A Collection of Neuroscience & Cognition Articles

AMERICAN PSYCHOLOGICAL ASSOCIATION

DEAR MEMBER,

At APA, we understand how challenging it can be for you to stay abreast of the latest and best academic research on psychology.

To help you save time, APA is compiling noteworthy articles from a variety of research disciplines. This booklet features some of the most influential scholars and cutting-edge scientific researchers on topics that range from increasing cognitive reserves to an analysis of learning and recall.

When curating this selection of articles for you, APA's scientific staff drew from half a dozen scholarly APA journals and publications that focus on the latest neuroscience and cognition research. APA's Journals Program houses hundreds of academic papers in this field, and you can access them by visiting the journals area on our website at www.apa.org/pubs/journals and browsing by subject: Neuroscience & Cognition.

Moving forward, we will continue to bring you even greater access to cutting-edge research. And we'd like your assistance in this effort: Please share your feedback about this booklet at membership@apa.org. We look forward to hearing from you as we work on your behalf to make your APA membership more valuable and enjoyable.

Sincerely,

Ian King, MBA
Executive Director, Membership
American Psychological Association

CONTENTS

Modeling the Interplay Between Affect and Deliberation

George Loewenstein
Carnegie Mellon University

Ted O'Donoghue
Cornell University

Sudeep Bhatia
University of Warwick

Drawing on diverse lines of research in psychology, economics, and neuroscience, we develop a model in which a person's behavior is determined by an interaction between *deliberative processes* that assess options with a broad, goal-based perspective, and *affective processes* that encompass emotions and other motivational states. Our model provides a framework for understanding many departures from rationality discussed in the literature and captures the familiar feeling of being "of 2 minds." Most important, by focusing on factors that moderate the relative influence of the 2 processes, our model generates a variety of novel testable predictions. We apply our model to intertemporal choice, risky decisions, and social preferences.

Keywords: decision making, dual process, dual system, willpower, intertemporal choice, risk, social preferences

From the writings of the earliest philosophers to the present, there has been an almost unbroken belief that human behavior is best understood as the product of two interacting and often competing processes. Many recent dual process perspectives have focused on the differences between two different modes of thinking—for

example, controlled versus automatic processes (Shiffrin & Schneider, 1977), symbolic and associative processes (Sloman, 1996; Smith & DeCoster, 2000), impulsive and reflective processes (Lieberman, 2003; Strack & Deutsch, 2004), and System I and II (Kahneman & Frederick, 2002). In this article, we also propose a dual-process framework; however, our focus is on choice behavior rather than judgment. Following a long tradition of perspectives drawing a distinction between, for example, "passion versus reason," "the id and the ego," and more recently, "emotion and cognition," we argue that choice behavior can be seen as the product of two motivational processes, one more deliberative and focused on broader goals and the other more reflexive and driven by emotions and other motivational states.

Although both affect and deliberation have been the focus of considerable research, when it comes to formal modeling, one process—the more deliberative of the two—has received the lion's share of attention. Considerable intellectual time and energy has gone into formulating what are sometimes referred to as cognitive or rational-choice models of decision making, such as the expected-utility model and the discounted-utility model. Such models are consequentialist in character; they assume that people

choose between different courses of action based on the desirability of their consequences. Attempts to increase the realism of such models, many associated with the field of behavioral decision research, have generally adhered to the consequentialist perspective but modify assumptions about probability weighting, time discounting, or the specific form of the utility function.

A major reason for this focus is that the other process—affect—has long been viewed as erratic and unpredictable, and hence too complicated to incorporate into formal models. In recent years, however, there has been a renewed interest in emotion, which has revealed a number of systematic properties of both the determinants and consequences of affect. New research by social psychologists (Epstein, 1994; Sloman, 1996; Wilson et al., 2000), neuroscientists (Damasio, 1994; LeDoux, 1996; Panksepp, 1998; Rolls, 1999) and decision researchers (Lerner & Keltner, 2000, 2001; Loewenstein, 1996; Loewenstein et al., 2001; Mellers et al., 1997; Peters & Slovic, 2000; Pham, 1998; Slovic et al., 2002) has led to a better understanding of the role that affect plays in decision making, much of it lending new support to historical dual-process views of human behavior. As of yet, however, there have been few attempts to develop formal models of behavior that incorporate these insights, and in particular to address how affect and deliberation interact to determine human behavior.

We propose a formal dual-process model in which a person's behavior is the joint product of a *deliberative system* that assesses options in a consequentialist fashion and an *affective system* that encompasses emotions such as anger and fear and motivational states such as hunger, sex, and pain. The model provides a new conceptual framework for understanding many of the documented departures from the standard rational-choice model discussed in behavioral decision research, behavioral economics, and judgment and decision making research. At the same time, it captures the familiar feeling of being "of two minds"—of simultaneously thinking one should behave one way while actually behaving in a different way (see, e.g., Milkman et al., 2008). Most important, by focusing on factors that moderate the relative influence of the two processes, the model generates a number of novel testable predictions.

A Dual-Process Model of Behavior

In psychology, the dual-process models that are closest in spirit to our own are Metcalfe and Mischel's hot/cool model (1999) and Fazio and Towles-Schwen's (1999) MODE model. Metcalfe and Mischel (1999) distinguish between a "hot emotional system" and a "cool cognitive system" and assume that a person's behavior depends on which system is dominant at a particular moment. Fazio and Towles-Schwen's (1999) MODE model similarly distinguishes two types of attitude-to-behavior processes, spontaneous processing and deliberative processing, with implicit, automatically activated attitudes guiding spontaneous processing, and explicit attitudes guiding deliberative processing.

Economists, too, have developed dual-process models of human behavior along these lines (Benhabib & Bisin, 2005; Bernheim & Rangel, 2004; Fudenberg & Levine, 2006; Shefrin & Thaler, 1988; Thaler & Shefrin, 1981; and an earlier version of the current article, Loewenstein & O'Donoghue, 2004). While our model overlaps with these models in ways we will discuss, all of these models (except the one on which the current model is based) focus exclusively on intertemporal choice. In this article, we apply our model to a variety of decision-making domains, including intertemporal choice, in which some of our assumptions—particularly affective myopia—overlap with those made by these other economic approaches.

Our model is also informed and motivated by evidence from neuroscience on the functional specificity of different regions of the human brain. Evolutionarily older brain regions, such as the limbic system, which includes areas such as the amygdala and the hypothalamus, evolved to promote survival and reproduction, incorporate affective mechanisms (MacLean, 1990). In contrast, the seemingly unique human ability to choose deliberately, by focusing on broader goals, relies on the prefrontal cortex (Damasio, 1994; Lhermitte, 1986; Miller & Cohen, 2001), the region of the brain that expanded most dramatically in the course of human evolution (Manuck et al., 2003). Indeed, these results have led to dual-system frameworks for the neuroscience of decision making. These focus on the distinction between valuation-based choices and goal-directed choices, with the former being processed primarily in areas such as the

amygdala and the ventromedial prefrontal cortex, and the latter being processed primarily in areas such as the dorsolateral prefrontal cortex (Daw, Niv, & Dayan, 2005; Hare, O'Doherty, Camerer, Schultz, & Rangel, 2008; see also Bechara, Damasio, Tranel, & Damasio, 1997 for an alternate but complementary approach to studying the role of emotions in decision making). Of course, there are many important distinctions between different dual-process accounts, and both the functional and neurobiological properties of these different systems are still up for debate (see, e.g., Kable & Glimcher, 2009).[1]

Our use of the term affect differs from many lay definitions, which tend to focus on the subjective feeling states associated with emotions. In our usage, the defining characteristic is that affects carry "action tendencies" (Frijda, 1986)—for example, anger motivates us to aggress, pain to take steps to ease the pain, and fear to escape (or in some cases to freeze). This perspective is consistent with accounts from evolutionary psychologists (Cosmides & Tooby, 2000), according to which affects are "superordinate programs" that orchestrate responses to recurrent situations of adaptive significance in our evolutionary past (see Loewenstein, 2007 for a discussion of the utility of such a definition).

Our use of the term affect is also related to the distinction between expected emotions and immediate emotions (Loewenstein & Lerner, 2003; Loewenstein et al., 2001; Rick & Loewenstein, 2008). Expected emotions are emotions that are anticipated to occur in the future as a result of decisions but are not experienced in the moment. As expected consequences of decisions, to the extent that they are taken into account, therefore, expected emotions will enter into deliberation. Indeed, one interpretation of the standard consequentialist model of decision making is that people seek to create positive expected emotions and avoid negative expected emotions. Immediate emotions, in contrast, are experienced at the moment of decision and might be completely unrelated to the decision at hand, in which case they are referred to as "incidental" (Bodenhausen, 1993). Perhaps most important, although they are experienced while making a decision, immediate emotions are not affected by the choice that is made, and thus, under the usual rational-choice perspective, should be irrelevant to choices. But numer-

ous studies have found that immediate emotions do influence decision making (Ariely & Loewenstein, 2006; Lerner & Keltner, 2000; Lerner et al., 2004; Raghunathan & Pham, 1999; Wilson & Daly, 2003). A natural interpretation of the affective system in our model is that it captures the influence of immediate emotions.

Finally, our use of the term affect (in contrast to deliberation) can be illustrated by the distinction that Kent Berridge (1996) draws between "wanting" and "liking." Wanting refers to an immediate motivation to acquire something or engage in some activity. Liking, in contrast, refers to how much one actually ends up enjoying the good or activity. Under this interpretation, our affective system makes decisions based on wanting, whereas our deliberative system makes decisions based on liking. Berridge indeed finds that wanting and liking are mediated by different, albeit overlapping, neural systems.

Note that our distinction between affect and deliberation does not imply that basic cognitive processes, such as those involved in object representation, memory, and attention, are absent in affective decision making. It is clear that these processes must play a role in any type of decision. We use the labels deliberation and affect primarily as labels to help organize two different types of motivations. Human behavior is driven by many different motivations in the brain, and restricting attention to two is clearly a simplification. Our point is that it can be a useful simplification to focus on two types of motivations, some that are more reactive and long-term goal-oriented (which we label "deliberation"), and others that are more reflexive and influenced by emotions and short-term drives (which we label "affect").

To formalize our approach, we assume that there are two "objective functions" operating simultaneously. Specifically, consider an individual who must choose an option x out of some choice set X. On one hand, the affective system

---

[1] Throughout this article, we will use findings in neuroscience to motivate our framework of dual-process decision making. However, it is important to note that brain processes do not always map one-to-one onto psychological processes or behaviors. A more rigorous link between these findings, and the framework that we propose in this article, needs to be based on a formal model of the neurobiological basis of decision making (see, e.g., Yechiam, Busemeyer, Stout, & Bechara, 2005, for a discussion).

is motivated to engage in certain behaviors, and we capture these motivations with a motivational function, $M(x, a)$. The variable a captures the intensity of affective motivation. If the affective system alone were completely in charge of behavior, the affective system would "choose" $x^A \equiv \text{argmax}_{x \in X} M(x, a)$, which we refer to as the *affective optimum*. On the other hand, the deliberative system evaluates behavior with a broader and more goal-oriented perspective, and we capture the desirability of actions as perceived by the deliberative system with a utility function, $U(x)$. If the deliberative system alone were completely in charge of behavior, the deliberative system would choose $x^D \equiv \text{argmax}_{x \in X} U(x)$, which we refer to as the *deliberative optimum*. Typically, however, neither system is completely in charge of behavior. Hence, to make predictions, we must incorporate sources of divergence between the two systems, and explain how the two systems interact to generate behavioral outcomes.

Environmental Stimuli

Both systems are influenced by environmental stimuli. In some cases, the two systems will respond to the same stimuli with similar motivational tendencies. For example, during a break at a conference, the availability of a snack might create a surge of hunger in the affective system and be perceived by the deliberative system as a welcome opportunity to recharge before the next session. However, because the two systems operate according to quite different principles, in other situations the same stimulus can influence the two systems differently. If the conferee is on a diet, for example, the availability of the snack might also remind her of that fact, leading to a divergence of affective and deliberative motivation.

Existing research points to a number of factors that influence the strength of affective motivations while affecting the goals of the deliberative system much less if at all. Perhaps most important is the *temporal proximity* of reward and cost stimuli: Affective motivations are intense when rewards and punishments are immediate but much less intense when they are temporally remote. Deliberation is, in contrast, much less sensitive to immediacy. The importance of immediacy for affect has been documented in countless studies. Berns et al. (2006), for example, scanned the brains of subjects as they were waiting to receive electric shocks of different intensities. They found that several affective regions known to respond to the experience of pain (such as the posterior insula, the amygdala, and the caudal anterior cingulate cortex) also responded to the anticipation of pain, and that the activation of these regions increased dramatically as the shock approached in time. Ichihara-Takeda and Funahashi (2006) similarly found that the activity in the orbitofrontal cortex, an area associated with the experience of affective reward, reached its peak immediately prior to the arrival of the reward. In contrast, deliberative areas, such as the dorsolateral prefrontal cortex, did not show this type of time dependence.

In addition to temporal proximity, various forms of *nontemporal proximity* have similar effects (Lewin, 1951). Thus, for example, a tempting snack is more likely to evoke hunger to the extent that it is nearby, visible, or being consumed by someone else. Early evidence on the role of nontemporal proximity comes from a series of classic studies conducted by Walter Mischel and colleagues (see, for instance, Mischel et al., 1972, 1989, 2003). Children were presented with a snack and told they could receive a larger snack if they waited until the experimenter returned. In a baseline treatment, children had the larger delayed snack positioned in front of them as they waited for the experimenter. Relative to this baseline treatment, children were able to delay significantly longer when the larger snack was not present, or even when the larger snack was present but covered. Research on construal-level theory (Trope & Liberman, 2003) also documents a distinction between proximate and nonproximate factors and provides evidence that level of construal plays a role in the relationship between nontemporal proximity and affective responses.

A third factor is the *vividness* of stimuli, by which we mean the ability to conjure the experience in mind. Researchers who study the impact of incidental emotions have become increasingly expert at evoking emotion, and many of the manipulations play on vividness by, for instance, showing people movies of an emotion-evoking event (Lerner et al., 2004), having people write essays in which they imagine themselves in a situation (Lerner & Keltner, 2000), playing music (Blood & Zatorre, 2001; Halberstadt & Niedenthal, 1997), or even through the

artful use of odors (Ditto et al., 2006; Zald & Pardo, 1997). The ability to evoke emotion through vividness suggests that vividness of different choice object and different experiences may play a crucial role in driving the responses of the affective system (see, however, Taylor & Thompson, 1982, for a discussion of limits on the impact of vividness on judgment).

To incorporate these three effects into our model, the motivational function, $M(x,a)$, incorporates a variable a that captures the *intensity* of affective motivations. In general, the larger is a, the stronger will be the affective motivations. In abstract terms, if the affective system prefers an option x over an option x', then an increase in affective intensity increases affective motivation in the sense that the difference $M(x, a) - M(x',a)$ increases with a. For some choice problems, there will be competing affective motivations, in which case a should be thought of as a vector of good-specific affective intensities. For instance, if one must make trade-offs between money and cookies, it would be natural to assume that $a = (a_M, a_C)$, where a_M is affective intensity for money, and a_C is affective intensity for cookies. Each affective intensity influences the motivation for its associated good.[2]

Behavioral Outcomes

A range of evidence suggests that the affective system holds a primacy in determining behavior—that is, the affective system has default control of behavior, but the deliberative system can step in to exert its influence as well. For instance, Joseph LeDoux and his colleagues (LeDoux, 1996) have demonstrated that fear responses are influenced by two separate neural pathways from the sensory thalamus to the amygdala (a lower-brain structure that plays a critical role in fear responses). One pathway goes directly from the sensory thalamus to the amygdala, and the second goes first from the sensory thalamus to the neocortex and from there to the amygdala. Moreover, they also discovered that the direct pathway is about twice as fast as the indirect pathway. As a result, rats can have an affective reaction to a stimulus before their cortex has had the chance to perform more refined processing.

When deliberation gets involved, what determines the extent to which it influences behavior? There is, in fact, compelling evidence that

deliberation does not easily take full control. Rather, when in conflict with affect, deliberative control, to the extent that it is possible, requires an expenditure of effort. The most important evidence along these lines comes from research by Baumeister and colleagues on willpower (for a summary, see Baumeister & Vohs, 2003), by which they mean an inner exertion of effort required to implement some desired behavior. Their basic contention is that such willpower is a resource in limited supply (at least in the short run), and that depletion of this resource by recent use will reduce a person's ability to implement desired behaviors. Baumeister's basic willpower paradigm involves having subjects carry out two successive, unrelated tasks that both require willpower and comparing the behavior on the second task to that of a control group that had not performed the first task. The general finding is that exerting willpower in one situation tends to undermine people's propensity to use it in a subsequent situation. In one representative study, for example, subjects who sat in front of a bowl of cookies without partaking subsequently gave up trying to solve a difficult problem more quickly than did subjects who were not first tempted by the cookies.

Because the target behaviors in Baumeister's studies— for example, not eating cookies or trying to solve a difficult puzzle—typically involve pursuit of broader goals, whereas not doing these behaviors typically involves indulging affective motivations, we believe there is a natural interpretation of these results for our model. Specifically, it is attempts by the deliberative system to override affective motivations that require an inner exertion of effort or willpower. Subsequently, if a person's willpower is depleted by recent use, the deliberative system will have less influence over behavior. Consistent with this view, a related line of research shows that simply making decisions can undermine willpower (Baumeister & Vohs, 2003).

Hence, one situation in which affect will have more sway over behavior is when the delibera-

---

[2] Both here and in the section on risky decision making, we talk in terms of good-specific affective intensities. This language, however, should be viewed as a shorthand for an underlying model with (a) multiple types of affects (e.g., hunger, greed, fear, etc.) and (b) different types of goods that are differentially affected by different types of affect.

tive system is "worn out" from past willpower use. A second, related, situation is when the deliberative system is currently occupied by unrelated cognitive tasks. Research has shown that having subjects perform simple cognitive tasks—an intervention labeled "cognitive load"—undermines efforts at self-control. In one study, Shiv and Fedorikhin (1999) had subjects memorize either a 7-digit number (high cognitive load) or a 2-digit number (low cognitive load) before presenting them with a choice between cake (a high-calorie food) and fruit (a low-calorie food). Fifty-nine percent chose the cake in the high-load condition, but only 37% in the low-load condition.

To formalize these ideas, we assume that the deliberative system makes the final choice, but it must make this choice subject to having to exert effort—willpower—to control affective motivations. We capture this cognitive effort by assuming that, to induce some behavior different from the affective optimum (i.e., to choose an $x \neq x^A$), the deliberative system must exert an effort cost, in utility units, of $h(W, \sigma) * [M(x^A, a) - M(x, a)]$. This formulation assumes that the further the deliberative system moves behavior away from the affective optimum, the more willpower is required. The factor $h(W, \sigma) > 0$ represents the cost to the deliberative system of mobilizing willpower—that is, the higher is $h(W, \sigma)$, the larger is the cognitive effort required to induce a given deviation from the affective optimum.

Based on our discussion, we incorporate two factors that make it more costly for the deliberative system to exert willpower. The first is the person's current *willpower strength*, which we denote by W. This variable is meant to capture the current stock of willpower reserves; we assume that h is decreasing in W, so that as one's willpower strength is depleted the deliberative system finds it more difficult (more costly) to influence the affective system. Our analysis in this article will focus on one particular implication with regard to willpower strength: The more willpower a person has used in the recent past, the more her current willpower strength will be depleted, and hence exerting willpower becomes more costly. The second factor that makes it more costly for the deliberative system to exert willpower relates to competing cognitive demands (such as those induced by cognitive load), which we denote by σ. Thus, we will assume that h is increasing in σ: If a person's deliberative system is distracted by unrelated cognitive tasks, exerting willpower becomes more costly.

General Implications

We now combine the elements of our formalization to derive general implications of our model. To make a choice, the deliberative system trades off the desirability of actions—as reflected by its utility function $U(x)$—against the willpower effort required to implement them. Hence, the deliberative system will choose the action $x \in X$ that maximizes $U(x) - h(W, \sigma) * [M(x^A, a) - M(x, a)]$. Because the affective optimum x^A is not affected by the person's actual choice, this is identical to maximizing:

$$V(x) \equiv U(x) + h(W, \sigma) * M(x, a) \qquad (1)$$

It follows that the person will choose an option that is somewhere in between the deliberative optimum and the affective optimum (when x is a scalar, either $x^D \geq x \geq x^A$ or $x^A \geq x \geq x^D$). Exactly where behavior falls will depend on the cost of mobilizing willpower as captured by $h(W, \sigma)$. As the cost of willpower decreases, behavior will be closer to the deliberative optimum, and as it increases, behavior will be closer to the affective optimum.

