizing. However, both models miss the drop in the relative frequency of maximizing that the low spans exhibit on the third block, which explains the lower fit for low spans (see Figure 2).

Before the shift, the decay variant of the model predicted a higher frequency of maximizing with a higher decay parameter value, representing faster forgetting, and thereby capturing the behavior of low spans. The noise variant of the model captures the behavior of low spans with the lower value of noise because lower noise predicts a higher frequency of maximizing, representing a more deterministic response. Therefore, both variants of the model allow for the prediction of a difference in maximizing behavior for low and high spans, although based on different mechanisms. However, the decay parameter *d* was not able to fully capture the magnitude of the gap separating the curves.

A clear difference between the predictions of the two variants of the models emerged after the shift. Faster decay also led to increased maximizing after a shift. Thus, according to the decay variant, low spans should perform better both before and after a shift. Moreover, the predicted fast decay advantage is even more pronounced after the shift than before. However, the opposite prediction was observed for the noise variant of the model. Lower noise yielded decreased and not increased maximizing after a shift. The chunks with the highest activations before the shift favor the wrong choice after the shift. Thus, it is likely that a chunk is retrieved that results in an incorrect (i.e., nonmaximizing) answer after a shift, the lower the noise, the more so. Thus, according to the noise variant, high spans who did worse before a shift should outperform low spans after the shift.[4] Figure 2 shows the predictions of the two variants of the model.

Discussion

With the simulations, we tried to make differential predictions between the predictive behavior hypothesis (modeled with noise) and the small-sample hypothesis (modeled with decay). One could argue that the decay model is not a strict translation of the small-sample hypothesis, because the decay model assumes that people rely on samples biased to include more recent items, whereas the small-sample hypothesis assumes that people rely on random samples (see, e.g., Karrev, 2004). Psychologically, it appears more plausible that if people, owing to capacity limitations, have to rely on a sample of data, the sample will tend to include more recent cases rather than randomly sampling from all cases. It is simply that older cases are harder to retrieve. This assumption is embedded in ACT-R's mechanisms for retrieval competition and is endorsed by other researchers by its inclusion in their own computational models of cognition (e.g., Erev, 1998; Rieskamp, Busemeyer, & Laine, 2003; Yechiam & Busemeyer, 2005). But even small random samples are more likely to reveal the shift than large random samples, as we found out in additional simulations (see additional materials on the Web at http://dx.doi.org/10.1037/0278-7393.32.5.966.supp).[5] Thus, also strictly translating the random sample procedure of the small-sample hypothesis results in the same predictions as our decay model. Therefore, we think it is appropriate to model the small-sample hypothesis with differences in decay *d.*

When fitting the two model variants to the data, the noise variant had a slightly better fit in predicting participants' behavior. We could not improve the fit of the decay variant by only varying

[3] For convenience, we present the theory here in its complete form, although it was formalized after running the experiments. Initially, we started out with the informal hypothesis that if a low short-term memory capacity helps people in detecting correlations, then it should also help them in detecting a change in the correlation. The predictive behavior hypothesis was developed after Experiment 1.

[4] Note, however, that this only holds until the activation of the chunk representing the correct (i.e., maximizing) answer is strengthened enough so that it surpasses the activation of the chunk representing the wrong answer. Then, lower noise would turn out to be beneficial once more. That is, the disadvantage after a shift resulting from lower noise will only hold as long as the relative frequency of maximizing is below .5, on average. Therefore, this noise variant of the model predicts that, over time, low spans catch up with high spans, and even outperform them after many trials after the shift.

[5] To find out whether small random samples are better able to detect a shift in the correlational structure of the environment than larger samples, we simulated the late-shift condition of our experiment with 256 trials with a correlation of $\Phi = .375$ followed by 128 trials with a correlation of $\Phi = -.375$. We were interested in how fast the shift would be detected in random samples with sizes *n*, varying between 4 and 10. On each trial after the shift, we randomly sampled *n* previous trials without replacement and computed the sample correlation to see whether it indicated a "correlation more extreme than that of the population" (Kareev et al., 1997, p. 278). That is, the sample correlation had to be more negative than $\Phi = -.375$ to count as detection. The smaller the sample size *n,* the higher was the probability of detecting the shift.

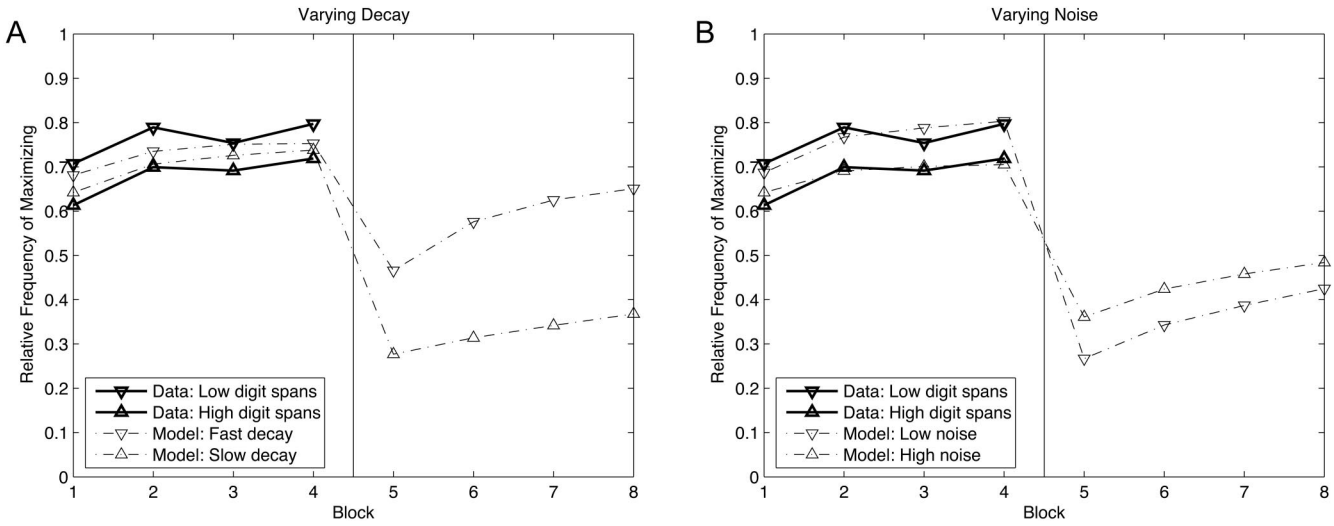

Figure 2. Model predictions of (A) the decay and (B) the noise variant.

decay *d* because the parameter values in the decay variant are at the extremes of the reasonable parameter value space. The extreme values signal a problem with the decay variant because usually they are not set too differently from the default value of $d = .5$. These results already support the predictive behavior hypothesis represented with the noise variant of the model. However, both model variants provide a good fit to Kareev's data. A more decisive comparison can be provided by considering the two qualitatively different predictions that the two variants of the model make after a shift in the environment occurs. Therefore, we think that a correlation detection task (or probability learning task, as we conceptualize it) that includes a shift in the environment will assist in deciding which of the two hypotheses of the low capacity advantage on correlation detection (small-sample hypothesis vs. predictive behavior hypothesis) is more likely. If short-term memory capacity affects people's perception of contingencies (or conditional probabilities) in the manner suggested by the small-sample hypothesis (Kareev et al., 1997), then it should be captured by the decay parameter. The model makes the clear prediction that this should result in a low capacity advantage after a shift. If, however, lower short-term memory capacity fosters simple maximizing, then the data should be congruent with the predictions made by varying the noise parameter. Thus, there should be a low capacity disadvantage after a shift. To test these predictions, we conducted two experiments.