Although we interpret our model as reflecting that the deliberative system chooses behavior subject to willpower costs, there is a second interpretation of our model that is more consistent with our discussion of affective primacy. Because the deliberative optimum x^D and the affective optimum x^A, are not affected by the person's actual choice, maximizing $V(x)$ is equivalent to minimizing $[U(x^D) - U(x)] + h(W, \sigma) * [M(x^A, a) - M(x, a)]$. Hence, our model can be interpreted as the minimization of a weighted sum of two costs: a cost to the deliberative system from not getting its optimum x^D, and a cost to the affective system from not getting its optimum x^A. In this interpretation, $h(W, \sigma)$ captures the relative weights of the two systems.

While we have motivated our model as a dual-process approach, in the end behavior is determined by a single "objective" function, $V(x)$. What is the value, then, of the dual-process approach? One way in which the dual-

process approach is useful is that it provides a natural interpretation of many behavioral outcomes. When evaluating risky prospects, people might cognitively believe that they should weight probabilities linearly, but then make choices that reflect an insensitivity to probabilities. When weighing some intertemporal indulgence such as a tasty but highly caloric morsel or a willing but forbidden sexual partner, people might cognitively think that the indulgence is not worth the future costs, but then indulge nonetheless. Our model provides a natural interpretation: People's beliefs for what they ought to do reflect only the objectives of the deliberative system, whereas actual behavior is influenced by affective motivations as well. In other words, many deviations from the standard prescriptive models of decision making can be interpreted as coming from the motivations of the affective system.

A second way in which the dual-process approach is useful is that it provides a template for interpreting research from neuroscience. Recent research in neuroscience, particularly in the subdiscipline of neuroeconomics, often focuses on where we see brain activity when people make decisions. And, while neuroscientists are often interested in more fine partitions, a frequent focus is on the extent to which activity occurs in the prefrontal cortex or in evolutionarily older brain systems, such as the amygdala, the hypothalamus, and other parts of the limbic system. To the extent that our deliberative system is roughly meant to capture activity in the prefrontal cortex whereas our affective system is roughly meant to capture activity in the evolutionarily older brain systems, according to our model such research can be used to shed insight on the different objectives of the two systems. Indeed, we have already discussed some neuroscientific research in this way, and do so further in the discussion of specific applications.[3]

But perhaps the most important value of the dual-process approach is that it generates testable predictions. These predictions are perhaps most clear when a person faces a binary choice between two options. Suppose a person is choosing between an option x and an option x', where option x is the deliberative optimum (i.e., $U(x) > U(x')$). According to our model, the person will choose the former when $U(x) + h(W, \sigma)M(x, a) > U(x') + h(W, \sigma)M(x', a)$, or $U(x) - U(x') > h(W, \sigma)[M(x', a) - M(x, a)]$.

First note that if option x is also the affective optimum, i.e., $M(x, a) > M(x', a)$, then the person will clearly choose option x. Hence, assume instead that option x' is the affective optimum, i.e., $M(x', a) > M(x, a)$. From the inequality, two general predictions follow:

General Prediction #1: If a person faces a binary choice between options x and x' where option x is the deliberative optimum while option x' is the affective optimum, then willpower depletion or unrelated cognitive demands such as cognitive load increase the cost of exerting willpower [increase $h(W, \sigma)$] and therefore make it less likely that the person chooses the deliberative optimum (option x).

General Prediction #2: If a person faces a binary choice between options x and x' where option x is the deliberative optimum and option x' is the affective optimum, then if increased affective intensity increases the affective preference for option x' over option x [i.e., if increased a increases the difference $M(x', a) - M(x, a)$], then affective intensity makes it less likely that a person chooses the deliberative optimum (option x). If, instead, affective intensity decreases the affective preference for option x' over option x [i.e., if increased a decreases the difference $M(x', a) - M(x, a)$], then an increase in affective intensity makes it more likely that a person chooses the deliberative optimum (option x).

In the next three sections, we apply our model to three specific domains: intertemporal choice, risky decision making, and social preferences. In each domain, we make specific assumptions about the objectives of the two systems and use these to derive specific predictions of our model. In some cases, we find existing evidence that supports these predictions, but in others we propose them as testable, but as yet untested, predictions of the model.

Intertemporal Choice

The most straightforward application of our model is to intertemporal choices—decisions that involve tradeoffs between current and future outcomes. Suppose that each option x in the choice set X generates a stream of payoffs $x_1, x_2,$

---

[3] Note of course that many fMRI studies in neuroscience are correlational, and that activation in a particular brain area cannot always justify inferences regarding the underlying psychological processes at play in observed behavior. That said, we believe that our approach is a desirable first step in incorporating neuroscientific research into the study of emotion and deliberation in preferential choice, and that much of this research serves as a valuable complement to the psychological and behavioral findings that we discuss in this article.

..., x_T, where payoff x_t is received in period t, and all payoffs involve the same type of choice option. For simplicity, and for comparison to standard approaches in economics, we assume that both the affective and the deliberative systems display standard exponential discounting. However, we additionally assume that the affective system is more myopic than the deliberative system (and sometimes consider the special case where the affective system cares only about immediate outcomes), and we also assume that increased affective intensity makes the affective system more myopic.[4]

Formally, we assume that the deliberative system's utility function is $U(x) = x_1 + \delta_D x_2 + \ldots + [\delta_D]^T x_T$; that is, exponential discounting with discount factor δ_D. The affective system's motivational function is $M(x, a) = x_1 + \delta_A(a)x_2 + \ldots + (\delta_A(a))^T x_T$; that is, exponential discounting with discount factor $\delta_A(a)$. We further assume that $\delta_A(a) < \delta_D$, and that increased affective intensity a implies a smaller $\delta_A(a)$ and thus more myopia. Putting these together, the decision maker will choose x to maximize:

$$V(x) = [x_1 + \delta_D x_2 + \ldots + [\delta_D]^T x_T] + h(W, \sigma)$$
$$*[x_1 + \delta_A(a)x_2 + \ldots + [\delta_A(a)]^T x_T]$$

(2)

Our assumption that the affective system is driven primarily by short-term payoffs, whereas the deliberative system cares about both short-term and longer-term payoffs is similar to that made by existing dual-process theories of intertemporal choice in economics (Benhabib & Bisin, 2005; Bernheim & Rangel, 2004; Fudenberg & Levine, 2006; Shefrin & Thaler, 1988; Thaler & Shefrin, 1981). There is considerable evidence in support of this assumption. On the deliberative side, Frederick (2003) asked subjects how they believed they should respond to outcomes occurring at different times, and most people generally believed that time discounting is not normatively justified—that outcomes should receive the same weight regardless of when they occur. This suggests that people perceive their own impulsivity as contrasting with what they believe to be reasonable.

On the affective side, when animals are presented with intertemporal choices, they are extremely myopic. There is a long literature that demonstrates extreme myopia in pigeons and rats. Indeed, it has been found that species of New World monkeys are willing to wait less than 20 s for a food reward that is three times as large (Stevens et al., 2005). Monkeys that are closer, evolutionarily, to humans show less although by human standards still extreme levels, myopia (Tobin et al., 1996). In a related vein, children have been shown to be more myopic than adults, with children and teenagers exhibiting much steeper discount functions that individuals in their 20s and 30s (Steinberg et al., 2009). To the extent that animal and child behavior can be used to shed insight on the motivations of humans' affective system, this evidence suggests that the affective system is myopic, and that concern for longer-term outcomes are a product of the deliberative system.

More convincing evidence comes from neuroscience. McClure et al. (2004, 2006) scanned subjects' brains using fMRI while they made choices between smaller-sooner rewards versus larger-later rewards. All of these choices produced activation in prefrontal regions associated with deliberation (such as the dorsolateral and ventrolateral prefrontal cortex); however, when one of the options involved an immediate reward, brain regions associated with affective processing, such as the ventral striatum and medial orbitofrontal cortex, also became activated. Moreover, in situations in which an immediate reward was one of the options, higher relative activation of the affective regions increased the likelihood that the subject would choose the immediate reward.

Similar results are suggested by Bjork et al. (2009), who found that delay discounting can be predicted by the size of the decision maker's lateral prefrontal cortex. Figner et al. (2010) also found that experimentally disrupting prefrontal areas associated with deliberation (particularly the lateral prefrontal cortex) led to an increased choice of immediate rewards over delayed rewards. This disruption did not, however, alter choices between delayed rewards, suggesting that deliberative processing plays a

---

[4] Our key predictions, I-1 and I-2, rely only on the assumption that the affective system is more myopic than the deliberative system, and not on the assumption of exponential discounting. We assume exponential discounting to highlight how our framework can give rise to hyperbolic discounting, even if neither system exhibits hyperbolic discounting (as we discuss later).

fundamental role in directing nonmyopic choice.

Last, considerable research on addiction and self-control has documented a discrepancy between an addict's short-term desires (involving, e.g., the consumption of an addictive substance), and an addict's long-term goals (which seek to regulate cravings and stop the use of these addictive substances; see, e.g., Goldstein, 2001 for a discussion). This pattern of behavior strongly supports the assumptions of affective myopia and deliberative far-sightedness that we propose in this article.

Equation 2 yields several important predictions. First, maximizing Equation 2 is equivalent to maximizing $\tilde{V}(x) = x_1 + \sum_{t=1}^{\infty} D(t)x_{1+t}$ with:

$$D(t) = \frac{(\delta_D)^t + h(W, \sigma)(\delta_A(a))^t}{1 + h(W, \sigma)}.$$

Note that $D(t)$ is a discount function reflecting the discounting associated with a payoff with delay t. This formulation (with $\delta_D > \delta_A(a)$) implies both discounting (i.e., that $D(0) = 1 > D(1) > D(2) \ldots$) and declining discount rates (i.e., $D(0)/D(1) > D(1)/D(2) > D(2)/D(3) \ldots$). In addition, in the special case where $\delta_A(a) = 0$, maximizing Equation 2 is equivalent maximizing $x_1 + \beta\delta x_2 + \ldots + \beta\delta^T x_T$, where $\beta = 1/(1+h(W, \sigma)) < 1$. This is the well-known beta-delta function used by Laibson (1997) and others, as an analytical tractable simplification of hyperbolic discounting.[5]

Hence, our model, with the assumption that affective discounting provides a natural interpretation—or reinterpretation—of (quasi) hyperbolic discounting. Specifically, even if the deliberative system discounts exponentially, because behavior is also influenced by a more myopic affective system, people will be more impatient when facing now versus near-future trade-offs than they will be when facing future versus further-future trade-offs—which is the essence of hyperbolic discounting. This formulation also implies that a decrease in willpower, increase in cognitive demands, or increase in affective intensity will lead to a higher value of β without changing the effective δ. The quasi-hyperbolic form defined here is consistent with a number of intertemporal preference reversals (e.g., Ainslie, 1975; Kirby, 1997), with de-

clining (average) discount rates (e.g., Ben-Zion, Rapoport, & Yagil, 1989; Thaler, 1981), as well as with evidence (e.g., Frederick et al., 2002) suggesting that the magnitude of discounting is based on the distinction between now and the future—and in particular, that people exhibit nearly constant discounting when facing two future trade-offs.

Beyond providing an alternative account of hyperbolic time discounting, Equation 2 also generates testable predictions by applying the two general predictions of our model:

> *Intertemporal Choice Prediction #1 (I-1):* An increase in $h(W, \sigma)$ will lead to more myopic behavior.

> *Intertemporal Choice Prediction #2 (I-2):* Any factor that increases the intensity of the affective motivation for the immediate payoff will lead to more myopic behavior.

The increases to myopic behavior listed in Predictions I-1 and I-2 will affect choice only when the decision involves tradeoffs between immediate and future payoffs. Willpower, cognitive load, or affective intensity will not alter tradeoffs involving two or more future payoffs. In addition, note that Predictions I-1 and I-2 also hold for the more general model, which allows the affective system to discount exponentially (but with a lower discount factor than that displayed by the deliberative system).

There is existing evidence on Predictions I-1 and I-2. For instance, Vohs and Heatherton (2000) investigated how willpower depletion affects the amount of ice cream people eat when asked to taste and rate three flavors. To the extent that eating ice cream involves immediate benefits and future costs, eating more ice cream can be taken to reflect increased myopia. In support of I-1, they found that, among dieters, willpower depletion led subjects to eat more ice cream. However, they found no effect among nondieters. In addition, Vohs and Faber (2007) found that willpower depletion led to increased impulse buying, and Vohs et al., (2008) found

---

[5] Mathematically hyperbolic discounting is described with the discount factor $1/(1+kt)$. The beta-delta approximates this formulation in discrete time. For β and δ between 0 and 1, the decision maker will be present biased when choosing between immediate and delayed rewards, but will discount exponentially when choosing between different delayed rewards. Note that some scholars have argued against discounting models, in favor of attribute tradeoff models (Scholten & Read, 2010).

that willpower depletion increased procrastination. Finally, more direct evidence of the impact of willpower depletion on delay discounting is documented by Vohs et al. (2013). Individuals who performed depletion tasks prior to making intertemporal choices were more likely to choose smaller, immediate rewards over larger, delayed rewards.

The Shiv and Fedorikhin (1999) study earlier in this article provides support for the cognitive demands Prediction of I-1—specifically, cognitive load makes subjects more prone to choose cake over fruit, reflecting increased myopia. Benjamin, Brown, and Shapiro (2013) provide more direct evidence. They asked Chilean high school juniors to make a series of short-term trade-offs and long-term trade-offs for monetary payoffs. Relative to control subjects, subjects who answered these questions while under cognitive load showed nontrivial reductions in short-term patience. In contrast, cognitive load had no effect on long-term patience.

I-2 captures a host of predictions based on the different factors, discussed above, that increase affective intensity. Most straightforwardly, our model predicts that nontemporal proximity of immediate outcomes should play a large role in elicited discount rates. Thus, for example, the extent that an immediate reward can be seen or smelled will affect the magnitude of discount rates that people's behavior reveals, which is consistent with the research by Mischel and colleagues described earlier in this article. Note that Mischel's results are puzzling when viewed from the perspective of hyperbolic discounting. As time passes, and thus the delay between the immediate smaller snack and the delayed larger snack shrinks, children become less willing to wait (which is why many children initially decide to wait, but then "bail out")— exactly the opposite of what hyperbolic discounting would predict. Willpower depletion, however, provides a natural explanation. Specifically, as time passes and the person's willpower is slowly depleted, eventually they no longer have enough willpower to support further delay.

Indeed, our framework provides a natural formalization of this behavioral pattern. Specifically, let τ denote the time for which a child has been waiting, let $W(\tau)$ denote the willpower remaining at time τ, and make the natural as-

sumption that have $dW/d\tau < 0$—because waiting takes willpower and thus willpower declines over time. Letting x denote the deliberative optimum (waiting) and x' denote the affective optimum (getting the snack now), the person will wait only if $U(x) - U(x') > h(W(\tau), \sigma)$ $[M(x', a) - M(x, a)]$. As time passes (τ increases) and willpower depletes ($W(\tau)$ declines), this condition becomes less and less likely to hold.

Our framework can similarly explain why decision makers are more likely to succumb to temptation when they are repeatedly confronted with tempting choices. Not only are temptation and willpower likely to fluctuate over time, allowing for more opportunities for temptation to overcome willpower, but also because resisting temptation depletes willpower, and doing so repeatedly depletes it proportionately.

Giordano et al. (2002) provide additional evidence in support of I-2. They measured the time discounting of heroin addicts for both money and heroin, both when the addicts were satiated (after they had received treatment with an opioid agonist) and when they were deprived (before receiving treatment). They observed greater time discounting for heroin than for money, and greater discounting of both types of reward when the addicts were opioid-deprived than when they were satiated. Johnson et al. (2007) similarly found that smokers discount cigarettes more than they discount money or health, and Rosati et al. (2007) found that individuals are more impatient for food relative to money. These results are consistent with our framework as long as heroin, food and cigarettes have higher affective intensity than money (and have even higher affective intensity when decision makers are in a state of craving or hunger).

Finally, I-2 predicts that people who have particularly strong affective reactions to stimuli will exhibit more myopic behavior. In fact, direct support for this prediction comes from research by Hariri et al. (2006), who found that people who exhibited larger affective reactions to random monetary gains and losses in one experimental session (as measured by neural activation in the ventral striatum) also showed increased myopia when trading off

immediate versus future monetary payoffs in a different experimental session.

Risky Decision Making

A second natural application of our model is to choices between risky prospects. To apply our two-system approach to risky decision making, we must make assumptions about how the two systems respond to risks. For the deliberative system, a natural assumption is that risks are evaluated according to their expected utility (or perhaps expected value). Indeed, most researchers, as well as knowledgeable lay people, agree that expected-utility theory is the appropriate prescriptive theory to use for evaluating risks (for a discussion, see Bleichrodt et al., 2001). It is less obvious what drives the affective system, but we suggest that insensitivity to probabilities and loss aversion—two prominent features in many descriptive theories of risk preferences (Starmer, 2000)—derive from the affective system.

Suppose that each option x in the choice set X is a lottery $x \equiv (x_1, p_1; \dots; x_N, p_N)$, where outcome x_i occurs with probability p_i. We assume that the deliberative system's utility function is $U(x) = \sum p_i u(x_i)$. In some subsequent analyses, we assume that $u(x_i) = x_i$ (i.e., the deliberative system cares about expected value) whenever choosing between monetary gambles. This does not affect any of our results; it only makes it easier to illustrate the effects of incorporating the affective system into a model of risky choice.

The affective system, in contrast, has motivational function $M(x, a) = \sum w(p_i) v(x_i, a)$, where $w(p_i)$ is a nonlinear probability-weighting function, and $v(x_i, a)$ is a value function that incorporates loss aversion. For simplicity, we assume the value function is $v(x_i, a) = au(x_i)$ if x_i is a gain, and $v(x_i, a) = a \lambda u(x_i)$ if x_i is a loss, where the variable $\lambda > 1$ reflects the degree of loss aversion. For the probability-weighting function, many of our results don't require a specific assumption, and thus we often use a generic $w(p)$. However, we believe that the key feature is that the affective system is less sensitive to probabilities than the deliberative system, that is, $dw/dp < 1$ for $p \in (0, 1)$. In our analysis, we sometimes use the specific exam-

ple of $w(p) = c + bp$ for all $p \in (0, 1)$, $w(0) = 0$, and $w(1) = 1$, where $c > 0$ and $c < 1 - b$.[6]

Incorporating these functions into Equation 1, the person will choose the option x that maximizes

$$V(x) = \sum p_i u(x_i) + h(W, \sigma) * \left[\sum w(p_i) v(x_i, a) \right].$$

(3)

The assumptions underlying Equation 3 come from diverse lines of research from a range of disciplines and resemble some of the assumptions made in Mukherjee (2010). There is strong physiological evidence that supports our contention that the affective system exhibits insensitivity to variation in probabilities. Studies that measure fear by means of physiological responses such as changes in heart rate and skin conductance—which primarily reflect activity in the affective system—find that reactions to an uncertain impending shock depend on the expected intensity of the shock but not the likelihood of receiving it (except if it is zero; Bankart & Elliott, 1974; Deane, 1969; Elliott, 1975; Monat et al., 1972; Snortum & Wilding, 1971). Other evidence supports the idea that emotional responses result largely from *mental images* of outcomes (Damasio, 1994). Because such images are largely invariant with respect to probability—one's mental image of winning a lottery, for example, depends a lot on how much one wins but not that much on one's chance of winning—emotional responses tend to be insensitive to probabilities.

There is also evidence that supports our contention that loss aversion derives from the affective system. For instance, Chen et al. (2006) introduced a currency into a colony of capuchin monkeys, presented the monkeys with gambles, and found that the monkeys displayed loss aversion. To the extent that animal behavior is indicative of the output of the human affective system, this result suggests that loss aversion, and the behaviors that it generates, derives from the affective system. Of course, there are many other differences between human and animal risk taking that are attributable to factors other

---

[6] Note that our assumption $dw/dp < 1$ for all $p \in (0, 1)$ will require a discontinuity at $p = 0$ and at $p = 1$, much as was suggested in the original version of prospect theory (Kahneman & Tversky, 1979).

than affective strength. For example, Weber et al. (2004) found that some differences between human and animal behavior in the domain of risk disappear when differences in reward learning are controlled for.