Experiment 1

Experiment 1 was designed to assess the impact of short-term memory capacity on behavior in an extended version of the correlation detection task used by Kareev et al. (1997, Experiment 1). To test our model predictions empirically, we added shifts in the correlational structure of the task (i.e., reversals of the correlations), which made it necessary to conduct a computer version of the task. To obtain a more complete picture of people's cognitive capacities, we administered measures of working memory in addition to the digit span short-term memory task used by Kareev et al.

The idea for using these additional working memory measures was that they allow for testing an additional hypothesis, regarding performance after a shift, not captured by the models. Performance after a shift will depend not only on detecting the change but also on how susceptible people are to proactive interference; that is, how strongly information that they have learned so far will interfere when people attempt to learn new information or when they attempt to adapt their behavior to this new information. Kane and Engle (2000) found that people with a low working memory capacity are more susceptible to proactive interference. Therefore, one could imagine that low spans, even if they detected the shift earlier, are not able to adapt their behavior to this shift appropriately because they are more susceptible to proactive interference. Such an effect could negate a possible advantage, resulting from an earlier detection of correlation. This alternative hypothesis is, in a sense, the opposite of the decay model in ACT-R. While the decay model assumes faster forgetting for low spans, and thereby a recency effect, the proactive interference hypothesis assumes a stronger primacy effect for low spans. That is, it assumes that low spans, owing to proactive interference, put too much weight on old information and thereby fail to adapt to a changing environment.

Method

Participants. Eighty students (42 female, 38 male) with an average age of 24 years ($SD = 3.5$) participated in the experiment. They were paid 7 euros (about U.S. $9) for participation, plus a bonus depending on their performance.

Design and procedure. Each participant was tested individually in a quiet room. We retained the task order of the original Kareev et al. (1997) study. First, short-term memory capacity was measured with a digit span forward task (as in Kareev et al.). Participants were required to verbally repeat sequences of digits that were read to them by the experimenter at a pace of approximately one digit per second. After correct repetition, the length of the sequence increased by one digit, whereas a failure terminated the task. Digit span capacity was determined by the highest number of correctly repeated digits. After the digit span forward task, participants were seated in front of a computer, where the correlation detection task was presented to them. This task was a computer adaptation of the correlation

detection task used by Kareev et al. (Experiment 1). Participants sequentially encountered red and green envelopes on the computer screen. Each time, they had to predict whether the envelope contained a coin marked with an X or an O. They received a 3-s feedback after each trial and were paid 3 euro cents (about U.S. 4 cents) for each correct prediction. Kareev et al. similarly rewarded their participants. Overall, there were 384 trials divided into three seamless blocks (consisting of 128 trials each), in each of which envelopes were drawn randomly without replacement.[6] With regard to differences in performance between high and low spans, the conditions with a correlation of $\Phi \approx |.4|$ had, on average, the largest effect size in Experiment 1 by Kareev et al. Therefore, we decided to administer a condition with a correlation of that size. For all participants, the first block in our experiment corresponded to the symmetric condition with $\Phi = |.375|$. Within this block, each participant encountered an identical distribution of color–symbol combinations consisting of 44 Xs (68.75%) and 20 Os (31.25%) in red envelopes and 20 Xs and 44 Os in green envelopes.

There were four conditions that were identical in the first block but differed according to whether shifts in the correlational structure (i.e., in the probabilities of outcomes given the color of the envelope) occurred in the second or in the third block. A shift always consisted of reversing the correlation, resulting in $\Phi = -.375$. That is, the distribution of symbols within the envelopes was exactly reversed, so that there were 20 Xs and 44 Os in red envelopes, and 44 Xs and 20 Os in green envelopes, in blocks after a shift. This large shift has the methodological advantage of leading to very distinct predictions of the two hypotheses we want to test against each other. Given the probabilistic nature of the task, anything less than this could have been too difficult for the participants to detect. There was no cue to indicate shifts in the correlational structure.

In the first *constant condition,* no shift occurred; in the second *early shift condition,* a shift occurred after the first block; in the third *late shift condition,* a shift occurred after the second block; and in the fourth *back shift condition*, there was a shift after the first block and a shift back to the initial correlation after the second block. These conditions are displayed in Table 1.

The motivation of the different conditions was the following. The constant condition is useful to see how the low capacity advantage, if replicable, develops over time. Because we do not know when a change would affect participants strongly, we think it is useful to also have an early and a late shift condition, independent of which model is more appropriate. If people catch on to a change in the environment quickly, then it is interesting to see how they catch another change as is provided in the back shift condition.

After the correlation detection task, we administered a counting span and an operation span task (Engle et al., 1999) as additional working memory measures. The main difference between short-term and working memory is that short-term memory only requires storage, whereas working memory additionally requires processing (Miyake et al., 2001). The counting span task consisted of counting aloud the objects on the screen and remembering the number for a later test. After several trials, participants had to recall all the numbers from the last two to six trials. For the operation span task, people had to evaluate simple mathematical equations and read aloud words that appeared with the equations on the screen. After two to five trials, they had to write down the words from these trials.

Results and Discussion

For all analyses of behavior in the different blocks, conditions were collapsed if they were comparable up until this point in the experiment. That is, the analyzed block and all previous blocks had to share the same correlational structure. For example, behavior in Block 2 after an early shift can be pooled across the early and the back shift conditions. Table 2 summarizes all correlations between the different capacity measures and the relative frequency of maximizing behavior on the different blocks.

Table 1
Conditions in Experiment 1: Positive or Negative Correlations in the Blocks

Condition	Block 1	Block 2	Block 3
Constant	+	+	+
Early shift	+	−	−
Late shift	+	+	−
Back shift	+	−	+

Replication. Analyzing the first block, which was comparable for all participants, allowed us to check whether we could replicate the low capacity advantage observed by Kareev et al. (1997). In keeping with the original analysis, we split the participants into two groups according to their median digit span capacity. Because it was not clear whether to treat those with median scores as high or low spans, we decided to exclude them. We believe this adds less noise than Kareev et al.'s procedure of randomly categorizing participants with a median value as high and low digit spans. Low digit spans ($M = 73.82$, $SD = 5.40$) performed better on the task than high digit spans ($M = 68.75$, $SD = 9.30$), $t(43.37) = 2.50$, $p = .02$, with corrected degrees of freedom due to higher variance for high spans, $F(1, 54) = 10.68$, $p < .01$. The mean difference corresponds to an effect size of Cohen's $d = 0.67$. This effect size is lower, compared with the corresponding condition of Kareev et al., with an effect of $d = 0.94$. As we deliberately picked a condition with a comparatively large effect size, some regression to the mean is likely to occur. In Kareev et al., the overall effect size was $d = 0.33$. Thus, the effect size in the present study was somewhere between the overall effect size Kareev et al. had observed and that which was observed in the conditions closest to our own. In sum, the original finding could successfully be replicated.

The variance for high digit spans was higher because their prevalence of maximizing was lower, on average. A group of participants who adopted perfect maximizing would have the same expected performance. In contrast, a group of participants who did not adopt maximizing would, on average, perform less well, compared with the maximizing group, but would also show much more variance in performance, which could, in principle, vary between 0% and 100% accuracy.