There is also neuroscientific evidence. Tom et al. (2007) collected fMRI data while subjects decided whether to accept or reject gambles that involved a chance to win or lose various amounts of money. The gambles differed in the magnitudes of the gains and losses, and the researchers found that affective regions, such as the striatum and medial orbitofrontal cortex, react to these changes. Moreover, these regions display a neural loss aversion: The increase in activity when the gain amount increases is smaller than the decrease in activity when the loss amount increases. Similar results have also been documented by Weber et al. (2007), who found that the amygdala is differentially active when decision makers are parting with goods. In addition, Sokol-Hessner, Camerer, and Phelps (2013) found that the reappraisal of choices involving loss aversion generates increased activity in the dorsolateral and ventrolateral prefrontal cortex and reduced activity in the amygdala. Regulating loss aversion, according to this research, involves the suppression of emotion by the deliberative system.

The relationship of affect with loss aversion has also been shown to be responsible for non-risky reference dependence anomalies, such as the endowment effect. Particularly, Knutson et al. (2008) found that activity in limbic system regions, such as the nucleus accumbens, which play an important role in loss averse behavior, also predict individual susceptibility to the endowment effect. Individuals who showed increased affective sensitivity to losses were also most likely to display discrepancies between acceptable buy and acceptable sell prices.

Another piece of neuroscientific evidence for the role of affect in loss aversion comes from a study by Shiv et al. (2003), who compared healthy people; patients with brain lesions in regions related to emotional processing, such as the amygdala and the orbitofrontal cortex (they were normal on most cognitive tests, including tests of intelligence); and patients with lesions in regions unrelated to emotion. Patients with emotion-related lesions were more likely to select risky gambles (involving losses) than other subjects—that is, they exhibited less loss aversion—and ul-

timately earned more money, suggestive of the idea that the emotional processing regions that were damaged play a role in loss aversion. Moreover, whereas normal people and patients with lesions unrelated to emotion were influenced by their outcomes in previous rounds, patients with emotion-related lesions were not. These results have also been documented by De Martino, Kumaran, Seymour, and Dolan (2010). De Martino and coauthors estimated loss aversion coefficients for two individuals with amygdala damage. Using a series of gambles with gains and losses ranging from $20 to $50, they found that estimated loss aversion coefficients for the two patients were very close to one, indicating an absence of loss aversion.

Predictions

The general model presented earlier will yield predictions that reflect three rough intuitions. First, because insensitivity to probabilities and loss aversion derive from the affective system, willpower depletion or unrelated cognitive demands, such as cognitive load, will magnify these behavioral tendencies. Second, if a person faces a choice between lotteries for which all outcomes involve the same type of good and thus the same affective intensity, then an increase in that affective intensity will also magnify insensitivity to probabilities and loss aversion. Finally, if a person faces a choice between lotteries that involve different goods and thus different affective intensities, then the effects of affective intensity are good-specific. To translate these rough intuitions into specific predictions, we apply our model, as specified in Equation 3, to specific risky choices.

Monetary Certainty Equivalent for Monetary Gambles

Suppose a person faces a simple gamble (Z, p; $0,1–p$) with $Z > 0$, and we elicit the person's monetary certainty equivalent—that is, the certain amount CE such that the person is indifferent between the gamble and that certain amount. Tests for nonlinear probability weighting often focus on these types of choices, and in particular on how overweighting of small probabilities should lead to $CE > pZ$, whereas underweighting of large probabilities should lead to $CE < pZ$.

This type of choice has two simplifying features. First, because all outcomes involve a monetary payoff, the same affective intensity for money, which we denote by a_M, is applied to all outcomes. Second, because $Z > 0$ and therefore $CE > 0$, and because we assume for monetary outcomes that $u(x) = x$, we can ignore loss aversion, and the value function merely becomes $v(x_i, a) = a_M x_i$. Hence, according to our model, the monetary certainty equivalent is determined by $CE + h(W, \sigma)[a_M CE] = pZ + h(W, \sigma)[w(p)a_M Z]$, which yields that $CE = \tilde{w}(p)Z$ where

$$\tilde{w}(p) = \begin{cases} 0 & if\ p = 0 \\ \dfrac{p + h(W, \sigma)\,a_M\,w(p)}{1 + h(W, \sigma)\,a_M} & if\ p \in (0, 1) \\ 1 & if\ p = 1 \end{cases}$$

Much as in expected utility and prospect theory, the certainty equivalent is derived from multiplying the magnitude of the outcome by a weight that is a function of the probability of that outcome. However, the probability weighting function for each of the three models is different. Under expected utility with linear utility for money, $CE = pZ$ (i.e., linear weighting of probabilities), and under prospect theory with a linear value function in the gain domain, $CE = \tilde{w}(p)Z$, where $\tilde{w}(p)$ is prospect theory's probability-weighting function. Figure 1 presents an example of an effective weighting function implied by our model when the affective system's weighting function $w(p)$ is assumed to be linear with a positive intercept and a slope less than 1 (reflecting an insensitivity to probability changes). In particular, it depicts (a) the weight used by the deliberative system (p), (b) the weight used by the affective system, $w(p)$, and (c) the effective weight used for decisions, $\tilde{w}(p)$. Notice that our model is closer in spirit to Kahneman and Tversky's original formulation in being ill defined at the extremes (in fact, Barseghyan et al., 2013 estimates probability weighting from data on insurance deductible choices and seems to find support for Kahneman and Tversky's original formulation).

Like prospect theory, if affective motivations generate an overweighting of small probabilities and underweighting of large probabilities, as reflected in $w(p)$, then our model predicts $CE > pZ$ for $p < \pi$ and $CE < pZ$ for $p > \pi$. However, unlike expected utility and prospect theory, which assume fixed probability weighting functions, our model generates novel testable predictions for factors that should alter probability weighting and hence the certainty equivalent:

Risky Choice Prediction #1 (R-1): When generating a certainty equivalent for simple monetary gambles, an increase in $h(W, \sigma)$ will increase CE when $CE > pZ$ and decrease CE when $CE < pZ$.

Risky Choice Prediction #2 (R-2): When generating a certainty equivalent for simple monetary gambles, an increase in the intensity of affective motivation for
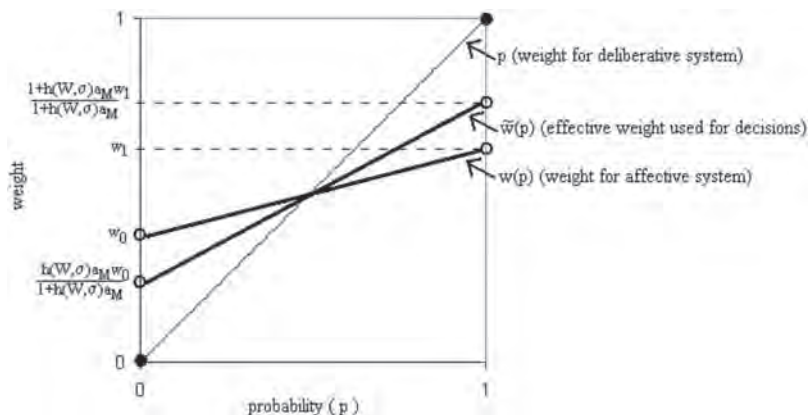

Figure 1. Effective probability weighting function $\tilde{w}(p)$ predicted by our model for certainty equivalents for monetary gambles when the affective system has probability weighting function $w(p) = w_0 + (w_1 - w_0) * p$ with $w_0 > 0$ and $w_1 < 1$.

money will increase CE when $CE > pZ$ and decrease CE when $CE < pZ$.

Intuitively, because deliberation argues for $CE = pZ$, if $CE > pZ$ then the affective system is dragging CE upward, and therefore when willpower depletion, cognitive load, or affective intensity for money give more sway to affect, it will drag CE further upward. Analogously, if $CE < pZ$, then the affective system is dragging CE downward, and therefore when willpower depletion, cognitive load, or affective intensity give more sway to affect, it will drag CE further downward. Hence, our model generates sharp predictions for these simple decisions; unfortunately, we know of no existing evidence on such effects.

Monetary Certainty Equivalent for Simple Nonmonetary Gambles

Suppose a person faces a simple gamble (x, $p;0$, $1 - p$), where x is a nonmonetary good such as a plate of cookies, and again we elicit the person's monetary certainty equivalent for this gamble. Because this choice involves two distinct goods—for example, money versus cookies—we must distinguish between affective intensity for money, a_M, and affective intensity for x, which we denote by a_x. According to our model, the monetary certainty equivalent is determined by $CE + h(W, \sigma)[a_M CE] = pu(x) + h(W, \sigma)[w(p)a_x u(x)]$, which yields that $CE = \tilde{w}(p)u(x)$ where

$$\tilde{w}(p) = \begin{cases} 0 & if\ p = 0 \\ \dfrac{p + h(W, \sigma)\, a_X\, w(p)}{1 + h(W, \sigma)\, a_M} & if\ p \in (0, 1) \\ 1 & if\ p = 1 \end{cases}$$

Figure 2 depicts the effective weighting function $\tilde{w}(p)$ here using the same affective system's weighting function $w(p)$ from Figure 1. While the effective weighting function here has the same qualitative shape as that in Figure 1, there is one important difference: whereas in Figure 1, $\tilde{w}(p) < p$ for p close enough to one, in Figure 2, it is possible to have $\tilde{w}(p) > p$ for all $p < 1$. Intuitively, there are two forces at work. First, just as for the certainty equivalent for simple monetary gambles, the affective system overweighs small probabilities, which tends to drag the

CE upward, and the affective system underweights large probabilities, which tends to drag the CE downward. Second, and unique for this case, affective intensity for the nonmonetary good might be larger than affective intensity for money, which tends to drag the CE upward. For low probabilities, these two effects reinforce each other, and thus affect drags the CE upward. For high probabilities, in contrast, the two forces oppose each other. If the former dominates, affect drags the CE downward (panel A of Figure 2); if the latter dominates, affect drags the CE upward (panel B of Figure 2). Because the impact of willpower depletion and unrelated cognitive demands, such as cognitive load, depend on whether affect is dragging the CE upward or downward, our model yields somewhat different predictions for the certainty equivalent for simple nonmonetary gambles than for simple monetary gambles (we are not aware of any existing evidence on these predictions):

Risky Choice Prediction #3 (R-3): When generating a certainty equivalent for simple nonmonetary gambles, an increase in $h(W, \sigma)$ will increase CE when p is small, but when p is large, the effect is ambiguous.

Because the choice is between money versus a nonmonetary good, the implications of affective intensity are good-specific. In particular, an increase in the affective intensity for x will increase the affective system's motivation for x without changing its motivation for money, and thus increase the certainty equivalent. Analogously, an increase in the affective intensity for money will increase the affective system's motivation for money without changing its motivation for x, and thus decrease the certainty equivalent.

Risky Choice Prediction #4 (R-4): When generating a certainty equivalent for simple monetary gambles, any factor that increases the intensity of the affective motivation for the non-monetary good will increase CE, whereas any factor that increases the intensity of the affective motivation for money will decrease CE.

The excessive reaction to affectively charged but unlikely outcomes that is predicted by our model can be seen in numerous domains of behavior, from gold rushes to market manias to the mating behavior of young adults. Less anecdotally, Ditto et al. (2006) offered participants choices between gambles for the chance
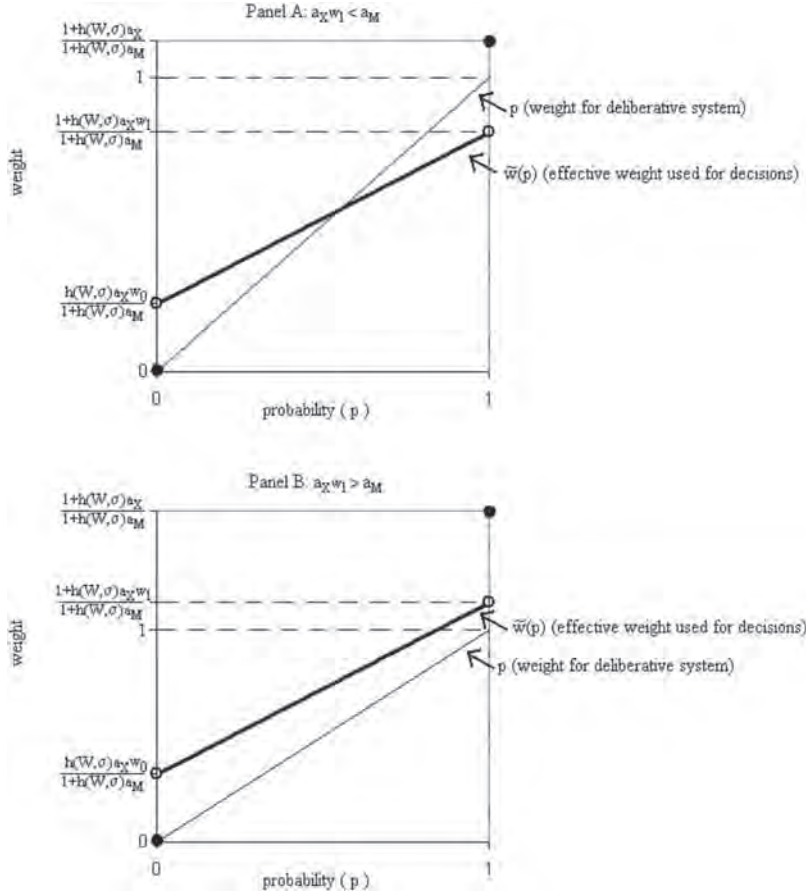
Figure 2. Effective probability weighting function $\tilde{w}(p)$ predicted by our model for certainty equivalents for nonmonetary gambles when the affective system has probability weighting function $w(p) = w_0 + (w_1-w_0)^*p$ with $w_0 > 0$ and $w_1 < 1$.

to win chocolate chip cookies and various fixed outside options. Half of the participants were only told about the cookies, whereas for the other half the cookies were freshly baked in the lab and placed in front of participants as they made their decision. Just as our model (R-4) predicts that increased affective intensity for cookies will increase the monetary certainty equivalent, it also predicts that increased affective intensity for cookies will make people more likely to accept the gamble over an outside option. This is exactly what is found by Ditto et al. (2006, though their results hold only for the high risk gambles).

Another study (Rottenstreich & Hsee, 2001) compared certainty equivalents for simple gambles that involve affect-rich outcomes (such as vacations and electric shocks) with certainty equivalents for simple gambles that involve affect-poor outcomes (such as money). In each case, they found that the certainty equivalent for the affect-rich outcome was larger than the certainty equivalent for the affect-poor outcome when the probability was very low (1%), but this result was reversed when the probability was very high (99%). From these results, they concluded that probability-weighting for affect-rich outcomes is more S-shaped than probability-weighting for affect-poor outcomes. In our model, affective intensity for the nonmonetary good does not directly translate into an effect on the probability-weighting function. Even so, these results are consistent with our model. In particular, according to our model, the increase

in probability from 1% to 99% will have a bigger effect on the affect-poor outcomes than on affect-rich outcomes as long as the deliberative system, which is the system influenced by the probability change, has a stronger reaction to the affect-poor outcome. For the case of gains, this means the deliberative system must prefer the affect-poor outcome (i.e., the utility of the affect-poor outcome is more positive), and in the case of losses, it means the deliberative system must prefer the affect-rich outcome (i.e., the utility of the affect-poor outcome is more negative). For the gambles studied by Rottenstreich and Hsee, both seem plausible.

Risk Preferences for Mixed (Gain-Loss) Gambles

Suppose a person must choose whether to accept a gamble (G,1/2; –L,1/2) with $G,L>0$. Unlike the previous decision, such gambles involve both gains and losses, and thus loss aversion becomes relevant. According to our model, the person will accept when $\left[\frac{1}{2}(G) + \frac{1}{2}(-L)\right] + h(W, \sigma)[\pi(a_M G) + \pi(-a_M \lambda L)] > 0$, where $\pi = w(1/2)$ here. This generates the following predictions:

> *Risky Choice Prediction #5 (R-5):* When facing 50–50 gain-loss gambles with $L < G < \lambda L$, an increase in $h(W, \sigma)$ will make it more likely that the person rejects the gamble.

> *Risky Choice Prediction #6 (R-6):* When facing 50–50 gain-loss gambles with $L < G < \lambda L$, any factor that increases the affective intensity for money will make it more likely that the person rejects the gamble.

If $G \leq L$ then both systems prefer to reject, and if $G \geq \lambda L$ then both systems prefer to accept, and so in either case willpower depletion, cognitive load, and affective intensity are irrelevant. The interesting case occurs when $L < G < \lambda L$—when the gamble has a small but positive expected value—in which case the deliberative system prefers to accept while the affective system prefers to reject. In such cases, willpower depletion, cognitive load, or affective intensity all increase the influence of loss aversion and make it more likely that the person will reject the gamble. For simplicity, we have restricted the previous example to the settings where both the gain and the loss outcomes are equally likely. However, these insights hold for more general gambles as well (in which the

effect of willpower depletion, cognitive load, or affective intensity will depend on gain and loss probabilities, in addition to gain and loss magnitudes).

Although we are not aware of any evidence of the impact of willpower depletion on risk-taking behavior, Benjamin et al. (2013) provide some indirect evidence on the effects of unrelated cognitive demands, such as cognitive load. In addition to asking the time preference questions described previously, they also asked their subjects to make a series of risky choices. Relative to control subjects, subjects who answer these questions while under cognitive load showed substantial reductions in risk taking behavior. Similar results have also been documented by Whitney et al. (2008) who found that the probability of choosing a risky gamble over a safe gamble reduced under cognitive load. To the extent that small-stakes risk aversion derives from loss aversion (Rabin, 2000; Rabin & Thaler, 2001), these results are consistent with the prediction that increasing cognitive load will lead to increased loss aversion (Prediction R-5).[7]

The Endowment Effect

Even though it is not an example of risky decision making, the endowment effect—the tendency to value an object more highly when one owns it—is commonly attributed to loss aversion (e.g., Tversky & Kahneman, 1991), and thus our model has implications for the endowment effect. Suppose, as in many experimental demonstrations of the endowment effect, that we elicit two reservation values: (a) The *selling price* P_S is the price such that, if the person is initially endowed with an object, she will be indifferent between keeping the object and receiving $$P_S$. (b) The *choice price* P_C is the price such that, if the person is initially not endowed with an object, she will be indifferent between gaining the object and receiving $$P_C$. The typical finding in experiments is that, even though the choices are

---

[7] Rabin and Thaler note that the preference for small-scale safe outcomes over small-scale risky outcomes observed in many laboratory experiments is inconsistent with the type of risky behavior observed for larger real-world stakes. Loss-aversion is a mechanism that can resolve this inconsistency.

the same—leaving the experiment with an object or with some money—the selling price is significantly larger than the choice price. To formalize this situation within our model, we assume, as done previously, that the deliberative system values money P as P and values the object x as $u(x)$. The deliberative system is not influenced by one's endowment. The affective system, in contrast, is sensitive to one's endowment. Specifically, when endowed, the affective system views the choice as [*gain P_S, lose $u(x)$*] versus [*no changes*]; and when not endowed, the affective system views the same choice as [*gain P_C*] versus [*gain $u(x)$*]. Hence, the selling price P_S and the choice price P_C are determined by $P_S + h(W, \sigma)[a_M P_S - \lambda a_x u(x)] = u(x) + h(W, \sigma)[0]$ and $P_C + h(W, \sigma)[a_M P_C] = u(x) + h(W, \sigma)[a_x u(x)]$, which generates $P_S/P_C = [1 + h(W, \sigma)\lambda a_x]/[1 + h(W, \sigma) a_x]$. Given $\lambda > 1$, not surprisingly our model yields an endowment effect ($P_S/P_C > 1$). More important, our model makes several predictions with regard to the endowment effect.

> *Endowment Effect Prediction #7 (R-7):* Any increase in $h(W, \sigma)$ will increase the magnitude of the endowment effect (increase P_S/P_C).