Because the performance depends to a certain degree on chance, we decided to focus on the relative frequency of maximizing. For each participant, we computed the proportion of trials in which participants chose the option corresponding to maximizing (i.e., choosing X if red and O if green before the shift, and vice versa after the shift). A value of .5 reflects random behavior, a value close to the frequency of the more frequent event in the environment (68.75%) reflects probability matching, and a value of 1 reflects perfect maximizing. We argue that this measure is less noisy than the performance because it is independent of the outcome of a decision (although it naturally correlates with performance; $r = .89$, $p < .01$). We think that this measure is easier to grasp intuitively than the measure Kareev et al. (1997) used, which

[6] We wanted to be as close as possible to Kareev et al.'s (1997) Experiment 1, in which people drew envelopes from a real bag, also without replacement.

Table 2
Summary of Results in Experiment 1

	Maximizing						
	Preshift			Postshift			
				Early		Late	Back
Measure	Block 1	Block 2	Block 3	Block 2	Block 3	Block 3	Block 3
Digit span							
r	−.23	−.16	−.44	.13	−.04	.49	−.11
p	.04	.32	.05	.41	.87	.03	.64
Counting span							
r	.01	−.15	−.21	−.14	−.08	.19	−.12
p	.96	.37	.38	.38	.73	.43	.61
Operation span							
r	−.06	−.17	−.08	−.03	−.15	−.13	−.14
p	.60	.30	.73	.85	.53	.58	.56
n	80	40	20	40	20	20	20

they originally called perceived correlation. The relative frequency of maximizing is correlated by 1 to perceived correlation, and for our analyses, it made no difference which measure was applied.

In the analysis reported above, we used a median split to correspond with Kareev et al.'s (1997) analysis. However, median splits decrease statistical power and can introduce error, primarily because the inherent variability of the predictor is reduced (Irwin & McClelland, 2003). Therefore, in the following analyses, we computed correlations to include all levels of digit span capacity where this was applicable. The low digit span capacity advantage was also reflected in a negative correlation between digit span capacity and preshift maximizing on the first block ($r = –.23$, $p = .04$), indicating that low digit spans show maximizing more frequently in this block. The course of preshift maximizing on the first 128 trials is depicted in Figure 3.

Postshift trials. In the trials after a shift, the correlation in the environment was reversed. Therefore, maximizing now consisted of choosing the opposite object, given a color (e.g., O is now the maximizing answer, given red, because the maximizing answer was X previously). In contrast to the small-sample hypothesis, there was no relation between digit span capacity and postshift maximizing behavior on the early postshift block ($r = .13$, $p = .41$), and even a high digit span capacity advantage, indicated by a positive correlation between digit span capacity and postshift maximizing on the late postshift block, was observed ($r = .49$, $p = .03$). These results are contrary to the prediction of the

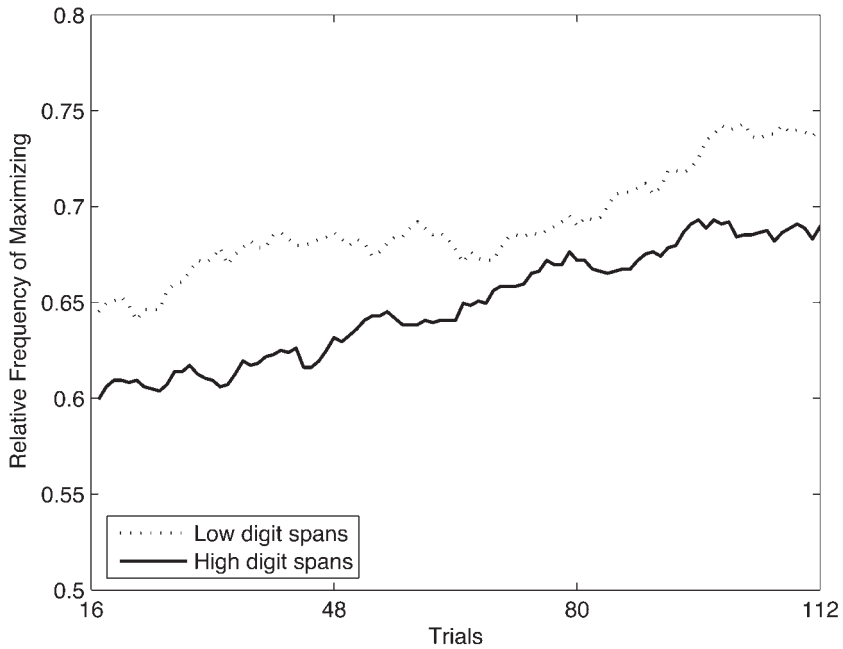

Figure 3. Preshift maximizing on Block 1, Experiment 1. The amount of maximizing is averaged within a moving window of 32 trials and is reported separately for high and low digit spans as derived by the median split.

decay variant of our model implementing the small-sample hypothesis. According to the decay variant model, the low digit span capacity advantage, corresponding to a fast decay parameter value of the model, leads to an even more pronounced advantage after a shift. Instead, the data revealed either no effect or the opposite, and are thereby congruent with the predictions made by the noise variant model representing the predictive behavior hypothesis. Postshift maximizing behavior was only related to digit span capacity in the late shift condition, and here, the correlation was positive. That is, high digit spans adopted maximizing with a higher relative frequency after the late shift. This condition is depicted in Figure 4.

Other working memory measures. Naturally, digit span capacity was correlated with both counting span ($r = .24$, $p = .03$) and operation span ($r = .24$, $p = .03$). However, the other working memory measures were unrelated to pre- and postshift behavior. That is, neither the low digit span capacity on preshift trials nor the high digit span capacity advantage on postshift trials could be captured by those measures. If the high digit span capacity advantage on the postshift trials were due to higher proactive interference of the low digit spans, this should be captured with one of the other working memory measures, which were also used by Kane and Engle (2000). Therefore, we are confident that this high digit span capacity advantage on postshift trials indeed favors the predictive behavior hypothesis (although we try to more carefully rule out the proactive interference hypothesis in Experiment 2; see below).

Preliminary Conclusion

The overall picture supports the hypothesis that it is not people's perception of correlation that differs between people with high and low digit span capacity, but differences in predictive behavior (i.e., differences in how consistently they maximized their payoffs). There was a low digit span capacity advantage before a shift, but no difference or even a high digit span capacity advantage that emerged after a shift. Thus, the data are not at all congruent with the decay variant of the model, but they are congruent with the noise variant, and thereby, our assumption that differences in predictive behavior are of importance.

However, we did not find a high digit span capacity advantage after an early shift, but only after a late shift. Because the sample size of the late shift condition in which we found a postshift high digit span capacity advantage is small ($n = 20$), we should interpret this finding with care.

Experiment 2

The second experiment was a slightly refined version of the first, intended to replicate the important results of Experiment 1. Now, we know that a change in the correlational structure of the environment only reveals differences between high and low spans after many trials. Therefore, we only implemented the late shift condition in which we found a high capacity advantage. We also wanted to more strongly rule out the alternative hypothesis that high spans were at an advantage after a shift because they were less susceptible to proactive interference. In Experiment 1, we only addressed this question by assessing additional working memory measures that were shown to be related to proactive interference (Kane & Engle, 2000). However, Kane and Engle used extreme group comparisons and a large sample size (192 and 216 participants, respectively) to show the modest relation between working memory and susceptibility to proactive interference. That is, there could have been proactive interference that we did not capture with

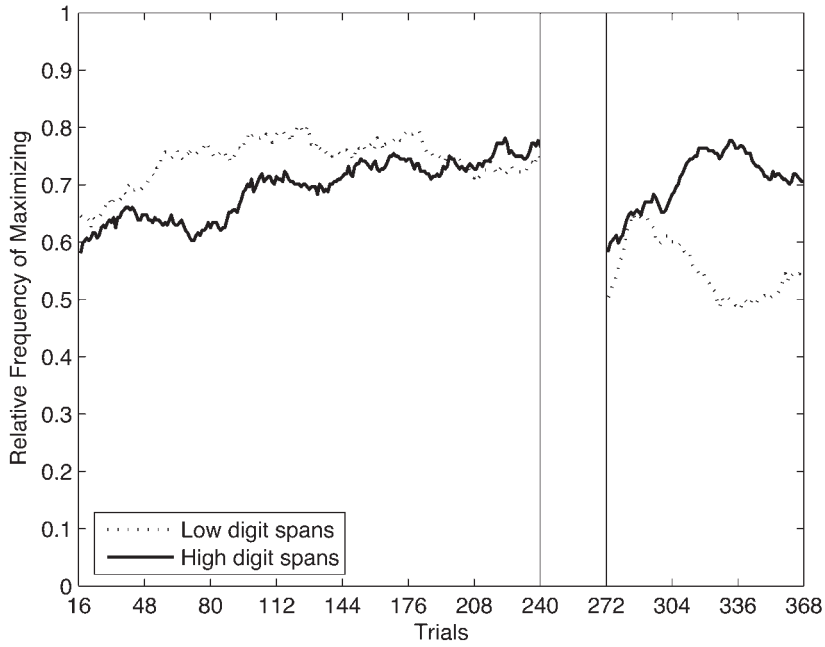

Figure 4. Maximizing on all trials, Experiment 1, late shift condition. Low and high digit spans were averaged separately across trials within a moving window of 32 trials. To prevent an overlap between trials before and after the shift in this window, we started averaging again after the shift, which is indicated by the two lines. That is, the last depicted data point before the shift consists of the last 32 trials before the shift, and the first depicted data point after the shift consists of the first 32 trials after the shift.

our working memory measures. Therefore, we assessed susceptibility to proactive interference directly.

Method

Participants. Eighty students (51 female, 29 male) with an average age of 24 years ($SD = 3.6$) participated in the study. They were paid 9 € (about U.S. $11.50) for participation, plus a bonus depending on their performance (identical to Experiment 1, per correct trial 3 € [U.S. 4] cents).

Design and procedure. Each participant was tested individually in a quiet room. Again, we kept the task order as in the original study by Kareev et al. (1997), starting with the digit span forward task to measure short-term memory capacity. This time, digit strings were digitally recorded beforehand, so that participants listened to identical audio files instead of listening to an experimenter reading the digits to them. The correlation detection task consisted of only the late shift condition of Experiment 1, with a shift seamlessly occurring after two blocks. Colors of the envelopes and keys on the keyboard (e.g., whether X was left or right) were counterbalanced. For a more detailed description of the task, see Experiment 1.

We included the counting span task again (see Experiment 1). Furthermore, we assessed susceptibility to proactive interference (Kane & Engle, 2000), which we considered to be a possible alternative explanation for the high digit span capacity advantage after a shift in Experiment 1. This task consisted of learning three word lists with words that belong to one category (professions) and one word list that belongs to another category (animal names). The words were presented successively, and participants had to recall as many words as possible after each list. It is usually observed that performance decreases over the course of the three word lists from one category (*proactive interference*) and then increases again on the last word list (*proactive interference release*).

Results and Discussion

A repeated measures analysis revealed no difference between the counterbalanced conditions with regard to maximizing in the three blocks, $F(5, 127) = 0.76$, $p = .58$. Therefore, all counterbalancing conditions were merged. It was surprising that the original low capacity advantage on preshift maximizing could not be found in Experiment 2. There was no significant correlation between digit span capacity and preshift maximizing on Block 1 ($r = -.08$, $p = .50$) and on Block 2 ($r = -.10$, $p = .38$). There was also no postshift high digit span capacity advantage; postshift maximizing on Block 3 was unrelated to digit span capacity ($r = .10$, $p = .39$).

Neither proactive interference nor its release could predict any behavior. That is, postshift maximizing really does not seem to be a function of susceptibility to proactive interference at all. Proactive interference was not correlated with digit span or counting span. Surprisingly, counting span was positively correlated to preshift maximizing on Block 2 ($r = .27$, $p = .02$).

Because both experiments were almost identical in structure, this result surprised us. Therefore, we suspected that some peculiarity of our sample in Experiment 2 might be responsible. Digit span capacity and counting span capacity were comparable between the experiments. The only demographic variables assessed were age and sex. The only difference between the samples from the two experiments that struck us was the larger proportion of women in Experiment 2, compared with Experiment 1 (63.8% vs. 52.5%), which suggested that we should explore sex differences in a post hoc analysis.

Post Hoc Analyses of Sex Differences

One reason for the different results might be based on sex differences because a different proportion of men and women participated in Experiment 2. We decided to merge the data sets from our two experiments to have a reasonable sample size to analyze men and women separately.

Merging the data sets only makes sense for blocks that are identical in both position (i.e., first, second, third) and learning history for both experiments, which is the case for the first two preshift blocks and the late postshift block. It results in sample sizes of $n = 160$ for preshift Block 1, $n = 120$ for preshift Block 2, and $n = 100$ for postshift Block 3 from the late shift condition. For all other blocks, we do not have an appropriate sample size to further divide them by sex. Individual difference measures assessed in both experiments were digit span and counting span.

An examination of the correlations between digit span and counting span, on the one hand, and relative frequency of maximizing, on the other hand, separately for men and women, indeed revealed a sex difference. The preshift low digit span capacity advantage and the postshift high digit span capacity advantage only existed for men but not for women. For women, there was even a positive correlation between counting span and preshift maximizing (see Table 3).

To illustrate this, Figure 5 depicts the relative frequency of maximizing, separately for men and women from the late shift condition in Experiment 1 and from Experiment 2 in which only the late shift condition was conducted. Men and women were separately divided into high and low digit spans with a median split (based on all participants), and the relative frequency of maximizing is averaged within a moving window of 32 trials.

The difference lies in the interaction. Men and women did not differ on absolute levels of relative frequency of maximizing ($M_{\text{Men}} = 0.64$; $M_{\text{Women}} = 0.65$), $F(1, 158) = 0.22$, $p = .64$, or performance ($M_{\text{Men}} = 70.82\%$; $M_{\text{Women}} = 70.99\%$), $F(1, 158) = 0.02$, $p = .89$, on Block 1, which we chose for this comparison because there are comparable data for all participants on this block. Note, however, that digit span capacity was higher for men ($M_{\text{Men}} = 6.22$; $M_{\text{Women}} = 5.85$), $F(1, 158) = 4.32$, $p = .04$, which was not the case for counting span ($M_{\text{Men}} = 0.71$; $M_{\text{Women}} = 0.69$), $F(1, 158) = 0.34$, $p = .56$. There was a correlation between digit span and counting span for men and women ($r = .26$, $p = .03$ and $r = .21$, $p = .04$).

Fortunately, we were able to test whether this sex difference is a peculiarity of our samples or something that may be more general because Kareev provided the original data set from Kareev et al.'s (1997) Experiment 1. It included a total of 112 participants (64 women, 48 men). Note that this experiment did not include a shift in the correlational structure, so that we could only test whether the sex difference on preshift trials also holds there. It does: There only is a (negative) correlation between performance and digit span capacity for men ($r = -.28$, $p = .06$) but not for women ($r = .06$, $p = .66$). The same holds for the correlation between digit span and the absolute strength of perceived correlation (which corresponds to the variable we call maximizing), which only existed for men ($r = -.29$, $p = .05$) but not for women ($r = -.05$, $p = .68$). Here, men and women did not differ with respect to performance, $F(1, 110) = 0.44$, $p = .51$; the absolute strength of perceived correlation, $F(1, 110) = 1.44$, $p = .23$; and digit span capacity, $F(1, 110) = 1.54$, $p = .22$.