> *Endowment Effect Prediction #8 (R-8):* Any factor that increases the intensity of the affective motivation for the object will increase the magnitude of the endowment effect (increase P_S/P_C).

Because the endowment effect is driven by the affective system, willpower depletion or unrelated cognitive demands, such as cognitive load, will magnify the endowment effect. Moreover, because affective intensity for the object will magnify the impact of the affective system, it will also magnify the endowment effect. We know of no evidence for R-7. But there is support for the role of affective intensity for the object, Prediction R-8. Considerable research suggests that the endowment effect is more pronounced for outcomes such as changes in health status (see, for instance, Thaler, 1980). In one meta-analysis, Horowitz and McConnell (2002) found that, whereas the mean ratio of willingness to accept relative to willingness to pay for ordinary private goods was 2.9, this ratio was 10.1 for goods involving health and safety. While health outcomes differ from other outcomes in many ways, they are frequently associated with strong emotional reactions, and are thus more vulnerable to the effect of loss aversion and related features of the affective system.

Discussion

Taking account of the interplay between affect and deliberation helps to make sense of several important behavioral effects in the literature on decision making under risk, and it also leads to novel predictions about specific behaviors. Beyond the phenomena and predictions just outlined, the same framework could potentially shed light on and generate novel predictions concerning a variety of risk-related phenomena. For instance, the model can be used to understand the effects of temporal proximity on risk-taking. There is a great deal of evidence that temporal proximity is an important determinant of fear responses. As the prospect of an uncertain aversive event approaches in time, fear tends to increase even when cognitive assessments of the probability or likely severity of the event remain constant (Loewenstein, 1987; Roth et al., 1996). Similarly, after the moment of peak risk recedes into the past (e.g., after a near-accident), fear lingers for some period, but dissipates over time. Evidence that temporal proximity can influence risk behaviors comes from studies wherein people initially agree to do various embarrassing activities in exchange for payment, but then closer to the time when the activity has to be performed, change their minds (Van Boven et al., 2005). Moreover, consistent with changes in the affective state of fear being the cause, subjects who were shown a film clip designed to induce fear (from Kubrick's "The Shining") right before they made their initial decision were much less likely to choose to perform, and hence less likely to change their minds when the so-called "moment of truth" arrived.

Social Preferences

Humans experience a wide range of social emotions, from powerful empathic responses, such as sympathy and sadness, to more negative emotions, such as anger and envy. To give a flavor for how our two-system perspective can be applied to social preferences, in this section we apply our model to one specific social motive—altruism—and its associated affect—sympathy. The perspective we suggest is that the deliberative system has a stable concern for others driven by moral and ethical principles for

how one ought to behave. The affective system, in contrast, is driven toward anything between pure self-interest and extreme altruism depending on the degree of sympathy that is triggered.[8]

Suppose that each option x in the choice set X is a pair of payoffs $x = (x_S, x_O)$, where x_S is a payoff for oneself and x_O is a payoff for another person. The deliberative system puts some stable weight ϕ on the other person's payoff, and so its utility function is $U(x) = x_S + \phi x_O$. The affective system, in contrast, puts a variable weight on the other person's payoff that depends on the degree of sympathy that the person currently feels toward the other. Because the degree of sympathy is naturally interpreted as the intensity of affect, the affective system's motivational function is $M(x, a) = x_S + a x_O$.

Incorporating these functions into Equation 1, the person will choose the option x that maximizes

$$V(x) = [x_s + \phi x_o] + h(W, \sigma)[x_s + a x_o]. \quad (4)$$

One motivation for the assumptions in this section comes from studies of other-regarding behavior in animals, which, again, we take as evidence for what drives the affective system. Animals, including monkeys and rats, can be powerfully moved by the plight of others (for an overview, see Preston & de Waal, 2002). At the same time, other-regarding behavior is not always observed in animals. Masserman, Wechkin, and Terris (1964), for instance, found that prosocial behavior in primates (aiding another animal that was being subjected to electric shocks) was more likely in animals that had experienced shock themselves, was enhanced by familiarity with the shocked individual, and was nonexistent when it was a different species of animal. Perhaps stretching the terminology used in this article we can interpret these findings as a decrease in proximity leading to reduced concern for others.

Research by Joshua Greene and colleagues (Greene et al., 2001, 2004) provides neural evidence on our perspective. They compared how people react to "personal" moral judgments, which involve doing personal harm to another—for example, pushing a person in front of a trolley to stop it from hitting five other people—with how they react to "impersonal" moral judgments—for example, flicking a switch so that the trolley turns to another track and only

hits one person instead of five. They proposed that such judgments are made using a combination of cognitive processes that argue for utilitarian judgments and emotional processes that deter one from doing direct harm to others. Consistent with this view, they found that affective regions of the brain, such as areas of the temporal sulcus and posterior cingulate, are activated more for personal moral judgments than for impersonal moral judgments, whereas deliberative areas, such as the dorsolateral prefrontal cortex, are activated more in the opposite setting (and it has long been known that people are less likely to make the utilitarian judgment for the personal moral dilemma).[9] In the same vein, more recent research has shown that patients with brain damage to affective regions, such as the ventromedial prefrontal cortex, are more likely to make utilitarian, impersonal moral judgments, even in highly personal settings (Koenigs et al., 2007).

Predictions

Maximizing Equation 4 is equivalent to maximizing $\tilde{V}(x) = x_S + \tilde{\phi}(a)x_O$, where $\tilde{\phi}(a) = [\phi + h(W, \sigma)a]/[1 + h(W, \sigma)]$. Hence, the person's choice will reflect an effective concern for others that is a weighted average of the deliberative concern ϕ and the affective concern a. Moreover, the affective system can push behavior toward more or less concern for others relative to the deliberative optimum. In situations where there is very little sympathy triggered in the affective system, the affective system will push behavior closer to pure self-interest—as reflected by $a < \phi$ implying $\tilde{\phi}(a) < \phi$. In contrast, in situations where there are very high levels of sympathy triggered in the affective system, the affective system will push behavior toward more altruism—as reflected by $a > \phi$ implying $\tilde{\phi}(a) > \phi$.

---

[8] Note that in general, sympathy and altruism are not identical: Altruism may stem from sympathy, if behavior is controlled by the affective system, or it may stem from moral principles, if behavior is controlled by the deliberative system.

[9] Though note that Greene et al. (2004) suggest that the relationship between cognition and emotion may not be this simple; that is, certain limbic areas, such as the anterior cingulate cortex, may also be involved in detecting conflict between emotion and cognition, and in recruiting prefrontal cortex control of emotional regions to resolve this conflict.

To generate testable predictions, we apply the general predictions of our model.

> *Social Choice Prediction #1 (S-1):* An increase in $h(W, \sigma)$ will increase $\tilde{\phi}(a)$ when affective intensity is high $(a > \phi)$ and decrease $\tilde{\phi}(a)$ when affective intensity is low $(a < \phi)$.

> *Social Choice Prediction #2 (S-2):* Any factor that increases the intensity of the affective motivation will increase $\tilde{\phi}(a)$.

S-1 reflects that the effects of willpower depletion or unrelated cognitive demands, such as cognitive load, depend on the degree of sympathy experienced. Specifically, when a person experiences little or no sympathy our model predicts that willpower depletion or cognitive load should reduce the likelihood of an altruistic act. In contrast, when a person experiences high sympathy our model predicts that willpower depletion or cognitive load should increase the likelihood of an altruistic act.

Gailliot et al. (2007) provide support for the effects of willpower depletion when sympathy is low. Specifically, in a task involving hypothetical questions about charitable giving and helping behavior toward strangers—both arguably low-sympathy situations—they found that subjects with higher willpower depletion were indeed less altruistic. There is also evidence on the effects of affective intensity (S-2). Perhaps the most direct evidence comes from a study by Batson et al. (1995) on empathy-induced altruism. They manipulated subjects' empathy toward a target individual by having them read a short description of that individual's need while taking an objective perspective (low empathy) or while trying to imagine how that individual feels (high empathy). They then gave subjects the opportunity to help the target despite the fact that doing so would violate some moral principle of justice such as random allocations or allocation based on need. Consistent with S-2, they found that subjects in the high-empathy treatment were much more likely to help the target individual.

S-2 also helps to explain why people treat statistical deaths differently than identifiable ones, since foreknowledge of who will die (or which group deaths will come from) creates a more vivid—and evocative—image of the consequences (see Schelling, 1968; Bohnet & Frey, 1999; Slovic, 2007, Small & Loewenstein, 2003, for an experimental demonstration).

A recent study by Small et al. (2007) provides further support for our perspective on the role of identifiability. Small et al. provided subjects with the opportunity to donate to a charity, and manipulated whether subjects were shown an identifiable victim (a picture and a description of a little girl) or a statistical victim (factual information about the overall problem). They also manipulated the extent to which people were primed to think more deliberatively. They found that deliberative thought decreased donations to the identifiable victim, but did not affect donations to the statistical victim. Under the plausible assumption that the affective system plays a major role in donations to the identifiable victim and but not in donations to the statistical victim, these results are what our model (Prediction S-2) predicts.

While we have focused our analysis solely on the simple social motive of altruism, researchers have discussed other social motives as well. For instance, there is a large literature that focuses on people's concerns for relative payoffs (Bolton & Ockenfels, 2000; Fehr & Schmidt, 1999; Loewenstein et al., 1989; Messick & Sentis, 1985). In principle, our model could be applied to these concerns as well; however, because both concerns would seem to have both a deliberative and an affective component, it is not entirely obvious what to assume about the motives of the two systems. Similarly, another area our approach could be applied to is game theoretic decision making. Groups of decision makers are frequently able to avoid rational, but inefficient, outcomes, such as defection in prisoner's dilemma type games. While there are many reasons why decision makers cooperate in this manner, affective impulses, that are especially pronounced with close others, may play an important role in explaining this behavior.

Discussion

There is a great deal of evidence that people's decisions are influenced by both affective and deliberative processes. Whereas standard consequentialist models focus, for the most part, on deliberative processes, our main contribution in this article has been to develop a formal model to incorporate affective processes. In particular, we have modeled the impact of affective processes using a motivation function that is myopic, that displays loss aversion and is insensitive

to probabilities, and that is influenced by sympathy and empathy concerns. The impact of this motivation function on behavior is increasing in the affective intensity of the stimuli in consideration, increasing in unrelated cognitive demands, such as cognitive load, and decreasing in the willpower possessed by the decision maker. We have shown that our model can explain a range of psychological, behavioral and neuroscientific results regarding intertemporal, risky and interpersonal decision making, and generates some new predictions that have not yet been tested.

Ours is not the first dual-process model of decision making. Metcalfe and Mischel's (1999) hot/cool model and Fazio and Towles-Schwen's (1999) MODE model, for example, propose that behavior is the product of two systems: one emotional and the other cognitive. Metcalfe and Mischel use their model to understand the effect of willpower, and Fazio and Towles-Schwen apply their model to attitude formation and other aspects of social judgment. Our model differs from these two important and influential approaches in its focus on preferential choice and its ability to make quantitative predictions in this domain.

These properties make our model similar to formal dual-process theories of intertemporal choice in economics (Benhabib & Bisin, 2005; Bernheim & Rangel, 2004; Fudenberg & Levine, 2006; Shefrin & Thaler, 1988; Thaler & Shefrin, 1981). Indeed some of our assumptions—such as those of affective myopia—resemble those made by these approaches, and many of the insights presented by these approaches hold for our model as well. What is unique about our model is its ability to make predictions across a number of different domains, including both intertemporal and risky choice, as well as social choice. These predictions rely on a small set of fundamental principles—such as the sensitivity of emotion to proximity and vividness, the consequentialist nature of deliberation, and the role of willpower in resolving the conflict between these two systems—principles that are firmly grounded in psychology and neuroscience. Our model can thus be seen as generalizing these existing approaches, and subsequently extending the descriptive and conceptual scope of dual process theory for preferential choice.

Our model also resembles a prior dual-process theory in psychology. Particularly Mukherjee (2010) builds upon an early version of our model (Loewenstein & O'Donoghue, 2004) to study risk preferences in detail. As in Loewenstein & O'Donoghue (2004), Mukherjee assumes a deliberative and an affective system interact to determine behavior, where each has its own objective function, and behavior is determined by a weighted sum of the two objective functions. Also as in Loewenstein & O'Donoghue (2004), for the domain of risk preferences, Mukherjee assumes that the deliberative system focuses on expected value whereas the affective system is influenced by loss aversion and a complete insensitivity to probabilities (Mukherjee further assumes that the affective system is also influenced by diminishing sensitivity). Mukherjee then investigates the implications of this model for a number of well-known decision problems that have emerged in the prospect theory literature: violations of stochastic dominance, the nature of risk attitudes, ambiguity aversion, the common consequence effect, the common ratio effect, and the isolation effect. However, Mukherjee's analysis does not focus on the impact of willpower, cognitive load, or affective intensity, which is a primary focus of our article. Moreover, when we apply our model to risk preferences, we focus on implications for a completely different set of risk contexts—specifically, for four frequently studied experimental paradigms: eliciting monetary certainty equivalents for monetary gambles, eliciting monetary certainty equivalents for nonmonetary gambles, decisions whether to accept or reject mixed (gain-loss) gambles, and the endowment effect.

There are a number of directions in which to further expand upon our framework. Perhaps the most important is to more fully explore the dynamics of willpower. We have provided an outline of how willpower can change during the time course of the decision process, leading to switches midway through choice; however, there are even more nuanced willpower dynamics. For instance, some, albeit preliminary, studies have found support for the idea that, in addition to being depleted in the short-term by exertion, willpower, like a muscle, may become strengthened in the long-term through repeated use (Muraven et al., 1999). More importantly, people's behavior might also reflect their at-

tempts to manage their use of willpower. There is in fact experimental evidence, in a version of the Baumeister paradigm, that people do have some awareness of the dynamic properties of willpower and take these into account in a strategic fashion (Muraven, 1998).

A second direction in which to expand our framework is to study people's assessments of their own behaviors. Because such assessments are an inherently cognitive task, they will naturally tend to exaggerate the role played by deliberation. In effect, one could say that the deliberative self egocentrically views itself as in control and commensurately underestimates the influence of affect (see Wegner & Wheatley, 1999). This failure to appreciate the role of affect in behavior can have a negative impact on efforts at self-control.

An implication of failing to appreciate the role of affect is that people will exaggerate the importance of willpower as a determinant of self-control. People who are thin often believe they are thin because of willpower, and that those who are less fortunate exhibit a lack of willpower. However, it is far more likely that those who are thin are blessed (at least in times of plentiful food) with a high metabolism or a well-functioning ventromedial hypothalamus (which regulates hunger and satiation). Indeed, obese people who go to the extraordinary length of stapling their stomach to lose weight often report that they have a sudden experience of "willpower" despite the obvious fact that stapling one's stomach affects hunger rather than willpower (Gawande, 2001). It is easy and natural for those who lack drives and impulses for drugs, food, and sex to condemn, and hence to be excessively judgmental and punitive, toward those who are subject to them—to assume that these behaviors result from a generalized character deficit, a deficiency in willpower. Similarly, the rich, who are not confronted with the constant task of reigning in their desires, are likely to judge the short-sighted behaviors of the poor too harshly. There is in fact recent evidence that people who are in elevated affective states tend to have a much more acute appreciation of the power of drives and the limitations of self-control than those who are affectively neutral states (Nordgren et al., 2007).

A third direction in which to expand our framework is to take it to specific domains in order to develop more detailed model specifi-

cations and quantitative predictions. Mathematical models have two types of goals: (a) developing precise qualitative predictions, and (b) developing precise quantitative predictions. Our analysis in this article has focused exclusively on the former—for example, deriving precise qualitative predictions for the directional impact of cognitive load, willpower depletion, and affective intensity on various behavioral outcomes. As such, we have imposed relatively little general structure on the deliberative utility function U, the affective motivational function M, and the cost function h for mobilizing willpower. But if researchers take our framework to specific domains, it will be natural to impose—or better yet estimate—a more fully specified model, and to use that model to generate more quantitative predictions. Such a quantitative analysis would also help in comparing our model with the nested baseline rational model (which would involve only the deliberative utility function, U).

After decades of domination by a cognitive perspective, in recent decades affect has come to the fore as a topic of great interest among psychologists. In this article, we attempt to integrate many of the findings from research conducted by psychologists and decision researchers interested in affect by proposing a formal model of interactions between affect and deliberation that can both explain existing findings and also generates testable but as yet untested predictions. If further testing substantiates these predictions, and hence the model, this could constitute the first step toward a formal theoretical perspective that integrates two major sides of human judgment and behavior.

References

Ainslie, G. (1975). Specious reward: A behavioral theory of impulsiveness and impulse control. *Psychological Bulletin, 82,* 463–496. http://dx.doi.org/10.1037/h0076860

Ariely, D., & Loewenstein, G. (2006). The heat of the moment: The effect of sexual arousal on sexual decision making. *Journal of Behavioral Decision Making, 19,* 87–98. http://dx.doi.org/10.1002/bdm.501

Bankart, C. P., & Elliott, R. (1974). Heart rate and skin conductance in anticipation of shocks with varying probability of occurrence. *Psychophysiology, 11,* 160–174. http://dx.doi.org/10.1111/j.1469-8986.1974.tb00836.x

Barseghyan, L., Molinari, F., O'Donoghue, T., & Teitelbaum, J. C. (2013). The nature of risk preferences: Evidence from insurance choices. *The American Economic Review, 103,* 2499–2529. http://dx.doi.org/10.1257/aer.103.6.2499

Batson, C. D., Klein, T. R., Highberger, L., & Shaw, L. L. (1995). Immorality from empathy-induced altruism: When compassion and justice conflict. *Journal of Personality and Social Psychology, 68,* 1042–1054. http://dx.doi.org/10.1037/0022-3514 .68.6.1042

Baumeister, R. F., & Vohs, K. D. (2003). Willpower, choice, and self-control. In G. Loewenstein, D. Read, & R. F. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 201–216). New York, NY: Russell Sage Foundation.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275,* 1293–1295. http://dx.doi.org/10.1126/science.275 .5304.1293

Benhabib, J., & Bisin, A. (2005). Modelling internal commitment mechanisms and self-control: A neuroeconomics approach to consumption-saving decisions. *Games and Economic Behavior, 52,* 460–492. http://dx.doi.org/10.1016/j.geb.2004.10.004

Benjamin, D. J., Brown, S. A., & Shapiro, J. M. (2013). "Who is 'behavioral'? Cognitive ability and anomalous preferences." *Journal of the European Economics Association, 11,* 1231–1255.

Ben-Zion, U., Rapoport, A., & Yagil, J. (1989). Discount rates inferred from decisions: An experimental study. *Management Science, 35,* 270–284. http://dx.doi.org/10.1287/mnsc.35.3.270

Bernheim, B. D., & Rangel, A. (2004). Addiction and cue-triggered decision processes. *The American Economic Review, 94,* 1558–1590. http://dx.doi .org/10.1257/0002828043052222

Berns, G. S., Chappelow, J., Cekic, M., Zink, C. F., Pagnoni, G., & Martin-Skurski, M. E. (2006). Neurobiological substrates of dread. *Science, 312,* 754–758. http://dx.doi.org/10.1126/science .1123721

Berridge, K. C. (1996). Food reward: Brain substrates of wanting and liking. *Neuroscience and Biobehavioral Reviews, 20,* 1–25. http://dx.doi .org/10.1016/0149-7634(95)00033-B

Bjork, J. M., Momenan, R., & Hommer, D. W. (2009). Delay discounting correlates with proportional lateral frontal cortex volumes. *Biological Psychiatry, 65,* 710–713. http://dx.doi.org/10.1016/ j.biopsych.2008.11.023

Bleichrodt, H., Pinto, J. L., & Wakker, P. P. (2001). Making descriptive use of prospect theory to improve the prescriptive use of expected utility. *Management Science, 47,* 1498–1514. http://dx.doi .org/10.1287/mnsc.47.11.1498.10248

Blood, A. J., & Zatorre, R. J. (2001). Intensely pleasurable responses to music correlate with activity in brain regions implicated in reward and emotion. *Proceedings of the National Academy of Sciences of the United States of America, 98,* 11818–11823. http://dx.doi.org/10.1073/pnas.191355898

Bodenhausen, G. V. (1993). Emotions, arousal, and stereotypic judgments: A heuristic model of affect and stereotyping. In D. M. Mackie & D. L. Hamilton (Eds.), *Affect, cognition, and stereotyping: Interactive processes in group perception* (pp. 13–37). San Diego, CA: Academic Press. http://dx.doi .org/10.1016/B978-0-08-088579-7.50006-5

Bohnet, I., & Frey, B. (1999). The sound of silence in prisoner's dilemma and dictator games. *Journal of Economic Behavior & Organization, 38,* 43–57. http://dx.doi.org/10.1016/S0167-2681(98)00121-8

Bolton, G., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *The American Economic Review, 90,* 166–193. http://dx.doi .org/10.1257/aer.90.1.166

Chen, M. K., Lakshminarayanan, V., & Santos, L. R. (2006). How basic are behavioral biases? Evidence from capuchin monkey trading behavior. *Journal of Political Economy, 114,* 517–537. http://dx.doi .org/10.1086/503550

Cosmides, L., & Tooby, J. (2000). Evolutionary psychology and the emotions. In M. Lewis & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (2nd ed., pp. 91–115). New York, NY: Guilford Press.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711.