Table 3
Exploring the Sex Difference of the Interaction Between Capacity and Maximizing

	Maximizing					
	Men			Women		
	Preshift		Postshift late	Preshift		Postshift late
Measure	Block 1	Block 2	Block 3	Block 1	Block 2	Block 3
Digit span						
r	−.19	−.43	.36	−.10	.13	.12
p	.12	<.01	.03	.37	.30	.34
Counting span						
r	.03	−.09	−.03	.18	.33	.21
p	.83	.53	.86	.09	<.01	.11
n	67	50	38	93	70	62

In summary, the low digit span capacity advantage on trials before a shift only exists for men, which was the case for the present experiments and Kareev et al.'s (1997) Experiment 1. In Experiment 2, the preshift low digit span capacity advantage developed over time and was stronger on preshift Block 2. In our view, this strengthens our argument that the difference between high and low digit spans lies in differences in predictive behavior. If the difference lied in perception, and thereby in the earlier detection of the correlation by low digit spans, as assumed by Kareev et al., then this difference should be more pronounced earlier rather than later. The postshift high digit span capacity advantage also existed only for men. That is, for women, digit span was unrelated to behavior. It is interesting to note that counting span was related to behavior, but in the opposite direction: It was positively correlated to the relative frequency of maximizing before a shift.

General Discussion

The goal of this article was to disentangle two potential explanations for the stunning finding of a low digit span capacity advantage on correlation detection (Kareev et al., 1997). Kareev et al.'s original explanation was that low digit spans perceive correlations as more extreme than they actually are because they base their estimates on smaller samples from the environment. Small samples statistically tend to overestimate correlations, and this overestimation can be advantageous in correlation detection. We have called this the *small-sample hypothesis*.

However, the small-sample hypothesis has been criticized theoretically (R. B. Anderson et al., 2005; Juslin & Olsson, 2005), and some studies dealing with contingency assessment provide conflicting evidence (e.g., Clément et al., 2002; Shanks, 1985, 1987). Therefore, we explored whether the low digit span capacity advantage found by Kareev et al. (1997) could be explained differently. Instead of assuming that people differ in their perception of correlations, we assumed that people differ in their predictive behavior. This *predictive behavior hypothesis* was inspired by revisiting the related probability learning literature, which revealed convergent evidence showing that the most successful predictive behavior (maximizing) can be related to a reduced or limited memory capacity.

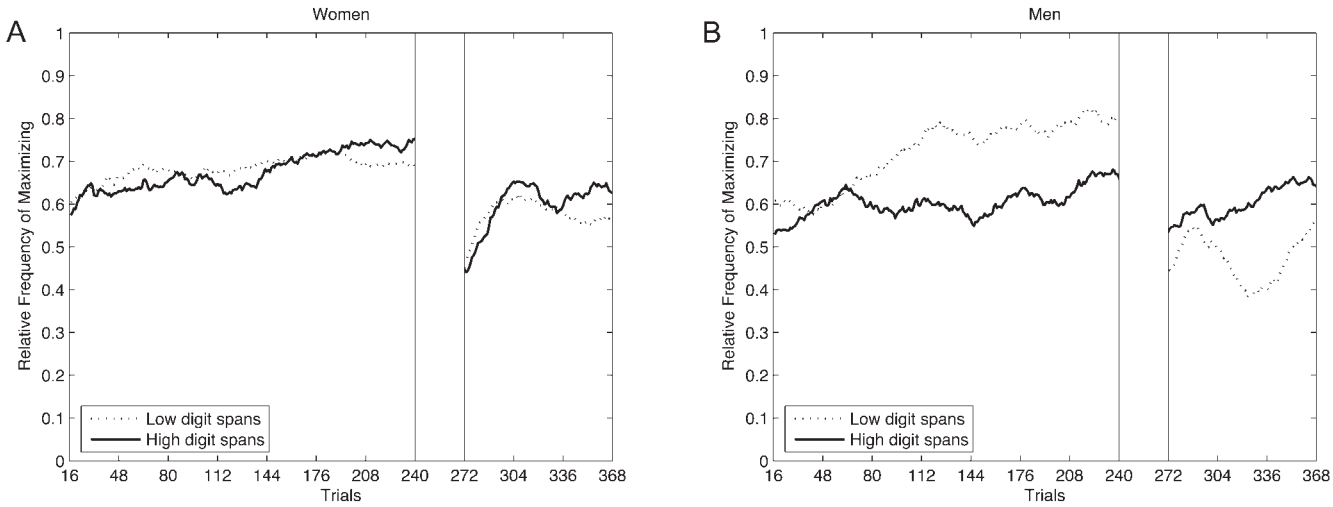

Figure 5. Maximizing separately for high and low digit spans, late shift condition, both experiments, for (A) women and (B) men. We averaged low and high digit spans separately across trials within a moving window of 32 trials and started averaging again after the shift, which is indicated by the two lines (see also Figure 4).

We therefore hypothesized that low digit spans are more likely to maximize their payoffs more consistently, resulting in the low digit span capacity advantage. Based on the initial learning trials as implemented by Kareev et al. (1997), one cannot distinguish conclusively between the small-sample hypothesis and the predictive behavior hypothesis. However, by instantiating both of these explanations in ACT-R models, it was possible to demonstrate that these hypotheses make different predictions about how participants will behave after a shift in the correlational structure of the environment. The model that implemented the small-sample hypothesis predicted a low digit span capacity advantage before and after a shift. In contrast, the model that implemented the predictive behavior hypothesis exhibited a low digit span capacity advantage before a shift but a high digit span capacity advantage after a shift.

Support for Differences in Predictive Behavior

The results of Experiments 1 and 2 replicate the low digit span capacity advantage found by Kareev et al. (1997), although these and the following results only held for men (see *A Puzzling Sex Difference* section, below). After a shift in the environment, however, we found either no difference between high and low digit spans or a high digit span capacity advantage. This is contradictory to the assumption made by the proponents of the small-sample hypothesis that high and low digit spans differ in their perception, but it is congruent with our assumption that this difference lies in predictive behavior.

However, if one wanted to keep the perceptual argument, one could argue that low spans are less likely to engage in further sampling, because they perceive the correlation as more extreme, and hence reach a conclusion faster. High spans, in contrast, keep sampling, because they have observed a weaker correlation and are less committed to their estimate, and hence are less at a disadvantage when a change takes place. Nevertheless, we see an advantage of our explanation is its consistency with similar findings of a low capacity advantage in the binary choice probability learning literature (e.g., Wolford et al., 2004). The typical binary choice task is very similar to the task at hand, but a sample-based perceptual argument cannot hold, because one only needs to acquire information about proportions. In contrast to correlations, sample proportions are unbiased estimators of population proportions. Thus, our account covers those highly related results, which the perceptual argument does not. Moreover, the results for men revealed that the low capacity advantage developed over time (at least in Experiment 2). It was stronger on Block 2 than on Block 1, which is, in our view, further counterevidence for the small-sample hypothesis. According to it, the low digit span capacity advantage lies in the early detection of strong correlations due to the small-sample bias to overestimate these correlations. Thus, it should plausibly have the largest effect early in the experiment.