Deane, G. E. (1969). Cardiac activity during experimentally induced anxiety. *Psychophysiology, 6,* 17–30. http://dx.doi.org/10.1111/j.1469-8986 .1969.tb02879.x

De Martino, B., Camerer, C. F., & Adolphs, R. (2010). Amygdala damage eliminates monetary loss aversion. *Proceedings of the National Academy of Sciences of the United States of America, 107,* 3788–3792. http://dx.doi.org/10.1073/pnas .0910230107

Ditto, P. H., Pizarro, D. A., Epstein, E. B., Jacobson, J. A., & MacDonald, T. K. (2006). Visceral influences on risk-taking behavior. *Journal of Behavioral Decision Making, 19,* 99–113. http://dx.doi .org/10.1002/bdm.520

Elliott, R. (1975). Heart rate in anticipation of shocks which have different probabilities of occurrences. *Psychological Reports, 36,* 923–931. http://dx.doi .org/10.2466/pr0.1975.36.3.923

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist, 49,* 709–724. http://dx.doi.org/10.1037/0003-066X.49.8.709

Fazio, R. H., & Towles-Schwen, T. (1999). The MODE model of attitude-behavior processes. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology* (pp. 97–116). New York, NY: Guilford Press.

Fehr, E., & Schmidt, K. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics, 114,* 817–868. http://dx.doi.org/10.1162/003355399556151

Figner, B., Knoch, D., Johnson, E. J., Krosch, A. R., Lisanby, S. H., Fehr, E., & Weber, E. U. (2010). Lateral prefrontal cortex and self-control in intertemporal choice. *Nature Neuroscience, 13,* 538–539. http://dx.doi.org/10.1038/nn.2516

Frederick, S. (2003). Measuring intergenerational time preference: Are future lives valued less? *Journal of Risk and Uncertainty, 26,* 39–53. http://dx.doi.org/10.1023/A:1022298223127

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature, 40,* 351–401. http://dx.doi.org/10.1257/jel.40.2.351

Frijda, N. H. (1986). *The emotions*. New York, NY: University Press.

Fudenberg, D., & Levine, D. K. (2006). A dual self model of impulse control. *The American Economic Review, 96,* 1449–1476. http://dx.doi.org/10.1257/aer.96.5.1449

Gailliot, M. T., Baumeister, R. F., DeWall, C. N., Maner, J. K., Plant, E. A., Tice, D. M., . . . Schmeichel, B. J. (2007). Self-control relies on glucose as a limited energy source: Willpower is more than a metaphor. *Journal of Personality and Social Psychology, 92,* 325–336. http://dx.doi.org/10.1037/0022-3514.92.2.325

Gawande, A. (2001, July 9). The man who couldn't stop eating. *New York Times Magazine,* 66–75.

Giordano, L. A., Bickel, W. K., Loewenstein, G., Jacobs, E. A., Marsch, L., & Badger, G. J. (2002). Mild opioid deprivation increases the degree that opioid-dependent outpatients discount delayed heroin and money. *Psychopharmacology, 163,* 174–182. http://dx.doi.org/10.1007/s00213-002-1159-2

Goldstein, A. (2001). *Addiction: From biology to drug policy* (2nd ed.). New York, NY: Oxford University Press.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44,* 389–400. http://dx.doi.org/10.1016/j.neuron.2004.09.027

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293,* 2105–2108. http://dx.doi.org/10.1126/science.1062872

Halberstadt, J. B., & Niedenthal, P. M. (1997). Emotional state and the use of stimulus dimensions in judgment. *Journal of Personality and Social Psychology, 72,* 1017–1033. http://dx.doi.org/10.1037/0022-3514.72.5.1017

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., & Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *The Journal of Neuroscience, 28,* 5623–5630.

Hariri, A. R., Brown, S. M., Williamson, D. E., Flory, J. D., de Wit, H., & Manuck, S. B. (2006). Preference for immediate over delayed rewards is associated with magnitude of ventral striatal activity. *The Journal of Neuroscience, 26,* 13213–13217. http://dx.doi.org/10.1523/JNEUROSCI.3446-06.2006

Horowitz, J., & McConnell, K. (2002). A review of WTA-WTP studies. *Journal of Environmental Economics and Management, 44,* 426–447. http://dx.doi.org/10.1006/jeem.2001.1215

Ichihara-Takeda, S., & Funahashi, S. (2006). Reward-period activity in primate dorsolateral prefrontal and orbitofrontal neurons is affected by reward schedules. *Journal of Cognitive Neuroscience, 18,* 212–226. http://dx.doi.org/10.1162/jocn.2006.18.2.212

Johnson, M. W., Bickel, W. K., & Baker, F. (2007). Moderate drug use and delay discounting: A comparison of heavy, light, and never smokers. *Experimental and Clinical Psychopharmacology, 15,* 187–194. http://dx.doi.org/10.1037/1064-1297.15.2.187

Kable, J. W., & Glimcher, P. W. (2009). The neurobiology of decision: Consensus and controversy. *Neuron, 63,* 733–745.

Kahneman, D., & Frederick, S. (2002). Representativeness revisited: Attribute substitution in intuitive judgment. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics & biases: The psychology of intuitive judgment* (pp. 49–81). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098.004

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47,* 263–291. http://dx.doi.org/10.2307/1914185

Kirby, K. (1997). Bidding on the future: Evidence against normative discounting of delayed rewards. *Journal of Experimental Psychology: General, 126,* 54–70. http://dx.doi.org/10.1037/0096-3445.126.1.54

Knutson, B., Wimmer, G. E., Rick, S., Hollon, N. G., Prelec, D., & Loewenstein, G. (2008). Neural antecedents of the endowment effect. *Neuron, 58,* 814–822. http://dx.doi.org/10.1016/j.neuron.2008.05.018

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446,* 908–911. http://dx.doi.org/10.1038/nature05631

Laibson, D. (1997). Golden eggs and hyperbolic discounting. *The Quarterly Journal of Economics, 112,* 443–478. http://dx.doi.org/10.1162/003355397555253

LeDoux, J. E. (1996). *The emotional brain: The mysterious underpinnings of emotional life.* New York, NY: Simon & Schuster.

Lerner, J. S., & Keltner, D. (2000). Beyond valence: Toward a model of emotion-specific influences on judgment and choice. *Cognition and Emotion, 14,* 473–493. http://dx.doi.org/10.1080/026999300402763

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81,* 146–159. http://dx.doi.org/10.1037/0022-3514.81.1.146

Lerner, J. S., Small, D. A., & Loewenstein, G. (2004). Heart strings and purse strings: Carryover effects of emotions on economic decisions. *Psychological Science, 15,* 337–341. http://dx.doi.org/10.1111/j.0956-7976.2004.00679.x

Lewin, K. (1951). *Field theory in social science.* New York, NY: Harper & Borthers.

Lhermitte, F. (1986). Human autonomy and the frontal lobes. Part II: Patient behavior in complex and social situations: The "environmental dependency syndrome." *Annals of Neurology, 19,* 335–343. http://dx.doi.org/10.1002/ana.410190405

Lieberman, M. D. (2003). Reflective and reflexive judgment processes: A social cognitive neuroscience approach. In J. P. Forgas, K. R. Williams, & W. von Hippel (Eds.), *Social judgments: Implicit and explicit processes* (pp. 44–67). New York: Cambridge University Press.

Loewenstein, G. (1987). Anticipation and the valuation of delayed consumption. *The Economic Journal, 97,* 666–684. http://dx.doi.org/10.2307/2232929

Loewenstein, G. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes, 65,* 272–292. http://dx.doi.org/10.1006/obhd.1996.0028

Loewenstein, G. (2007). Defining affect. *Social Sciences Information Sur les Sciences Sociales, 46,* 405–410. http://dx.doi.org/10.1177/05390184070460030106

Loewenstein, G., & Lerner, J. (2003). The role of emotion in decision making. In R. J. Davidson, H. H. Goldsmith, & K. R. Scherer (Eds.), *Hand-*book of affective sciences (pp. 619–642). New York, NY: Oxford University Press.

Loewenstein, G., & O'Donoghue, T. (2004). *Animal spirits: Affective and deliberative processes in economic behavior. Mimeo.* Ithaca, NY: Cornell University.

Loewenstein, G., Thompson, L., & Bazerman, M. (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology, 57,* 426–441. http://dx.doi.org/10.1037/0022-3514.57.3.426

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127,* 267–286. http://dx.doi.org/10.1037/0033-2909.127.2.267

MacLean, P. D. (1990). *The triune brain in evolution: Role in paleocerebral function.* New York, NY: Plenum Press.

Manuck, S. B., Flory, J. D., Muldoon, M. F., & Ferrell, R. E. (2003). A neurobiology of intertemporal choice. In G. Loewenstein, D. Read, & R. F. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 139–172). New York, NY: Russell Sage Foundation.

Masserman, J. H., Wechkin, S., & Terris, W. (1964). "Altruistic" behavior in rhesus monkeys. *The American Journal of Psychiatry, 121,* 584–585. http://dx.doi.org/10.1176/ajp.121.6.584

McClure, S. M., Ericson, K. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2006). *Time discounting for primary rewards.* Working Paper, Center for the Study of Brain, Mind, & Behavior and Department of Psychology, Princeton University.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards. *Science, 306,* 503–507. http://dx.doi.org/10.1126/science.1100907

Mellers, B. A., Schwartz, A., Ho, K., & Ritov, I. (1997). Decision affect theory: Emotional reactions to the outcomes of risky options. *Psychological Science, 8,* 423–429. http://dx.doi.org/10.1111/j.1467-9280.1997.tb00455.x

Messick, D. M., & Sentis, K. P. (1985). Estimating social and nonsocial utility functions from ordinal data. *European Journal of Social Psychology, 15,* 389–399.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: Dynamics of willpower. *Psychological Review, 106,* 3–19. http://dx.doi.org/10.1037/0033-295X.106.1.3

Milkman, K. L., Rogers, T., & Bazerman, M. H. (2008). Harnessing our inner angels and demons: What we have learned about want/should conflicts and how that knowledge can help us reduce short-sighted decision making. *Perspectives on Psycho-*

logical Science, 3, 324–338. http://dx.doi.org/10.1111/j.1745-6924.2008.00083.x

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. http://dx.doi.org/10.1146/annurev.neuro.24.1.167

Mischel, W., Ayduk, O., & Mendoza-Denton, R. (2003). Sustaining delay of gratification over time: A hot-cool systems perspective. In G. Loewenstein, D. Read, & R. F. Baumeister (Eds.), *Time and decision: Economic and psychological perspectives on intertemporal choice* (pp. 175–200). New York, NY: Russell Sage Foundation.

Mischel, W., Ebbesen, E. B., & Zeiss, A. R. (1972). Cognitive and attentional mechanisms in delay of gratification. *Journal of Personality and Social Psychology, 21,* 204–218. http://dx.doi.org/10.1037/h0032198

Mischel, W., Shoda, Y., & Rodriguez, M. I. (1989). Delay of gratification in children. *Science, 244,* 933–938. http://dx.doi.org/10.1126/science.2658056

Monat, A., Averill, J. R., & Lazarus, R. S. (1972). Anticipatory stress and coping reactions under various conditions of uncertainty. *Journal of Personality and Social Psychology, 24,* 237–253. http://dx.doi.org/10.1037/h0033297

Mukherjee, K. (2010). A dual system model of preferences under risk. *Psychological Review, 117,* 243.

Muraven, M. (1998). *Mechanisms of self-control failure: Motivation and limited resource.* PhD dissertation, Case Western Reserve University.

Muraven, M., Baumeister, R. F., & Tice, D. M. (1999). Longitudinal improvement of self-regulation through practice: Building self-control strength through repeated exercise. *The Journal of Social Psychology, 139,* 446–457. http://dx.doi.org/10.1080/00224549909598404

Nordgren, L. F., van der Pligt, J., & van Harreveld, F. (2007). Evaluating Eve: Visceral states influence the evaluation of impulsive behavior. *Journal of Personality and Social Psychology, 93,* 75–84. http://dx.doi.org/10.1037/0022-3514.93.1.75

Panksepp, J. (1998). *Affective neuroscience.* New York, NY: Oxford University Press.

Peters, E., & Slovic, P. (2000). The springs of action: Affective and analytical information processing in choice. *Personality and Social Psychology Bulletin, 26,* 1465–1475. http://dx.doi.org/10.1177/01461672002612002

Pham, M. T. (1998). Representativeness, relevance, and the use of feelings in decision making. *Journal of Consumer Research, 25,* 144–159. http://dx.doi.org/10.1086/209532

Preston, S. D., & de Waal, F. B. M. (2002). Empathy: Its ultimate and proximate bases. *Behavioral and Brain Sciences, 25,* 1–20.

Rabin, M. (2000). Risk aversion and expected-utility theory: A calibration theorem. *Econometrica, 68,* 1281–1292. http://dx.doi.org/10.1111/1468-0262.00158

Rabin, M., & Thaler, R. (2001). Anomalies: Risk aversion. *The Journal of Economic Perspectives, 15,* 219–232. http://dx.doi.org/10.1257/jep.15.1.219

Raghunathan, R., & Pham, M. T. (1999). All negative moods are not equal: Motivational influences of anxiety and sadness on decision making. *Organizational Behavior and Human Decision Processes, 79,* 56–77. http://dx.doi.org/10.1006/obhd.1999.2838

Rick, S., & Loewenstein, G. (2008). The role of emotion in economic behavior. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *The handbook of emotion* (3rd ed., pp. 138–156). New York, NY: Guilford Press.

Rolls, E. T. (1999). *The brain and emotion.* New York: Oxford University Press.

Rosati, A. G., Stevens, J. R., Hare, B., & Hauser, M. D. (2007). The evolutionary origins of human patience: Temporal preferences in chimpanzees, bonobos, and human adults. *Current Biology, 17,* 1663–1668. http://dx.doi.org/10.1016/j.cub.2007.08.033

Roth, W. T., Breivik, G., Jørgensen, P. E., & Hofmann, S. (1996). Activation in novice and expert parachutists while jumping. *Psychophysiology, 33,* 63–72. http://dx.doi.org/10.1111/j.1469-8986.1996.tb02109.x

Rottenstreich, Y., & Hsee, C. K. (2001). Money, kisses, and electric shocks: On the affective psychology of risk. *Psychological Science, 12,* 185–190. http://dx.doi.org/10.1111/1467-9280.00334

Schelling, T. C. (1968). The life you save may be your own. In S. B. Chase (Ed.), *Problems in public expenditure analysis* (pp. 127–162). Washington, DC: The Brookings Institute.

Scholten, M., & Read, D. (2010). The psychology of intertemporal tradeoffs. *Psychological Review, 117,* 925–944. http://dx.doi.org/10.1037/a0019619

Shefrin, H. M., & Thaler, R. H. (1988). The behavioral life-cycle hypothesis. *Economic Inquiry, 26,* 609–643. http://dx.doi.org/10.1111/j.1465-7295.1988.tb01520.x

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II Perceptual learning, automatic attending and a general theory. *Psychological Review, 84,* 127–190. http://dx.doi.org/10.1037/0033-295X.84.2.127

Shiv, B., & Fedorikhin, A. (1999). Heart and mind in conflict: The interplay of affect and cognition in consumer decision making. *Journal of Consumer Research, 26,* 278–292. http://dx.doi.org/10.1086/209563

Shiv, B., Loewenstein, G., Bechara, A., Damasio, H., & Damasio, A. (2003). *Investment behavior and the dark side of emotion. Mimeo.* Iowa City, IA: University of Iowa.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin, 119,* 3–22. http://dx.doi.org/10.1037/0033-2909.119.1.3

Slovic, P. (2007). "If I look at the mass I will never act": Psychic numbing and genocide. *Judgment and Decision Making, 2,* 79–95.

Slovic, P., Finucane, M., Peters, E., & MacGregor, D. G. (2002). The affect heuristic. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 397–420). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511808098.025

Small, D. A., & Loewenstein, G. (2003). Helping the victim or helping a victim: Altruism and identifiability. *Journal of Risk and Uncertainty, 26,* 5–16. http://dx.doi.org/10.1023/A:1022299422219

Small, D. A., Loewenstein, G., & Slovic, P. (2007). Sympathy and callousness: The impact of deliberative thought on donations to identifiable and statistical victims. *Organizational Behavior and Human Decision Processes, 102,* 143–153. http://dx.doi.org/10.1016/j.obhdp.2006.01.005

Smith, E. R., & DeCoster, J. (2000). Dual process models in social and cognitive psychology: Conceptual integration and links to underlying memory systems. *Personality and Social Psychology Review, 4,* 108–131.

Snortum, J. R., & Wilding, F. W. (1971). Temporal estimation and heart rate as a function of repression-sensitization score and probability of shock. *Journal of Consulting and Clinical Psychology, 37,* 417–422. http://dx.doi.org/10.1037/h0031874

Sokol-Hessner, P., Camerer, C. F., & Phelps, E. A. (2013). Emotion regulation reduces loss aversion and decreases amygdala responses to losses. *Social Cognitive and Affective Neuroscience, 8,* 341–350. http://dx.doi.org/10.1093/scan/nss002

Starmer, C. (2000). Developments in non-expected utility theory: The hunt for a descriptive theory of choice under risk. *Journal of Economic Literature, 38,* 332–382. http://dx.doi.org/10.1257/jel.38.2.332

Steinberg, L., Graham, S., O'Brien, L., Woolard, J., Cauffman, E., & Banich, M. (2009). Age differences in future orientation and delay discounting. *Child Development, 80,* 28–44. http://dx.doi.org/10.1111/j.1467-8624.2008.01244.x

Stevens, J. R., Hallinan, E. V., & Hauser, M. D. (2005). The ecology and evolution of patience in two New World monkeys. *Biology Letters, 1,* 223–226. http://dx.doi.org/10.1098/rsbl.2004.0285

Strack, F., & Deutsch, R. (2004). Reflective and impulsive determinants of social behavior. *Person-ality and Social Psychology Review, 8,* 220–247. http://dx.doi.org/10.1207/s15327957pspr0803_1

Taylor, S. E., & Thompson, S. C. (1982). Stalking the elusive "vividness" effect. *Psychological Review, 89,* 155.

Thaler, R. H. (1980). Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization, 1,* 39–60. http://dx.doi.org/10.1016/0167-2681(80)90051-7

Thaler, R. H. (1981). Some empirical evidence on dynamic inconsistency. *Economics Letters, 8,* 201–207. http://dx.doi.org/10.1016/0165-1765(81)90067-7

Thaler, R. H., & Shefrin, H. M. (1981). An economic theory of self-control. *Journal of Political Economy, 89,* 392–406. http://dx.doi.org/10.1086/260971

Tobin, H., Logue, A. W., Chelonis, J. J., Ackerman, K. T., & May, J. G. (1996). Self-control in the monkey Macaca mascicularis. *Animal Learning & Behavior, 24,* 168–174. http://dx.doi.org/10.3758/BF03198964

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science, 315,* 515–518. http://dx.doi.org/10.1126/science.1134239

Trope, Y., & Liberman, N. (2003). Temporal construal. *Psychological Review, 110,* 403–421. http://dx.doi.org/10.1037/0033-295X.110.3.403

Tversky, A., & Kahneman, D. (1991). Loss aversion in riskless choice: A reference dependent model. *The Quarterly Journal of Economics, 106,* 1039–1061. http://dx.doi.org/10.2307/2937956

Van Boven, L., Loewenstein, G., & Dunning, D. (2005). The illusion of courage in social predictions: Underestimating the impact of fear of embarrassment on other people. *Organizational Behavior and Human Decision Processes, 96,* 130–141. http://dx.doi.org/10.1016/j.obhdp.2004.12.001

Vohs, K., Baumeister, R. F., & Schmeichel, B. J. (2013). Motivation, personal beliefs, and limited resources all contribute to self-control. *Journal of Experimental Social Psychology, 49,* 184–188. http://dx.doi.org/10.1016/j.jesp.2012.08.007

Vohs, K. D., Baumeister, R. F., Schmeichel, B. J., Twenge, J. M., Nelson, N. M., & Tice, D. M. (2008). Making choices impairs subsequent self-control: A limited-resource account of decision making, self-regulation, and active initiative. *Journal of Personality and Social Psychology, 94,* 883–898. http://dx.doi.org/10.1037/0022-3514.94.5.883

Vohs, K. D., & Faber, R. J. (2007). Spent resources: Self-regulatory resource availability affects impulse buying. *Journal of Consumer Research, 33,* 537–547. http://dx.doi.org/10.1086/510228

Vohs, K. D., & Heatherton, T. F. (2000). Self-regulatory failure: A resource-depletion approach. *Psychological Science, 11,* 249–254. http://dx.doi.org/10.1111/1467-9280.00250

Weber, B., Aholt, A., Neuhaus, C., Trautner, P., Elger, C. E., & Teichert, T. (2007). Neural evidence for reference-dependence in real-market transactions. *NeuroImage, 35,* 441–447. http://dx.doi.org/10.1016/j.neuroimage.2006.11.034

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review, 111,* 430–445. http://dx.doi.org/10.1037/0033-295X.111.2.430

Wegner, D. M., & Wheatley, T. (1999). Apparent mental causation: Sources of the experience of will. *American Psychologist, 54,* 480–492. http://dx.doi.org/10.1037/0003-066X.54.7.480

Whitney, P., Rinehart, C. A., & Hinson, J. M. (2008). Framing effects under cognitive load: The role of working memory in risky decisions. *Psychonomic Bulletin & Review, 15,* 1179–1184. http://dx.doi.org/10.3758/PBR.15.6.1179

Wilson, M., & Daly, M. (2003). Do pretty women inspire men to discount the future? *Biology Letters Proceedings: Biological Sciences, 4,* 177–179.

Wilson, T. D., Lindsey, S., & Schooler, T. Y. (2000). A model of dual attitudes. *Psychological Review, 107,* 101–126. http://dx.doi.org/10.1037/0033-295X.107.1.101

Yechiam, E., Busemeyer, J. R., Stout, J. C., & Bechara, A. (2005). Using cognitive models to map relations between neuropsychological disorders and human decision-making deficits. *Psychological Science, 16,* 973–978.

Zald, D. H., & Pardo, J. V. (1997). Emotion, olfaction, and the human amygdala: Amygdala activation during aversive olfactory stimulation. *Proceedings of the National Academy of Sciences of the United States of America, 94,* 4119–4124. http://dx.doi.org/10.1073/pnas.94.8.4119

Recognition Without Awareness: Encoding and Retrieval Factors

Fergus I. M. Craik, Nathan S. Rose, and Nigel Gopie
Rotman Research Institute at Baycrest, Toronto, Ontario, Canada

The article reports 4 experiments that explore the notion of recognition without awareness using words as the material. Previous work by Voss and associates has shown that complex visual patterns were correctly selected as targets in a 2-alternative forced-choice (2-AFC) recognition test although participants reported that they were guessing. The present experiments sought to extend this earlier work by having participants study words in different ways and then attempt to recognize the words later in a series of 4-alternative forced-choice (4-AFC) tests, some of which contained *no* target word. The data of interest are cases in which a target was present and participants stated that they were guessing, yet chose the correct item. This value was greater than $p = .25$ in all conditions of the 4 experiments, demonstrating the phenomenon of recognition without awareness. Whereas Voss and colleagues attributed their findings with kaleidoscope patterns to enhanced processing fluency of perceptual attributes, the main factor associated with different levels of recognition without awareness in the present studies was a variable criterion for the subjective state accompanying selection of the "guess" option, depending on the overall difficulty of the recognition test. We conclude by discussing some implications of the results for the distinction between implicit and explicit memory.

Keywords: recognition, awareness, criterion shifts, implicit memory, explicit memory

In real-life decision situations we are often faced with alternatives that seem so equivalent that choice is extremely difficult. Under such circumstances our final selection may feel like an arbitrary choice, although in fact there may be implicit influences acting outside conscious control that bias us toward selecting one alternative over another. The observation that people can make correct choices while believing that they are selecting randomly has a long history in experimental psychology. Studies dating from the 19th century have consistently found that participants can make subtle perceptual discrimination judgments with above-chance accuracy despite claims that they are simply guessing (Adams, 1957; Voss & Paller, 2010). Voss and colleagues have recently provided evidence for a similar effect in recognition memory (Voss, Baym & Paller, 2008). Participants studied a series of kaleidoscope images and then attempted to recognize the studied items among a set of perceptually similar pairs. The study phase was performed under either full attention (FA) or divided attention (DA) conditions, and the recognition test was either a yes-no test (10 studied targets mixed with 10 similar foils) or a 2-AFC test (10 simultaneously presented target-foil pairs). In the yes-no test, recognition accuracy was good following encoding under FA conditions but very poor following DA at encoding, as one might expect in an explicit memory situation. Surprisingly, however, participants' performance on the forced-choice test was better following DA than FA at encoding. Further experiments revealed that when participants were asked to rate their forced-choice responses as being on the basis of some memory for the studied item or as random guesses, recognition accuracy was higher for responses judged to be guesses than for those thought to be based on memory.

These experiments thus provide evidence for substantial levels of recognition memory when participants believe they are simply guessing—that is, for recognition without awareness. This result was obtained only under very specific conditions, however—when encoding was performed under DA conditions, when the test was 2-AFC, when responding was under a tight time deadline (c. 2 sec from stimulus onset), and when the choice was between two perceptually similar visual patterns. There was essentially no evidence for the effect with a yes-no testing procedure or even with a forced-choice procedure when participants were given unlimited time to respond or when target stimuli were paired with a perceptually dissimilar foil (Voss et al., 2008, Experiments 3 & 4, respectively). A subsequent study revealed a further limitation; the effect was not obtained in the forced-choice procedure when participants were encouraged to respond accurately and guess only when absolutely necessary, although the original result reappeared when participants were encouraged to guess (Voss & Paller, 2010).

Voss and colleagues refer to their finding as "implicit recognition" and suggest that the underlying processes are different both

from those mediating explicit recollection *and* from those mediating feelings of familiarity. They comment that "Familiarity-based recognition is taken as an instance of explicit memory because familiarity responses entail the awareness of memory retrieval" (Voss et al., 2008, p. 458). In support of the claim that implicit recognition has a different mechanism they cite a further study (Voss & Paller, 2009) in which participants performed the forced-choice test for kaleidoscope patterns, encoded under either FA or DA conditions. Participants in this study assessed each recognition choice as being associated with some explicit recollection of the encoding phase ("remember" = R), with a more general feeling of familiarity (they simply "knew" it had been studied = K), or as a pure guess. Event-related potential (ERP) recordings were also made during the recognition test. The results confirmed earlier findings of higher levels of accuracy following DA at encoding, and also of greater than chance accuracy levels with "guess" responses, especially in the DA condition. Additionally, the pattern of behavioral results in guess decisions was distinct from the pattern observed with both R and K decisions, suggesting that the mechanism associated with implicit recognition is different from that associated with recognition with awareness. The ERP results supported this claim. Recognition responses accompanied by feelings of recollection or familiarity were associated with positive shifts in the late positive complex (600–900 ms) and in the P200 potential. In contrast, correct guess responses were associated with frontal-occipital negative potentials occurring 200–400 ms after stimulus onset. The authors speculate that the distinct mechanism underlying the phenomenon of recognition without awareness may reflect a stimulus-specific enhancement of perceptual fluency (e.g., Jacoby & Whitehouse, 1989), with this subtle change in processing yielding enough information to support a correct recognition choice, although not enough to give rise to any conscious feeling of remembering.

A major question arising from this work is whether the phenomenon of recognition without awareness can be demonstrated with material other than complex perceptual patterns and, if so, whether it is associated with similar neural mechanisms. Is implicit recognition found with verbal materials, for example? In one early experiment, Peynircioğlu (1990) had participants study a list of words, and then gave them a word-fragment completion test in which some fragments were from the list and others were new. Participants attempted to complete the fragments and also rated each fragment with regard to whether it was based on a list member or based on a new word. Considering only fragments that were *not* completed, a higher mean rating was given to list than lure words. Thus, apparently participants had some sense of familiarity for the fragments even in the absence of identification. Subsequent work by Cleary and Greene (2004, 2005) showed that when studied and unstudied words were presented too quickly to identify in a perceptual identification test, participants could still discriminate studied from unstudied items. The authors attributed the effect to a greater sense of familiarity associated with the briefly flashed studied words. The finding that recognition without identification is associated with a specific ERP signal (Voss & Paller, 2009) was confirmed and extended to verbal material in a study using the method of Peynircioğlu (1990) and reported by Ryals, Yadon, Nomi, and Cleary (2011). The two major findings were, first, that for unidentified word fragments the proportion attributed to the original list was greater for studied than unstudied

unidentified items; that is, recognition without identification (RWI) was again obtained. Second, the ERP correlate of the RWI effect was an N300 component of the evoked response, in agreement with Voss and Paller (2009) but using verbal materials and a yes-no recognition procedure. Ryals and colleagues concluded that their results confirmed the existence of unconscious recognition memory and that the RWI effect is indexed by the N300 ERP signature.

A study by Starns, Hicks, Brown, and Martin (2008) also found evidence for recognition without identification using verbal material. Their basic paradigm was to have participants study a list of words in which half of the words were printed in large font and half in small font (Experiment 1), or were rated for either pleasantness or imageability (Experiments 2 & 3). Participants were then given a recognition list composed of 50% studied words and 50% lures; additionally, half of the participants were informed that only 25% were targets and the other half informed that 75% were targets. Following this test, participants were re-presented with the original list and asked to decide the "source"—that is, whether each word had been in small or large font (or rated for pleasantness or imageability). The major finding was that participants' source judgments were above chance for words they had failed to recognize in the first test. Importantly, however, this effect was found only in the condition in which participants were informed that only 25% of the test words were targets. The authors concluded that the phenomenon of accurate source memory for unrecognized items is a reality, but that it occurs only under conditions in which a conservative response bias has been induced.

In summary, there is good evidence for the phenomenon of recognition without awareness, although the evidence associated with verbal materials is somewhat indirect in the sense that correct decisions about list membership were made on the basis of word fragments or the words themselves presented very briefly (Cleary and colleagues). Similarly, in the experiments by Starns et al. (2008) the evidence for recognition without identification comes from above-chance attribution of source rather than of the words themselves. One interesting question then is whether the phenomenon would extend to conditions in which participants correctly select words presented in full view despite claiming that they are simply guessing. This is the question addressed in the present experiments.

We became interested in these findings when considering the results of an earlier set of experiments reported by Gopie, Craik and Hasher (2011). In that study, younger and older adult participants first named the print color (red, green, blue, yellow) of a series of words as rapidly as possible; they were informed that the words themselves were irrelevant. This encoding phase was followed by a word fragment completion test, containing fragments of words from the "encoded" list as well as new word fragments. The higher completion rate for repeated words than for new words (priming effect) was greater for older adults (0.25) than for younger adults (0.10), in line with the notion that older adults fail to inhibit "irrelevant" information, which they can subsequently use if that information becomes useful (Hasher, Zacks & May, 1999). Surprisingly, however, this pattern reversed in a second experiment using the same color-naming initial phase, but with explicit instructions to "use words from the initial list where possible" in the fragment-completion test. Now younger adults had a priming score of 0.24 and older adults' score dropped to 0.08.

The same Age × Implicit/Explicit interaction was replicated in a further experiment.

What is the nature of the encoded verbal information in the incidental color-naming situation? It is well established that implicit verbal tests such as fragment completion are particularly sensitive to perceptual information (e.g., Craik, Moscovitch & McDowd, 1994; Schacter, Dobbins, & Schnyer, 2004), and the successful priming shown by older adults suggests that they may have encoded words in the color-naming phase in a perceptual manner. When younger adults performed the color-naming task under DA conditions, their subsequent fragment-completion performance resembled that of older adults (implicit completion = 0.22, explicit completion = −.03; Gopie et al., 2011, Experiment 3), again suggesting that the DA condition induced a somewhat superficial encoding of the words. This DA condition obviously resembles the DA conditions used by Voss and colleagues, and the finding of successful fragment completion subsequent to this type of encoding fits well with Voss and colleagues' characterization of their recognition without awareness as reflecting enhanced perceptual fluency. These initial findings prompted the question of whether recognition without awareness would be observed if the color-naming initial phase was followed by an *explicit* recognition test. Would participants select a previously viewed word at greater than chance levels while claiming that they were simply guessing? The following experiments investigated this possibility.

A further purpose of the present series of studies was to obtain more information about the types of representation associated with implicit recognition, and the factors that affect the size of the effect. The similarities and differences between effects obtained with words and with kaleidoscope patterns should suggest commonalities and limitations among the various representations underlying implicit recognition effects. The results may also point to differences in the representations associated with implicit and explicit memory for the studied items. Do such differences reflect the involvement of different memory systems, for example (e.g., Tulving & Schacter, 1990), or simply differences in the types or amounts of information the representations contain about the original episode (e.g., Chechile, Sloboda & Chamberland, 2012)?

With regard to factors that might influence the size of the effect, we were influenced by two findings from previous studies and one conjecture of our own. First, we presented words in an initial encoding phase under either full or divided attention conditions (FA or DA). The reasons for this followed the stronger effects observed by Voss and colleagues under DA conditions, also the possibility from the studies by Gopie et al. (2011) that DA induces a more superficial perceptual encoding. We speculated that recognition without awareness for words might also be stronger following such conditions. Second, the results of Starns et al. (2008) provided strong evidence that the effect would be greater under conditions of conservative responding in the test phase, so our results were examined with this point in mind. Finally, in line with the notions of encoding specificity (Tulving & Thomson, 1973) we hypothesized that the effect would be stronger to the extent that the encoding and test conditions were different, so this factor was also incorporated in our design. The rationale for this last point is described in the next paragraph.

To illustrate the phenomenon of recognition failure of recallable words, Tulving and Thomson (1973) first presented target words as response items in a paired-associate list. Next, in an apparently unrelated phase of the experiment, participants were given cue words and asked to generate four free associations to each cue word. The cues were chosen so that the generated associations often matched response words in the previous paired-associate list. In the third phase, participants were asked to read through the words they had generated and circle any they recognized as the previously learned words. Finally, they were given the original paired-associate stimulus words as cues to recall the appropriate responses. The main result was that participants often failed to recognize generated target words in Phase 3, although the same words were recalled in response to the original paired-associate cues in the final phase. The authors' interpretation of this striking result was that target words encoded specifically as responses in a paired-associate list were not perceived subjectively as the same words when encountered again as their own generated associations. Essentially, the context change between encoding and test acted to reduce recognition. Applying this thinking to the phenomenon of recognition without awareness, our conjecture was that a similar change of context between initial encoding and the recognition test might result in a failure of explicit recognition, but allow the participant to still select the correct target word in a forced-choice test by virtue of implicit recognition. This notion predicts that greater amounts of context change between encoding and test would be associated with increased recognition failure but also an increased likelihood of 'recognition without awareness.'

In order to provide a sensitive test of recognition memory, and to allow for the possibility of recognition without awareness, we used the testing procedure devised by Tulving and Thomson. In the present studies, participants were shown a series of 4-word sets and instructed to choose the one word they may have encountered in the first (encoding) phase. They were also told (correctly) that some of the 4-word sets contained *no* target, but that they must still select one word as the most likely to have been in the first phase. To make sense of this procedure, participants were also instructed to give a confidence rating for each choice, in which 2 = "fairly certain it was on the list," 1 = "possibly on the list," and 0 = "pure guess—I was forced to choose one." In a 4-AFC situation, chance responding will yield $p = .25$, so the interest in the present experiments is in cases in which target items *are* present, participants give a rating of 0, yet choose correctly at a level higher than 0.25. In line with the previous literature, we will refer to such outcomes as 'recognition without awareness.' Experiments 1 and 2 utilized the color-naming procedure reported by Gopie et al. (2011) as the encoding phase, and in the final two experiments we used the paired-associate learning task reported by Tulving and Thomson (1973).

Experiment 1

Method

Participants. The participants were 48 undergraduates from the University of Toronto who participated in the experiment for course credit. Their mean age was 18.9 years, and mean score on the Shipley vocabulary scale (Zachary, 1986) was 29.1. Participants were randomly assigned to the FA or DA condition, $n = 24$ per condition.

Design and procedure. The experiment consisted of two phases; incidental encoding followed by a 4-AFC recognition test. These phases were separated by a 10-min retention interval in which participants played the Tetris computer game. The first phase was described as a color-judgment experiment, in which participants were presented with a series of 40 common nouns whose font color was red, blue, green or yellow. The task was to judge the color of each word as it appeared in the computer monitor and to respond as rapidly as possible by pressing one of four response buttons. Following each response there was a 1,000 ms intertrial interval before the next word appeared. The words themselves were described as being irrelevant to the task. Half of the participants performed the task under FA conditions, and half under DA conditions. The DA task was to listen to a string of auditory digits presented at a 1.5 s rate, and detect targets defined as three successive odd digits (e.g., 7–9–5, 1–3–1, etc.). DA participants signaled detection of an auditory target by pressing the space bar.

After completing the color-judgment task, all 48 participants played the computer game Tetris for 10 min. They then all performed an explicit recognition memory task under full attention conditions. The recognition test contained 50 4-AFC trials; participants were instructed to select one of the four words in *all* cases—a word that may have been an ignored color word in the first phase. Of the 50 trials, 40 did contain one target word from the study phase, and 10 contained no target words. Participants were informed that on certain trials no target word would be present, but that they should always choose the word they judged to be the most likely from Phase 1. Participants were also asked to rate their choice as a word they were "fairly certain" had been in Phase 1, as "possibly there" or as a "pure guess." These ratings were coded as 2, 1, and 0, respectively.

Results

When a target word was present among the four choices, participants rated their choice as a "pure guess" on an average of 19.1 trials out of 40 (48%) in the FA condition and 20.9 trials (52%) in the DA condition. In these cases the proportions of correct choices were 0.27 and 0.33 for FA and DA conditions, respectively; that is, the proportions of correct selections were 0.27 and 0.33, given that a target word was present and that the selection was made with zero confidence. The proportions of "pure guess" judgments in this and the following experiments are given in Table 1 under the heading "Prop. 0." The values of correct selections given a confidence rating of zero are referred to as $p(c)|0$, and these values are also given in Table 1. The chance value is 0.25, and t tests showed that the 0.27 value was not significantly different from 0.25, $t(23) = 0.80$, $p > .05$, whereas the value of 0.33 was reliably different from chance, $t(23) = 4.82$, $p < .001$. Additionally, the DA value of 0.33 was significantly higher than the FA value of 0.27, $t(46) = 2.11$, $p < .05$. When a target was present in the set, the conditional probabilities of choosing it correctly given a confidence rating of 1 ("possibly there") or 2 ("fairly certain") were 0.35 and 0.49, respectively, for FA participants, and 0.32 and 0.43, respectively, for DA participants. Thus performance was above chance but far from perfect when participants claimed some memory awareness of their choices.

Table 1
Performance Measures in Experiments 1--4

Experiment	Hit rate	Prop."0"	p(c)|0	Prop. "1 + 2"	p(c)|1 + 2
Experiment 1					
FA	0.34	0.48	0.27	0.52	0.20
DA	0.34	0.52	0.33	0.48	0.13
Experiment 2					
FA	0.58	0.26	0.39	0.74	0.56
DA	0.35	0.39	0.30	0.62	0.19
Experiment 3					
FA	0.69	0.26	0.40	0.74	0.75
DA	0.57	0.31	0.34	0.69	0.59
Experiment 4					
FA	0.85	0.17	0.41	0.83	0.91
DA	0.84	0.15	0.43	0.85	0.88

Note. Overall hit rate is defined as correct selection of a target item regardless of confidence rating (0 + 1 + 2). Prop. "0" = proportion of choices rated zero; p(c)|0 = probability of choosing correctly, given a '0' confidence rating; Prop. "1 + 2" = proportion of choices rated 1 or 2; p(c)|1 + 2 = probability of choosing correctly, given a confidence rating of 1 or 2. These last values are corrected for chance responding (see text).