Estimation Versus Prediction

We think that explaining the low capacity advantage in this correlation detection task with differences in predictive behavior rather than with differences in the perception of correlation also reconciles this low capacity advantage with apparently conflicting empirical evidence. Assuming that people with a lower short-term memory capacity consider smaller samples, and thereby perceive correlations as more extreme, conflicts with findings that correlation estimates are higher with larger rather than with smaller samples (e.g., Clément et al., 2002; Shanks, 1985, 1987). For correlation estimation tasks, there is usually also no low capacity advantage reported, but rather the opposite (e.g., Shaklee & Mims, 1982). But estimation and prediction are two different processes. For the task at hand, precise estimation is not necessary. It suffices to figure out which symbol is more frequently associated with which color. Building on that, the simpler predictive behavior by low spans is more successful, at least until the shift.

Plausibility of the ACT-R Model

The results we found are congruent with the noise variant of our model that we use to implement the predictive behavior hypothesis. Noise has been used in ACT-R to account for different levels of explorative behavior (Taatgen et al., 2006). The noise parameter can account for differences in predictive behavior, capturing the result that high digit spans tend to probability match whereas low digit spans are more likely to maximize.

At first glance, it may not be clear how the noise parameter relates short-term memory capacity to the tendency to match the probabilities. The connection may be that people do not actually attempt to match probabilities, but rather that probability matching is the result of more complex predictive behavior, such as pattern search (e.g., Wolford et al., 2004). Looking for patterns in the envelope task requires tracking the order of events, which could place high demands on memory. In contrast, explicitly assessing a correlation requires just knowing the frequencies of color–symbol combinations, because the order in which these appeared is irrelevant. So the cause of the low span advantage could be that people with a low short-term memory capacity find it difficult to entertain complex patterns and, therefore, tend to settle on maximizing. Because noise is used to model the outcome of either simple predictive behavior (maximizing) or rather complex predictive behavior (pattern search, resulting in probability matching), we think it is reasonable to capture this short-term memory phenomenon with noise in ACT-R.

It is plausible that more complex predictive behavior helps people to adapt to a shift appropriately. But even though we use noise to model this complexity, this does not mean that more noise is always better and that more noise can always be interpreted as higher complexity. More noise can also mean unsystematic, random variability, which can be detrimental. If noise were too high, for example, then the predicted behavior would be approximately at chance level before and after a shift. Results that could be interpreted in this way were reported by Sanford and Maule (1973). They compared the relative frequency of maximizing between young and old people in a simple binary probability learning task, including a shift. Older people performed worse than young people before and after the shift. While young people reached probability matching or slightly overmatched both before and after the shift, old people stayed below the probability matching level even before the shift and were approximately at chance level after the shift. This could indicate high levels of noise (in the sense of too much random variation) for the old people, which clearly slows the adaptation to a shift.

Besides the psychological plausibility of the model, we were also interested in discussing the model in terms of signal detection theory. As Juslin and Olsson (2005) pointed out, it does not suffice to look at the hit rate, but one ultimately needs to consider the

posterior probability of a hit, which also takes false alarms into account. More specifically, we were interested in the question whether the noise and the decay models make different predictions about high and low spans with regard to sensitivity and response bias. To analyze the models' sensitivities, we compared how well they could distinguish signal trials (as described in the part about the ACT-R simulations) from noise trials (i.e., trials in which the distribution of symbols in the envelopes was random). To analyze the models' response biases, we looked at the false alarm rates of each model's responses on the noise trials, which indicate how likely it is that a model interprets noise as signal.

Presenting these signal detection analyses in detail would go beyond the scope of this article (for details, see additional materials on the Web at http://dx.doi.org/10.1037/0278-7393.32.5.966.supp). Essentially, the results suggest that both the noise and the decay models predict a higher sensitivity for the low spans compared with the high spans. That is, the low spans are better able to distinguish the signal from the noise trials. Furthermore, the noise model also predicts a higher response bias for low spans, whereas the decay model predicts rather the opposite (and also a much smaller difference in response bias). That is, the noise model, which we believe to more accurately describe the differences between high and low spans, predicts that low spans are more likely to interpret something as signal. These simulation results fit nicely with our interpretation that low spans explore less and settle more quickly on maximizing, whereas high spans are more careful in drawing conclusions and continue to explore longer.

Relations to Other Models

The implementation in ACT-R is based on the idea that instances of possible solutions are stored in memory that can be used to solve future trials (Logan, 1988). Therefore, it is interesting to consider similarities and differences to other instance-based models such as exemplar models. One exemplar model that has been used to explain multiple-cue probability judgment is ProbEx (probabilities from exemplars; Juslin & Persson, 2002). ProbEx could plausibly solve the envelope task by storing the outcome of each trial as a separate exemplar, with the exemplars containing information about the color of the envelope and the symbol within it. On each trial, it activates exemplars as a function of their similarity to the stimulus. Because each stimulus has only one feature (the color of the envelope), there are always a large number of exemplars that have exactly the same similarity. Thus, it would basically retrieve exemplars proportionally to their frequency of occurrence. ProbEx has a deterministic choice rule and will always select the option supported by more exemplars. But because the sampling of exemplars is probabilistic, ProbEx still could predict behavior similar to probability matching on the aggregated level. Another possibility for ProbEx to approach the task would be to store not only the outcomes as exemplars but also the answers as additional exemplars. This would make it similar to our ACT-R models, in which the activation of a chunk is also a function of outcome and behavior. In this manner, ProbEx would predict overmatching and asymptotically approach maximizing over time, because the proportion of exemplars representing the maximizing answer will grow steadily the more often it is chosen. Similar to our model without decay, ProbEx does not forget exemplars and so after the shift would have difficulty overcoming its initial response tendencies.

The exemplar-based random walk model (EBRW; Nosofsky & Palmeri, 1997) weighs recent exemplars more strongly and thus will be better able to capture a shift. This model is an extension of the general context model (GCM; Nosofsky, 1986), which predicts probability matching (Nosofsky, Kruschke, & McKinley, 1992). If, in a categorization task, GCM receives Category A feedback in 70% of the cases, it will predict Category A in 70% of the cases. Because EBRW includes GCM as a special case, it can also account for probability matching. However, it has a parameter that allows the theory to account for maximizing, namely the response criteria defining how much evidence the model needs before making a decision. The higher the response criterion, the more deterministic (i.e., maximizing) the choices predicted by the model will be, resulting in a greater sensitivity to differences between the two categories. Such a model would be very similar to our predictive behavior models, in which we modeled higher sensitivity with lower noise.

The results presented here could also be interpreted in the light of the RELACS model (reinforcement learning among cognitive strategies; Erev & Barron, 2005; see also Rieskamp & Otto, 2006). The model is able to capture a large variety of binary choice tasks. It assumes three different cognitive strategies that are involved in those tasks: fast best reply (i.e., select the action with the highest recent payoff), case-based reasoning (i.e., choose the best action that led to the best outcome in a similar case in the past), and slow best reply (i.e., a slow learning of the strategy likely to maximize earnings). The pattern search process could be modeled either with a high proportion of case-based reasoning or with a high exploration in slow best reply. Both of these implementations would predict a deviation from maximizing, similar to increased noise in the ACT-R model we applied.

Why Is Digit Span a Better Predictor Than Other Working Memory Measures?

We were surprised to find a relation only between behavior and short-term memory capacity assessed with a digit span test but not with one of the working memory measures, counting span or operation span. It is unlikely that this is due to a lack of reliability in these measures, because these tasks usually have a reliability, based on internal consistency, between .70 and .90 (with 0 and 1 being the borders *no reliability* and *perfect reliability*; Conway et al., 2005). As pointed out before, short-term memory tasks tend to emphasize the simple storage of information, whereas working memory tasks require the additional processing of the information being stored (Miyake et al., 2001).