Discussion

The major finding of interest was that participants in the DA condition did exhibit some degree of recognition without awareness. In that condition, 18 participants had values of correct choices rated as a "guess" that exceeded the chance level of 0.25 whereas only four participants had values less than 0.25. The finding of more recognition without awareness following DA conditions at encoding echoes the findings of Voss and colleagues, and is in line with the idea that this encoding condition may have yielded superficial perceptual encodings of words. Arguably, this type of encoding may be sufficient to choose target words correctly in a later forced-choice recognition test, but insufficient to yield the subjective experience of remembering. In order to obtain further evidence on this phenomenon we replicated the study in a second experiment, but with the one difference that participants were informed in the first color-judging phase that memory for the words would be tested later. Our assumption was that this change would result in more deliberate encoding of the words, and therefore an increase in hit rate in the recognition test. We also predicted that this stronger encoding would be associated with a decline in the proportion of "0" responses (because of the increased hit rate) and a reduction in the propensity to select target words while apparently guessing (following our assumption that intentional encoding in the first phase would result in a greater match between encoding and retrieval).

Experiment 2

Method

The design and procedure were exactly as in Experiment 1, including the FA and DA conditions, but with the one alteration that participants were informed before performing the color-judgment task that there would be a memory test for the colored words. The participants were again undergraduates who participated for course credit; 24 were tested in the FA condition (mean age = 19.0 years; Shipley vocabulary = 29.5) and 21 were tested

in the DA condition (mean age = 19.3 years; Shipley vocabulary = 30.1).

Results

When a target word was present, participants rated their choice as a "pure guess" 10.5 times, on average, out of a possible 40 trials in the FA conditions and 15.2 times, on average, in the DA condition. Thus the measures of proportion "0" were 0.26 and 0.39, respectively (see Table 1). In these "pure guess" cases, the mean proportions of correct choices were 0.39 in the FA condition and 0.30 in the DA condition (see Table 1). Although both of these $p(c)|0$ values exceed the chance value of 0.25, only the FA value was significantly higher than 0.25, $t(23) = 2.85$, $p < .01$; the DA value was not significantly greater than chance, $t(20) = 1.56$, $p > .05$. In addition, the FA value of 0.39 was not significantly higher than the DA value of 0.30, $t(43) = 1.54$, $p > .05$. The proportions of correct selections made with confidence ratings 1 and 2 when targets were present were 0.44 and 0.84, respectively for the FA condition, and 0.31 and 0.57, respectively for the DA condition. These "aware" values are understandably higher than the corresponding values in Experiment 1.

Three 2 (FA/DA) $\times$ 2 (Experiments 1 & 2) analyses of variance (ANOVAs) were also carried out to compare the values of hit rate, proportion "0" and $p(c)|0$ between the experiments. Hit rate was defined as the probability of selecting the correct target when one was present, regardless of confidence rating. The ANOVA on hit rates showed a significant effect of experiment, $F(1, 90) = 42.12$, $p < .001$, $\eta_p^2 = .32$, of FA/DA, $F(1, 90) = 33.47$, $p < .001$, $\eta_p^2 = 0.27$, and the interaction between the two factors, $F(1, 90) = 32.02$, $p < .001$, $\eta_p^2 = .26$. Table 1 indicates that these effects show that hit rates in Experiment 2 were generally higher than those in Experiment 1, and also that hit rates were higher for FA than DA conditions. However, these effects were modulated by a significant interaction between the factors; only the FA condition in Experiment 2 showed the benefit of intentional learning conditions. The ANOVA on proportion "0" scores revealed significant effects of experiment, $F(1, 90) = 21.06$, $p < .001$, $\eta_p^2 = .19$, and of FA/DA, $F(1, 90) = 4.78$, $p < .03$, $\eta_p^2 = .05$, but no interaction, $F(1, 90) = 1.11$, $p > .05$. That is, values of proportion "0" were higher for Experiment 1 than for Experiment 2, and somewhat higher for DA conditions than for FA conditions (see Table 1). The ANOVA on $p(c)|0$ values showed that neither the effect of Experiment, $F(1, 90) = 1.68$, $p > .05$ nor FA/DA ($F < 1.0$) was significant, but the interaction was statistically reliable, $F(1, 89) = 5.58$, $p = .02$, $\eta_p^2 = .06$. Table 1 shows that this last effect is attributable to the value for DA being higher than that for FA in Experiment 1, but that FA is greater than DA in Experiment 2.

Discussion

Our predictions for Experiment 2 relative to the first experiment were that the intentional learning instructions in the color-judgment phase would increase the hit rate, reduce the proportion of "0" confidence ratings, and also reduce the value of proportion correct, given a "0" rating [$p(c)|0$]. The first two predictions were borne out by the results, although the hit rate increase was found only for FA conditions. The prediction that $p(c)|0$ would decrease was *not* upheld, however. There was no main effect for experi-

ment, but the significant interaction between Experiment and FA/DA showed that $p(c)|0$ increased from 0.27 to 0.39 in the FA condition but declined slightly (from 0.33 to 0.30) in the DA condition. The speculation that recognition without awareness might decrease as a function of a better match between encoding and test conditions was, therefore, not supported by these results. Our assumption was that the intentional encoding instructions in Experiment 2 would be more similar than the incidental conditions in Experiment 1 to the intentional recognition conditions at test, and so $p(c)|0$ should decline from Experiment 1 to Experiment 2, which generally did not happen.

Another initial prediction was that the probability of recognition without awareness would be greater in DA than in FA conditions. This prediction was supported marginally in Experiment 1, but the probabilities of $p(c)|0$ reversed in Experiment 2 where the values were 0.39 for FA and 0.30 for DA. The difference was not significant, but was nevertheless in the wrong direction. The proportion of guess responses did drop substantially as the potential to learn the words in Phase 1 increased. In turn, this result raises the possibility that the subjective meaning of a guess response might change as a function of how well words were learned. The possibility that such a change in criterion might signal a shift to more conservative responding in line with the conclusions of Starns et al. (2008) is considered again after describing two further experiments.

Despite obtaining results from the two color-word experiments that gave little support to the notions of either context change or encoding under DA conditions as a basis for recognition without awareness, we decided to change the encoding paradigm before abandoning the ideas. Accordingly, we ran two experiments using a paradigm that was closer to the paradigm used by Tulving and Thomson (1973). One difference was that in our version the test words were provided rather than generated by the participants. The paradigm thus consisted of several paired associate lists in the encoding phase followed by a test phase consisting of a series of 4-AFC recognition tests. To encourage the use of the "pure guess" ("0") response, only half of the test trials contained a target, and participants were informed of this fact. Experiment 3 was the first study using this paradigm, and so was basically exploratory in nature.

Experiment 3

Method

Experiment 3 again contained an encoding phase followed by a test phase. In this case the first phase consisted of a series of paired-associate lists, and this phase was followed by a 4-AFC recognition test for words on the final list. The participants were 48 young adults (undergraduate students) who were allocated randomly to one of two conditions, FA and DA, during the learning phase. The FA condition had 24 participants (mean age = 18.8 years; years of education = 12.3) and the DA condition also had 24 participants (mean age = 18.6 years; years of education = 12.5). The materials used for the paired-associate lists were common words (mostly nouns) of 1–3 syllables and ranging in frequency from 10 (coin) to 1,207 (man) according to the Kučera and Francis (1967) norms. During the encoding phase, participants studied two lists of 24 paired associates (Lists 1 & 2) presented

visually at a 5 s rate with a 1 s interstimulus interval. Participants in the DA condition also performed the "successive-odd-digits" task presented auditorily while learning the lists. In this case the DA task was made slightly easier by asking participants to detect the presence of two successive odd digits. At the end of each list all participants completed a self-paced cued-recall test. List 3 had 48 paired associates presented in the same way, but at the end of presentation participants were informed that we were interested in the effects of time delay on memory, and that the recall test would come later. In the meantime, they played the computer game Tetris for 5 min.

Participants were then given 48 sets of four words on two sheets of paper. Half of the sets contained a response word from List 3; the other half contained no target words. There were two versions of the 4-AFC recognition test (A & B); 24 participants (12 FA and 12 DA) received Version A, which contained 24 List 3 response words, and the remaining 24 participants received Version B, which contained the target words not on A. Participants were asked to circle one word in each set of four, the word most likely to come from List 3. They were informed that only half of the 4-word sets contained a target, but they should always select one, guessing when necessary. They were also instructed to provide a confidence rating with each word, with 0, 1, 2 having the same meaning as in Experiments 1 and 2. Finally, they were given the cued-recall test for the original List 3.

Results and Discussion

Paired associate recall probabilities were 0.36, 0.61 and 0.43 for Lists 1, 2, 3, respectively, for the FA group, and 0.12, 0.36, and 0.26, respectively, for the DA group. Thus, as expected, the recall values for DA participants were consistently lower than the corresponding values for FA participants.

As shown in Table 1, hit rates were 0.69 and 0.57 for the FA and DA groups, respectively. Thus intentional learning of paired-associate responses presented at a relatively slow rate (6 s per pair) was associated with higher hit rates than those obtained from the first two experiments. Proportions of target words recognized correctly with confidence ratings 1 and 2 were 0.58 and 0.94, respectively, for the FA group, and 0.55 and 0.79, respectively, for the DA group. All of these values were reliably higher than 0.25, all values of $t > 6.50$. Table 1 also shows that the proportion of words selected with zero confidence on the 24 trials when a target word was present was 0.26 for FA participants and 0.31 for DA participants. When a target word was present and the selection was made with zero confidence, participants were correct with proportions 0.40 for the FA group and 0.34 for the DA group. These values are shown in Table 1 under the heading $p(c)|0$. The 0.40 value for FA was greater than the chance value of 0.25, $t(22) = 2.73$, $p < .01$; but the 0.34 value for DA was not reliably greater than chance, $t(23) = 1.60$, $p = .12$. From the point of view of the context change hypothesis, it is unclear whether the shift between paired-associate learning and the 4-AFC test is more or less than the shift between color-word naming and the test, so the final experiment was designed to provide a clearer test of this hypothesis. For now it may be noted that the value of $p(c)|0$ was again higher for the FA group than for the DA group, again providing no evidence for the notion that recognition without awareness is associated with DA at encoding. The third hypothesis, that the

incidence of recognition without awareness is restricted to conditions of conservative responding (Starns et al., 2008), is difficult to assess from these data; consideration of this possibility is deferred until the final experiment is described.

The major purpose of Experiment 4 was to provide a strong test of the context shift account by making conditions for the 4-AFC test as compatible as possible with the encoding conditions. This was accomplished by reminding participants of the original paired-associate pairs at the time of the recognition test. We did this by preceding each set of four words in the 4-AFC test with a stimulus word from the original learned list. When a target word was present in the set it was always preceded by its correct stimulus word from the original List 3 learning trial. Thus, if the original pair to be learned was moth-FOOD, the four words provided for the recognition test (BASE, FOOD, BOOK, FARM) would be preceded by "moth." When a target word was not present, the recognition set was composed of four new words preceded by a randomly chosen stimulus word from the original paired-associate list. By reinstating the learning context in this way we expected to increase the hit rate but greatly reduce the phenomenon of recognition without awareness.

Experiment 4

Method

Experiment 4 was a replication of Experiment 3, with the one change that each set of four words in the 4-AFC recognition test was preceded by a stimulus word from the 48 pairs to be learned in List 3. In the 24 cases that a target word was present, the stimulus word was its correct pair mate; the remaining 24 4-AFC cases (which contained no target words) were paired randomly with the remaining 24 stimulus words. As in Experiment 3, half of the participants were given the A set of 4-AFC choices and half were given Set B. Forty-eight participants age 18–28 years participated in the study. Half of them were assigned to the FA condition (mean age = 21.8 years, mean years of education = 14.8) and half to the DA condition (mean age = 21.3 years, mean years of education = 14.8). The DA group again performed the auditory monitoring task ("tap the table every time you hear 2 successive odd digits") while learning the initial three paired-associate lists.

Results and Discussion

The overall recognition performance in this cued 4-AFC situation was predictably high—85% correct for FA and 84% correct for DA participants (hit rates in Table 1). Nevertheless, participants did rate their confidence level as zero in a number of instances when a target was present; the proportions were 0.17 for the FA group and 0.15 for the DA group (proportion "0" in Table 1). For the 21 participants in the FA group who selected items with zero confidence, the proportion of correct choices was 0.41; this value was significantly greater than the chance value of 0.25, $t(20) = 3.31$, $p < .01$. For the 18 DA participants who selected items with zero confidence, the proportion was 0.43, and the associated significance value was $t(17) = 3.38$, $p < .01$. Clearly these values of $p(c)|0$ did not fall close to 0.25 as predicted, and are broadly comparable to the results of Experiment 3, despite the

apparent success of the contextual reinstatement manipulation—overall recognition rates rose from 69% to 85% for FA participants in Experiments 3 and 4, respectively, and from 57% to 84%, respectively, for DA participants.

Other results made sense in light of the easier conditions associated with the cued 4-AFC procedure. Probabilities of correct recognition given confidence ratings 1 and 2 were 0.60 and 0.99, respectively, for FA participants, and 0.61 and 0.99, respectively, for DA participants. Cued recall probabilities for Lists 1, 2, and 3 were 0.43, 0.62, and 0.54, respectively, for FA participants, and 0.20, 0.44, and 0.44, respectively, for DA participants. Thus the DA manipulation reduced recall values, as in the previous experiments, but it is interesting to note that the manipulation did not reduce recognition scores in this instance. Apparently, the combination of context reinstatement with the forced-choice procedure was sufficient to compensate for the poorer initial encoding revealed in the cued recall cases.

Three ANOVAs were conducted to compare the results of Experiments 3 and 4. Each was a 2 (Experiments 3 & 4) × 2 (FA vs. DA) between subjects analysis. For overall hit rates, there was a significant effect of Experiment, $F(1, 92) = 44.70$, $p < .001$, $\eta_p^2 = .33$, and marginally reliable effects of FA/DA, $F(1, 92) = 3.23$, $p < .08$, $\eta_p^2 = .03$, and the interaction between the factors, $F(1, 92) = 2.90$, $p < .10$, $\eta_p^2 = .03$. Table 1 shows that these effects signify that hit rates were higher in Experiment 4 than in Experiment 3, and that there was a trend for these values to be higher for FA than for DA, especially in Experiment 3. For the measure proportion "0," the ANOVA yielded a significant effect of Experiment, $F(1, 92) = 14.91$, $p < .001$, $\eta_p^2 = .14$, but no effects of either FA/DA ($F < 1.0$) or of the interaction, $F(1, 92) = 1.79$, ns. Table 1 shows that the proportion of zero responses was greater in Experiment 3 than in Experiment 4. For the measure $p(c)|0$, the effect of Experiment was not significant, $F(1, 82) = 2.63$, $p > .05$, and neither the effect of FA/DA ($F = 1.08$) nor the interaction ($F < 1.0$) approached significance.

But the major result of interest is that the greater amount of contextual reinstatement from Experiment 3 to Experiment 4 had no effect on the values of $p(c)|0$. The contextual reinstatement manipulation clearly worked, given the substantially higher levels of overall recognition in the present experiment, but there was no evidence for a reduction in 'recognition without awareness.' The hypothesis that recognition without awareness in these paradigms might be akin to the phenomenon of recognition failure in the Tulving and Thomson (1973) experiments was therefore not supported by the present results, or at least not in the version that proposes that the size of the effect should be reduced as the encoding and retrieval contexts are made more similar. Over the four experiments, there is thus little or no support for either the context change hypothesis or the notion that recognition without awareness is more likely to occur under conditions of DA at encoding. The remaining hypothesis, that the value of $p(c)|0$ rises as recognition decisions are made under more conservative criteria, is considered in the General Discussion that follows.

General Discussion

A consideration of the data from all four experiments (see Table 1) shows clearly that our measure of recognition without awareness [$p(c)|0$] does not vary systematically with FA/DA at encod-

ing, and there is also little evidence for the notion that $p(c)|0$ varies as a function of context shift between encoding and retrieval. However, another possibility stems from the idea that there may be criterion shifts in the likelihood of giving a zero confidence response. In particular, it seemed possible that the criterion may depend on the overall ease or difficulty of the final 4-AFC recognition test. Presumably easy tasks will yield many 1 and 2 judgments when targets are present, and relatively few 0 judgments. But for easy tasks, targets are typically rather obvious and will be rated 1 or 2. If an item is less obvious, it may be chosen but given a zero confidence rating when contrasted with easier items. This thinking predicts a relationship between overall difficulty of the recognition test and values of $p(c)|0$—easy tasks should give relatively few "0" judgments, but a high value of $p(c)|0$.

Table 1 shows that the overall hit rates rise generally from Experiment 1 to Experiment 4, indicating that there was a tendency for the tasks to become easier. The Table also shows a tendency for the proportion of "0" responses when a target was present to decline from the first to the last experiment (understandably, as participants made more confident 1 or 2 responses as task difficulty decreased) and also for the measure $p(c)|0$ to increase from Experiments 1 to 4. These trends were assessed by carrying out a series of Spearman's rho correlation coefficients among the variables, using the means of the eight conditions (4 Experiments × FA/DA). The correlation between hit rate and proportion "0" was rho (6) = −.94, $p < .01$ and the correlation between hit rate and $p(c)|0$ was rho (6) = +.93, $p < .01$ (Figure 1a and 1b, respectively). Additionally, the correlation between proportion "0" and $p(c)|0$ was rho (6) = −.91, $p < .01$. There is thus good evidence across the eight conditions that as the task became easier (measured by increasing hit rate), the proportion of "0" confidence responses declined and the values of $p(c)|0$ correspondingly increased. Also, it is the case that values of $p(c)|0$ increased systematically as the proportion of "0" responses declined.

Further insight into the processes operating in the experiments may be gained by considering the relations between hit rates and the proportions of responses given 1 or 2 ratings when a target was present, and also between hit rates and the proportions of these 1 or 2 responses that were actually correct [$p(c)|1 + 2$]. The proportions given either 1 or 2 confidence responses are simply the complements of the proportions given zero responses, and are shown in column 4 of Table 1. These values signal the occasions that participants thought they had chosen the correct item. In order to compare the values of proportion correct given 1 or 2 [$p(c)|1 + 2$] with the proportions chosen with 1 or 2 confidence ratings, we corrected values of $p(c)|1 + 2$ for chance. Specifically, for each condition we first calculated the proportion correct given a rating of 1 or 2; we then subtracted the chance value of 0.25 from that proportion, and divided the result by 1.0 minus chance (0.75). The resulting scale of proportions correct given a 1 or 2 rating thus runs from 0 to 1.0, as does the scale of proportions of 1 or 2 chosen. The corrected values of $p(c)|1 + 2$ are shown in column 5 of Table 1. These values, and also the proportions of 1 or 2 chosen, are plotted against overall hit rate in Figure 2.