We have argued that the lower performance of high digit spans is the result of their more complex predictive behavior. They search for patterns, but because there are none, they fail. Of our capacity measures, only digit span capacity was related to the prevalence of maximizing, suggesting that simple storage is particularly important for pattern search. If one adopts the strategy of rehearsing the sequence of events to search for patterns in it, then one constraint on pattern search is storing the sequence. Other processes involved in pattern search, such as hypothesizing about specific patterns, should be more strongly related to working memory (for the relation between working memory and hypothesis generation, see, e.g., Dougherty & Hunter, 2003). Finding no

relation between working memory and maximizing, however, suggests that this part of pattern search, at least for this task, did not tax the working memory capacity of even our low-span participants. So, the difference between simple storage (required for the digit span test) and more complex processing (additionally required for the working memory tasks) would explain why we only found a relation between digit span capacity and behavior.

A Puzzling Sex Difference

We found an intriguing sex difference in the interaction between digit span capacity and predictive behavior. Only men exhibited the low digit span capacity advantage before a shift and the high digit span advantage after a shift. In contrast, short-term memory capacity did not explain any variance in the behavior of women. This sex difference in the interaction between short-term memory and predictive behavior exists in the data from Experiments 1 and 2 and in Kareev et al.'s (1997) data.

In the probability learning literature, a sex difference with respect to the absolute amount of maximizing has been reported: West and Stanovich (2003) found that men were more likely to deliberately opt for a maximizing strategy when a typical probability learning task was described to them and they had to specify, in advance, what they would do. Furthermore, there are reports of sex differences favoring men in a similar task, the Iowa Gambling Task (Overman, 2004; Reavis & Overman, 2001). However, in the data reported here, there is no sex difference in decision making (here: maximizing behavior) per se, but only in the interaction between short-term memory capacity and maximizing.

Because maximizing behavior is comparable on average, it is likely that men and women do not differ with regard to the complexity of their predictive behavior on average. However, looking at the low digit spans only, it seems to be the case that women are better able to engage in complex behavior (resulting in nonmaximizing) despite a low digit span capacity. This suggests that women can draw on resources other than simple storage capacity, whereas men draw more exclusively on the simple storage that the digit span test presumably taps.

Because we do not have additional data, we can only speculate about what these resources could be. Reliable sex differences favoring women have been repeatedly shown on episodic memory tasks, particularly those with a verbal component (for an overview, see Herlitz, Nilsson, & Bäckman, 1997). More generally, females surpass males on tasks in which verbal processing of material is either required or can be used (Lewin, Wolgers, & Herlitz, 2001). This indicates that women could more easily engage in verbal processing to solve a task. Speculating about patterns in sequences of events is a task in which verbalization clearly is possible. Thus, women could draw on verbal episodic memory to search for patterns, whereas men are apparently more likely to depend on short-term memory to store these sequences. This would explain why digit span capacity only explains the variance in the maximizing behavior for men but not for women.

Conclusion

We reported counterevidence to the small-sample hypothesis of correlation detection. Our modeling and empirical results support the view that differences between high and low digit spans lie in differences in predictive behavior and not in differences in perception, although the whole effect only seems to exist for men. Yet, we agree with Kareev et al. (1997) that subtle benefits can follow from what are commonly seen simply as limitations (Hertwig & Todd, 2003; Schooler & Hertwig, 2005). For example, forgetting has been interpreted not simply as regrettable failures of the memory system but as reflecting statistical patterns with which information recurs in the environment (J. R. Anderson & Schooler, 1991).

Therefore, while we are, in general, sympathetic to the idea that limitations of cognitive capacities can serve adaptive functions, we disagree with Kareev et al.'s (1997) assertion that limitations in short-term memory amplify the detection of correlation by forcing people to rely on small samples. Instead, we think that these limitations foster simpler predictive behavior, choosing the more frequent option on every trial (maximizing) because of an incapability to apply more complex predictive behavior. In this experiment, simple predictive behavior, maximizing, is more successful than any other more complex predictive behavior, such as pattern search, when the correlational structure of the task is stable. However, applying this behavior consistently seems to put people at a disadvantage if the environment changes.

Not only in the laboratory but also in the real world, it often pays to explore alternatives, looking for changes and other patterns in the statistical structure of the environment. Ayton and Fischer (2004) speculated that the condition of constant probabilities that maximizing exploits rarely holds outside of psychological laboratories and casinos. Furthermore, the cost of missing a nonrandom sequence could well be higher than the price of detecting patterns where there are none (Lopes, 1982). Here we modeled what we take to be systematic exploration with random noise, but even random noise has been shown to be an effective way to escape local minima in optimization problems (Kirkpatrick, Gelatt, & Vecchi, 1983).

We started by testing the small-sample hypothesis, which considers less information to be helpful, and end by noting that noisy behavior, which can of course be harmful, has the potential to be beneficial. Like Kareev (2000), we emphasize that it is crucial to consider the match between a cognitive process and the environment in which it operates, because what works well in one environment may work poorly in another. No single strategy is optimal per se.

References

Alloy, L. B., & Tabachnik, N. (1984). Assessment of covariation in humans and animals: The joint influence of prior expectations and current situational information. *Psychological Review, 91,* 112–149.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review, 111,* 1036–1060.

Anderson, J. R., Bothell, D., Lebiere, C., & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language, 38,* 341–380.

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 25,* 1120–1136.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought.* Mahwah, NJ: Erlbaum.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Anderson, R. B., Doherty, M. E., Berg, N. D., & Friedrich, J. C. (2005).

Sample size and the detection of correlation: A signal detection account. *Psychological Review, 112,* 268–279.

Ayton, P., & Fischer, I. (2004). The hot-hand fallacy and the gambler's fallacy: Two faces of subjective randomness? *Memory & Cognition, 32,* 1369–1378.

Bauer, M. (1972). Relations between prediction- and estimation-responses in cue-probability learning and transfer. *Scandinavian Journal of Psychology, 13,* 198–207.

Bitterman, M. E., Wodinsky, J., & Candland, D. K. (1958). Some comparative psychology. *American Journal of Psychology, 71,* 94–110.

Brunswik, E. (1939). Probability as a determiner of rat behavior. *Journal of Experimental Psychology, 25,* 175–197.

Clément, M., Mercier, P., & Pasto, L. (2002). Sample size, confidence, and contingency judgement. *Canadian Journal of Experimental Psychology, 56,* 128–137.

Conway, A. R. A., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin and Review, 12,* 769–786.

De Houwer, J., & Beckers, T. (2002). A review of recent developments in research and theory on human contingency learning research. *Quarterly Journal of Experimental Psychology, 55B,* 289–310.

Derks, P. L., & Paclisanu, M. I. (1967). Simple strategies in binary predictions by children and adults. *Journal of Experimental Psychology, 73,* 278–285.

Dougherty, M. R. P., & Hunter, J. E. (2003). Hypothesis generation, probability judgment and individual differences in working memory capacity. *Acta Psychologica, 113,* 263–282.

Engle, R. W., Tuholski, S. W., Laughlin, J. E., & Conway, A. R. A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128,* 309–331.

Erev, I. (1998). Signal detection by human observers: A cutoff reinforcement learning model of categorization decisions under uncertainty. *Psychological Review, 105,* 280–298.

Erev, I., & Barron, G. (2005). On adaptation, maximization, and reinforcement learning among cognitive strategies. *Psychological Review, 112,* 912–932.

Estes, W. K. (1976). The cognitive side of probability learning. *Psychological Review, 83,* 37–64.

Estes, W. K., & Straughan, J. H. (1954). Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology, 47,* 225–234.

Fantino, E., & Esfandiari, A. (2002). Probability matching: Encouraging optimal responding in humans. *Canadian Journal of Experimental Psychology, 56,* 58–63.

Gonzalez, C., Lerch, F. J., & Lebiere, C. (2003). Instance-based learning in real-time dynamic decision making. *Cognitive Science, 27,* 591–635.

Goodnow, J. J. (1955). Determinants of choice-distribution in two-choice situations. *American Journal of Psychology, 68,* 106–116.

Herlitz, A., Nilsson, L.-G., & Bäckman, L. (1997). Gender differences in episodic memory. *Memory & Cognition, 25,* 801–811.

Herrnstein, R. J., & Loveland, D. H. (1975). Maximizing and matching on concurrent ratio schedules. *Journal of the Experimental Analysis of Behavior, 24,* 107–116.

Hertwig, R., & Todd, P. M. (2003). More is not always better: The benefits of cognitive limits. In D. Hardman & L. Macchi (Eds.), *Thinking: Psychological perspectives on reasoning, judgment and decision making* (pp. 213–231). Chichester, UK: Wiley.

Hinson, J. M., & Staddon, J. E. R. (1983). Matching, maximizing and hillclimbing. *Journal of the Experimental Analysis of Behavior, 40,* 321–331.

Humphreys, L. G. (1939). Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology, 25,* 294–301.

Irwin, J. R., & McClelland, G. H. (2003). Negative consequences of dichotomizing continuous predictor variables. *Journal of Market Research, 40,* 366–371.

Johnson-Laird, P. N. (1983). *Mental models: Towards a cognitive science of language, inference, and consciousness.* Cambridge, MA: Harvard University Press.

Juslin, P., & Olsson, H. (2005). Capacity limitations and the detection of correlation: Comment on Kareev (2000). *Psychological Review, 112,* 256–267.

Juslin, P., & Persson, M. (2002). PROBabilities from EXemplars (PROBEX): A "lazy" algorithm for probabilistic inference from generic knowledge. *Cognitive Science, 26,* 563–607.

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kane, M. J., & Engle, R. W. (2000). Working-memory capacity, proactive interference, and divided attention: Limits on long-term memory retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 336–358.

Kareev, Y. (1995a). Positive bias in the perception of covariation. *Psychological Review, 102,* 490–502.

Kareev, Y. (1995b). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56,* 263–269.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review, 107,* 397–402.

Kareev, Y. (2004). On the perception of consistency. *Psychology of Learning and Motivation: Advances in Research and Theory, 44,* 261–285.

Kareev, Y. (2005). And yet the small-sample effect does hold: Reply to Juslin and Olsson (2005) and Anderson, Doherty, Berg, and Friedrich (2005). *Psychological Review, 112,* 280–285.

Kareev, Y., Lieberman, I., & Lev, M. (1997). Through a narrow window: Sample size and the perception of correlation. *Journal of Experimental Psychology: General, 126,* 278–287.

Kirkpatrick, S., Gelatt, C. D., Jr., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220,* 671–680.

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working memory capacity?! *Intelligence, 14,* 389–433.

Lewin, C., Wolgers, G., & Herlitz, A. (2001). Sex differences favoring women in verbal and non-verbal, but not in visuospatial episodic memory. *Neuropsychology, 15,* 165–173.

Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review, 95,* 492–527.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 626–636.

Lovett, M. C. (1998). Choice. In J. R. Anderson & C. Lebiere (Eds.), *The atomic components of thought* (pp. 255–296). Mahwah, NJ: Erlbaum.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review, 63,* 81–97.

Miyake, A., Friedman, N. P., Rettinger, D. A., Shah, P., & Hegarty, M. (2001). How are visuospatial working memory, executive functioning, and spatial abilities related? A latent-variable analysis. *Journal of Experimental Psychology: General, 130,* 621–640.

Mutter, S. A., & Williams, T. W. (2004). Aging and the detection of contingency in causal learning. *Psychology and Aging, 19,* 13–26.

Myers, J. L. (1976). Probability learning and sequence learning. In W. K. Estes (Ed.), *Handbook of learning and cognitive processes: Approaches to human learning and motivation* (pp. 171–205). Hillsdale, NJ: Erlbaum.

Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimations in a probability learning situation. *Journal of Experimental Psychology, 57,* 294–298.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–

categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

Nosofsky, R. M., Kruschke, J. K., & McKinley, S. C. (1992). Combining exemplar-based category representations and connectionist learning rules. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 211–233.

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

Overman, W. H. (2004). Sex differences in early childhood, adolescence and adulthood on cognitive tasks that rely on orbital prefrontal cortex. *Brain and Cognition, 55,* 134–147.

Parr, W., & Mercier, P. (1998). Adult age differences in on-line contingency judgments. *Canadian Journal of Experimental Psychology, 52,* 147–158.

Reavis, R., & Overman, W. H. (2001). Adult sex differences on a decision-making task previously shown to depend on the orbital prefrontal cortex. *Behavioral Neuroscience, 115,* 196–206.

Rieskamp, J., Busemeyer, J. R., & Laine, T. (2003). How do people learn to allocate resources? Comparing two learning theories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29,* 1066–1081.

Rieskamp, J., & Otto, P. E. (2006). SSL: A theory of how people learn to select strategies. *Journal of Experimental Psychology: General, 135,* 207–236.

Sanford, A. J., & Maule, A. J. (1973). The concept of general experience: Age and strategies in guessing future events. *Journal of Gerontology, 28,* 81–88.

Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review, 112,* 610–628.

Shaklee, H., & Mims, M. (1982). Sources of error in judging event covariations: Effects of memory demands. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 8,* 208–224.

Shanks, D. R. (1985). Continuous monitoring of human contingency judgment across trials. *Memory & Cognition, 13,* 158–167.

Shanks, D. R. (1987). Acquisition functions in causality judgment. *Learning and Motivation, 18,* 147–166.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15,* 233–250.

Singer, E. (1967). Ability and the use of optimal strategy in decisions. *American Journal of Psychology, 80,* 243–249.

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188.

Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say "broke"? A model of learning the past tense without feedback. *Cognition, 86,* 123–155.

Taatgen, N. A., Lebiere, C., & Anderson, J. R. (2006). Modeling paradigms in ACT-R. In R. Sun (Ed.), *Cognition and multi-agent interaction: From cognitive modeling to social simulation* (pp. 29–52). Cambridge, UK: Cambridge University Press.

Vulkan, N. (2000). An economist's perspective on probability matching. *Journal of Economic Surveys, 14,* 101–118.

Weir, M. W. (1964). Developmental changes in problem-solving strategies. *Psychological Review, 71,* 473–490.

West, R. F., & Stanovich, K. E. (2003). Is probability matching smart? Associations between probabilistic choices and cognitive ability. *Memory & Cognition, 31,* 243–251.

Wilson, W. A., Jr., & Rollin, A. R. (1959). Two-choice behavior of rhesus monkeys in a noncontingent situation. *Journal of Experimental Psychology, 58,* 174–180.

Wolford, G., Miller, M. B., & Gazzaniga, M. (2000). The left hemisphere's role in hypothesis formation. *Journal of Neuroscience, 20* (RC64), 1–4.

Wolford, G., Newman, S., Miller, M. B., & Wig, G. (2004). Searching for patterns in random sequences. *Canadian Journal of Experimental Psychology, 58,* 221–228.

Yechiam. E., & Busemeyer, J. R. (2005). Comparison of basic assumptions embedded in learning models for experience based decision-making. *Psychonomic Bulletin and Review, 12,* 387–402.

Yellott, J. I., Jr. (1969). Probability learning with noncontingent success. *Journal of Mathematical Psychology, 6,* 541–575.

Received December 4, 2005
Revision received March 7, 2006
Accepted March 16, 2006 ■