The figure shows that both functions are well fit by linear functions, but with different slopes. At lower values of hit rate (difficult tasks) the proportions correct are around 0.15–0.20, whereas the proportions of selections made with 1 or 2 confidence ratings are between 0.45 and 0.55. That is, the relatively high
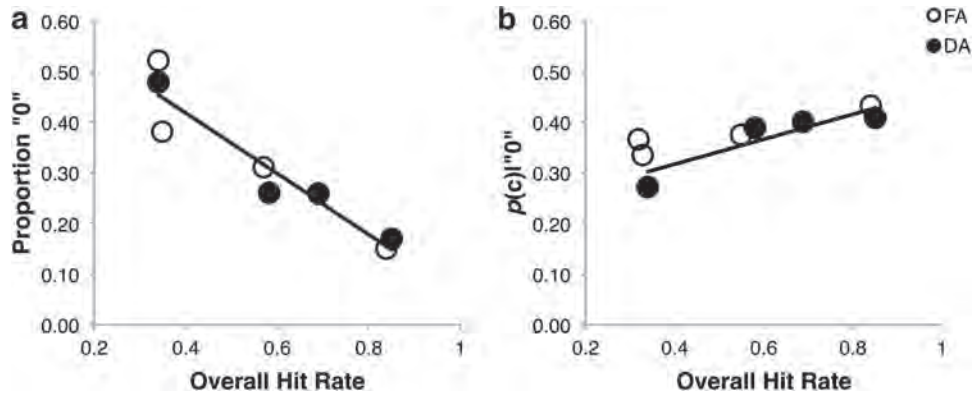
Figure 1. (a) Proportions given "0" ratings, and (b) p(c)|"0" (proportions correct given ratings of "0"), as a function of overall hit rate for the FA and DA conditions.

confidence levels are unwarranted by the proportions actually correct. This discrepancy reduces, however, as the tasks get easier, until at higher levels of hit rate (relatively easy tasks) the proportions of choices made with 1 or 2 ratings are slightly lower than the corresponding proportions correct. That is, participants are somewhat conservative in their allocation of confidence ratings at the easy end of task difficulty. This observation is in line with previous reports that stricter criteria are typically applied to strongly encoded stimuli and thus easier detection and recognition performance (Singer, 2009). On the assumption that trials themselves vary on a continuum of difficulty for each person in a given experiment, this pattern implies that for low values of hit rate participants allocate more ratings of 1 and 2 than they "ought to" given the difficulty level, so the remaining "0" allocations are given to the most difficult trials, and correspondingly show a low probability of being correct. When hit rates are high, however, the pattern reverses. Now participants allocate fewer ratings of 1 and 2 to choices than they might do given the relatively easy tasks, and so the remaining "0" allocations are given to trials that are also

relatively easy and so show a higher hit rate. In summary, we suggest that the strong correlation between overall hit rate and p(c)|0 is a function of a changing criterion for the allocation of "0" responses (Singer, 2009). The probability of 'recognition without awareness' increases as the tasks become easier, and participants adopt a more conservative criterion for claiming that they have chosen an item from the encoding list.

How general is this criterion account of recognition without awareness? It is clearly compatible with the results of Starns and colleagues (2008) who explicitly concluded that source memory for unrecognized items varied with the bias to respond "old" in recognition decisions. In their case the phenomenon appeared only under conditions that promoted conservative responding. The present account is probably less applicable to the studies by Cleary and Greene (2004, 2005) and by Ryals et al. (2011) who showed recognition of list membership in the absence of identification on the basis of processing word fragments or speeded processing of the words themselves. In these cases the recognition of list membership is probably due to an unconscious recognition memory process, possibly attributable to a minimal sense of familiarity as the authors suggest.

With regard to the studies of Voss and colleagues (Voss et al., 2008; Voss & Paller, 2009, 2010), we agree with those authors that the term "implicit memory" should be reserved for cases in which individuals' performance shows evidence of memory for previous events, yet they are unaware that their responses are based on memory. By this definition, choices accompanied by feelings of either recollection or familiarity (or given with either "remember" or "know" responses) are classified as cases of explicit memory. On the contrary, correct choices made but classified as 'guesses' are instances of implicit memory—recognition without awareness. Such instances were documented both by Voss and colleagues and in the present experiments.

Voss and Paller (2010) consider, but reject, the possibility that recognition without awareness is simply based on the processing of relatively weak representations that might otherwise evoke responses of familiarity or recollection (see also Voss, Lucas & Paller, 2012). Their arguments are based partly on different ERP signatures related to implicit memory compared to those associated with familiarity and recollection (Voss et al., 2012), but also to the changes in R, K, and guess responses between conditions in
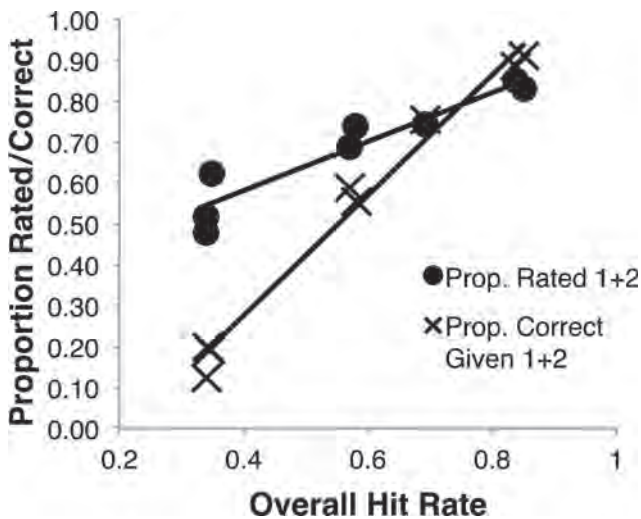

Figure 2. Proportions rated 1 or 2, and proportions correct given 1 or 2 ratings, as a function of overall hit rate.

which guessing was either encouraged or discouraged. In the latter case, the probability of a guess response being accurate was $p = .43$ (less than chance in a 2-AFC paradigm), whereas when guessing was encouraged the accuracy of responses classified as guesses rose to $p = .78$ (Voss & Paller, 2010). Interestingly, the proportions correct for R and K responses did not change systematically between the two encouragement conditions: for R responses the proportions correct were 0.80 under "confidence encouraged" instructions and 0.76 under "guessing encouraged" instructions; the corresponding proportions for K responses were 0.60 and 0.62, respectively. The authors argue that this result is not consistent with a simple shift in criterion. However, an alternative reading of the result is that whereas the encouragement to guess led to an increase in the rate of guessing (from 12% to 26%) and a concomitant decrease in the rates for R and K responses, the subjective criteria for K and R responses (based on proportions correct) remained relatively unchanged, whereas the encouragement to use the "guess" response allowed participants to select that response more often. It is also necessary to add that the subjective criterion for "guess" must have changed, such that under encouraging conditions items that might otherwise be classified as K or R are now classified more cautiously as guesses, with a consequent rise in the probability that such responses are correct. Our suggestion is therefore that encouragement to guess differentially changes the criterion for what constitutes a 'guess' response but leaves the subjective labels unchanged for K and R responses.

Conditions under which recognition without awareness was *not* observed in the studies reported by Voss and colleagues include recognition of kaleidoscope patterns using a yes-no procedure, using a 2-AFC procedure when the two choices were perceptually very different, and when participants were given an extended time to decide. All of these cases likely engender a deliberate conscious retrieval strategy rather than a reliance on perceptually based implicit recognition. Voss and Paller (2009) boost their case for a perceptual basis of their implicit memory demonstrations by showing that the neural correlates of the effect included frontal-occipital brain potentials at 200–400 ms post-stimulus-onset, potentials that were distinct from the late positive responses associated with judgments of recollection or familiarity. Given that these researchers used complex and relatively meaningless kaleidoscope patterns as stimuli it makes sense that recognition choices were made on the basis of perceptual processing.

In our own case the stimuli were words, the four choices presented on each test trial were not perceptually similar, and participants performed the sequence of 4-AFC test trials at their own pace, rather than under time pressure. In addition, Table 1 makes it clear that there were no systematic changes associated with encoding under full versus divided attention in our experiments. It thus seems clear that implicit recognition can occur with a variety of materials and under a variety of experimental conditions. One way of reconciling the present results with those of Voss and Paller is to suggest that the selection of a correct item is based on processing the relevant neural representation in all cases, although of course the nature of that representation will vary widely. We also suggest that such representations are the basis for correct selection for *both* implicit and explicit recognition memory; the difference between the two types is that explicit memory is accompanied additionally by some representation of the context of initial occurrence—either a nonspecific feeling of past occur-

rence in the case of K responses or specific recollection of context for R responses. Speculatively, this second type of representation may be associated with changes in the neural activations recorded as the late positive complex in the ERP signals reported by Voss and Paller (2009). In turn, various factors will contribute to the encoding and retrieval of such contextual representations; they may include such things as attention to contextual attributes during encoding, the associative relationship of the target item to its initial context, the degree to which the retrieval context matches the encoding context, and the extent to which the participant engages deliberate attempts to recollect the initial situation. These, of course, are among the factors studied by many researchers investigating the characteristics of explicit recollection.

The point we wish to stress here is that two distinct sets of factors may be operating to give rise to the phenomenon of recognition without awareness; one set contributes to the relative strengths (or degrees of fluent processing) of representations of target items and their lures, the second set contributes to the occurrence and adequacy of representations of the initial encoding contexts of these target items. According to this view, recognition without awareness will occur when item representations are strongly present (or are processed fluently), but contextual representations are weak or absent. In the Voss and Paller experiments, participants deliberately attempted to learn the kaleidoscope patterns so good item representations were established. It seems likely, however, that the corresponding contextual representations were poorly differentiated among the various very similar items, enabling participants to select a target item correctly but in the absence of any feeling of recollection that it was one they had studied. The encoding of well-differentiated contextual representations would be even less likely under divided attention conditions, and the retrieval of such representations would be poor under conditions that discouraged a deliberate analytic retrieval strategy—for example, the speeded 2-AFC conditions used in the Voss and Paller studies.

In the case of the current experiments, participants were presumably able to form good item representations under the intentional learning conditions of Experiment 2, 3, and 4. Context information was also available at encoding but this information may have been difficult to access at retrieval given that the items were now presented in a very different 4-AFC context. Additionally, we suggest that the probability of correctly selecting a target item with zero confidence varied with the overall difficulty of the particular recognition test, and a concurrent shift in the criterion associated with the subjective feeling of what constitutes a "guess" response. Voss and Paller (2012) suggest that their effects are based on fluency of perceptual processing of the encoded representations, and this seems very reasonable given that the items were complex and relatively meaningless visual patterns. In the present case, the correct selection of the target word when confidence was zero may also be attributable to the greater perceptual fluency of processing targets relative to lures (Jacoby and Whitehouse (1989). Alternatively, Chechile, Sloboda, and Chamberland (2012) have suggested that implicit and explicit recognition differ simply in the adequacy (e.g., strength, vividness) of the encoded representation, with weakly represented items being insufficient to support explicit recognition, but still sufficient to select the correct item while claiming that the choice was simply a guess. We add to the Chechile et al. model by suggesting that the criterion for a

"guess" response is variable as described earlier, but differ from them by suggesting that explicit recognition also involves the retrieval of a further representation of the initial context.

Conclusions

The four experiments presented in this article provide ample evidence for the reality of recognition without awareness; in this case, with words exposed for several seconds. In many ways the existence of the phenomenon is unsurprising as related effects have been reported over the years. One example is the ability of participants to make accurate psychophysical judgments (e.g., relative judgments of weight, length, and shape) while claiming that they were simply guessing (early studies reviewed by Adams, 1957). A second example is evidence for semantic processing of words in the absence of conscious identification of these words (e.g., Marcel, 1983; Stenberg, Lindgren, Johansson, Olsson & Rosén, 2000). In the present case we have suggested that variation in the strength of the effect is principally attributable to the general difficulty of the recognition decision in a particular experiment—easier decisions were associated with fewer guess responses but with higher values of recognition without awareness [higher values of p(c)|0]. The data are thus consistent with the notion that the subjective criterion for choosing correctly while stating that the choice was simply a guess is flexible, depending (among other possible factors) on the context of the overall recognition situation.

It may be asked what "criterion shift" means across different conditions that always involve the forced-choice procedure; does the participant not simply choose the subjectively strongest item in all cases? In answer, we emphasize that our use of the term "criterion shift" does not refer to whether or not participants make a choice—they choose on all trials—but rather to the subjective state accompanying the selection, and to the fact that this state varies as a function of overall task difficulty. Under easy conditions many trials yield obvious selections labeled 1 or 2. On the remaining trials participants can still select the correct item, but these cases feel relatively less obvious and are therefore labeled 0. Under difficult conditions participants make more "guess" selections, but in this case the general difficulty results in the selection of fewer target items; choice is much closer to chance responding.

Finally, unlike Voss and Paller (2009), we see no reason in our data to suggest that implicit and explicit recognition reflect different forms of memory. We argue rather that correct selection of a target item may be based on relative fluency of processing, or on the strength or adequacy of its encoded representation, and that these factors are likely to vary on a continuum. Additionally, however, the recognition process will evoke some representation of the item's previous context of occurrence. This representation will also vary in the degree to which it fully specifies the past event; inadequate representations may simply evoke a feeling of general "pastness" whereas more adequate representations will reinstate a conscious memory of the original event. In turn, these different degrees of adequacy will be associated, respectively, with the subjective impressions of familiarity and recollection. In cases where such additional contextual representations are not evoked, the participant may still choose the target item correctly, but now with no subjective feeling of explicit recognition. These cases may therefore be described as recognition without awareness.

References

Adams, J. K. (1957). Laboratory studies of behavior without awareness. *Psychological Bulletin, 54,* 383–405. http://dx.doi.org/10.1037/h0043350

Chechile, R. A., Sloboda, L. N., & Chamberland, J. R. (2012). Obtaining separate measures for implicit and explicit memory. *Journal of Mathematical Psychology, 56,* 35–53. http://dx.doi.org/10.1016/j.jmp.2012.01.002

Cleary, A. M., & Greene, R. L. (2004). True and false memory in the absence of perceptual identification. *Memory, 12,* 231–236. http://dx.doi.org/10.1080/09658210244000577

Cleary, A. M., & Greene, R. L. (2005). Recognition without perceptual identification: A measure of familiarity? *The Quarterly Journal of Experimental Psychology, 58,* 1143–1152. http://dx.doi.org/10.1080/02724980443000665

Craik, F. I. M., Moscovitch, M., & McDowd, J. M. (1994). Contributions of surface and conceptual information to performance on implicit and explicit memory tasks. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 864–875. http://dx.doi.org/10.1037/0278-7393.20.4.864

Gopie, N., Craik, F. I. M., & Hasher, L. (2011). A double dissociation of implicit and explicit memory in younger and older adults. *Psychological Science, 22,* 634–640. http://dx.doi.org/10.1177/0956797611403321

Hasher, L., Zacks, R. T., & May, C. P. (1999). Inhibitory control, circadian arousal, and age. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp. 653–675). Cambridge, MA: MIT Press.

Jacoby, L. L., & Whitehouse, K. (1989). An illusion of memory: False recognition influenced by unconscious perception. *Journal of Experimental Psychology: General, 118,* 126–135. http://dx.doi.org/10.1037/0096-3445.118.2.126

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English.* Providence, RI: Brown University Press.

Marcel, A. J. (1983). Conscious and unconscious perception: An approach to the relations between phenomenal experience and perceptual processes. *Cognitive Psychology, 15,* 238–300. http://dx.doi.org/10.1016/0010-0285(83)90010-5

Peynircioğlu, Z. F. (1990). A feeling-of-recognition without identification. *Journal of Memory and Language, 29,* 493–500. http://dx.doi.org/10.1016/0749-596X(90)90068-B

Ryals, A. J., Yadon, C. A., Nomi, J. S., & Cleary, A. M. (2011). When word identification fails: ERP correlates of recognition without identification and of word identification failure. *Neuropsychologia, 49,* 3224–3237. http://dx.doi.org/10.1016/j.neuropsychologia.2011.07.027

Schacter, D. L., Dobbins, I. G., & Schnyer, D. M. (2004). Specificity of priming: A cognitive neuroscience perspective. *Nature Reviews Neuroscience, 5,* 853–862. http://dx.doi.org/10.1038/nrn1534

Singer, M. (2009). Strength-based criterion shifts in recognition memory. *Memory & Cognition, 37,* 976–984. http://dx.doi.org/10.3758/MC.37.7.976

Starns, J. J., Hicks, J. L., Brown, N. L., & Martin, B. A. (2008). Source memory for unrecognized items: Predictions from multivariate signal detection theory. *Memory & Cognition, 36,* 1–8. http://dx.doi.org/10.3758/MC.36.1.1

Stenberg, G., Lindgren, M., Johansson, M., Olsson, A., & Rosén, I. (2000). Semantic processing without conscious identification: Evidence from event-related potentials. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 973–1004. http://dx.doi.org/10.1037/0278-7393.26.4.973

Tulving, E., & Schacter, D. L. (1990). Priming and human memory systems. *Science, 247,* 301–306. http://dx.doi.org/10.1126/science.2296719

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80,* 352–373. http://dx.doi.org/10.1037/h0020071

Voss, J. L., Baym, C. L., & Paller, K. A. (2008). Accurate forced-choice recognition without awareness of memory retrieval. *Learning & Memory, 15,* 454–459. http://dx.doi.org/10.1101/lm.971208

Voss, J. L., Lucas, H. D., & Paller, K. A. (2012). More than a feeling: Pervasive influences of memory without awareness of retrieval. *Cognitive Neuroscience, 3,* 193–207. http://dx.doi.org/10.1080/17588928.2012.674935

Voss, J. L., & Paller, K. A. (2009). An electrophysiological signature of unconscious recognition memory. *Nature Neuroscience, 12,* 349–355. http://dx.doi.org/10.1038/nn.2260

Voss, J. L., & Paller, K. A. (2010). What makes recognition without awareness appear to be elusive? Strategic factors that influence the accuracy of guesses. *Learning & Memory, 17,* 460–468. http://dx.doi.org/10.1101/lm.1896010

Zachary, R. A. (1986). *Shipley Institute of Living Scale*: *Revised manual.* Los Angeles: Western Psychological Services.

RESEARCH REPORT

The Tip-of-the-Tongue Heuristic: How Tip-of-the-Tongue States Confer Perceptibility on Inaccessible Words

Anne M. Cleary and Alexander B. Claxton
Colorado State University

This study shows that the presence of a tip-of-the-tongue (TOT) state—the sense that a word is in memory when its retrieval fails—is used as a heuristic for inferring that an inaccessible word has characteristics that are consistent with greater word perceptibility. When reporting a TOT state, people judged an unretrieved word as more likely to have previously appeared darker and clearer (Experiment 1a), and larger (Experiment 1b). They also judged an unretrieved word as more likely to be a high frequency word (Experiment 2). This was not because greater fluency or word perceptibility at encoding led to later TOT states: Increased fluency or perceptibility of a word at encoding did not increase the likelihood of a TOT state for it when its retrieval later failed; moreover, the TOT state was not diagnostic of an unretrieved word's fluency or perceptibility when it was last seen. Results instead suggest that TOT states themselves are used as a heuristic for inferring the likely characteristics of unretrieved words. During the uncertainty of retrieval failure, TOT states are a source of information on which people rely in reasoning about the likely characteristics of the unretrieved information, choosing characteristics that are consistent with greater fluency of processing.

Keywords: tip-of-the-tongue states, metacognition, attribution, heuristics, recognition without identification

The tip-of-the-tongue (TOT) state is the feeling of being on the verge of retrieving a word from memory when unable to do so (Brown, 1991, 2012; Schwartz, 2002). It is generally thought to result largely from attributions that people make based on other available information, such as retrieval of some of the unretrieved word's attributes (see Schwartz, 2002; Schwartz & Metcalfe, 2011, for a review). For example, if unable to retrieve a word but able to retrieve its first letter, one may infer from this that the word is on the verge of access. The present study is concerned with the reverse: whether participants infer from the presence of a TOT state itself an increased likelihood of certain word characteristics. Such a finding would underscore the need for teasing apart instances of actual access to a word's attributes and instances of mere inference of those attributes from the TOT state itself, thus, having important implications for the study of TOT states.

There is reason to hypothesize that TOT states can serve as a source of inference. As reviewed by Cleary, Staley, and Klein (2014), an inadvertent finding from the recognition-without-identification literature is that people infer from the presence of a TOT state an increased likelihood that an unretrieved word appeared on an earlier study list. Recognition-without-identification itself is old-new discrimination among unretrieved targets; it is shown by higher ratings of study-status likelihood for studied than for nonstudied unretrieved targets. For example, participants give higher ratings of likely answer study-status to general knowledge questions whose unretrieved answers were studied than to those whose unretrieved answers were not (Cleary, 2006). The inadvertent finding to emerge from the recognition-without-identification literature is that higher study-status likelihood ratings are given during TOT than non-TOT states, regardless of the actual study-status of the unretrieved targets (e.g., Cleary & Specker, 2007; Cleary, 2006; Cleary, Konkel, Nomi, & McCabe, 2010; Cleary & Reyes, 2009). In short, during TOT states, people are biased to judge unretrieved targets as studied. This TOT bias is distinguishable from recognition-without-identification, occurring in situations where the recognition-without-identification effect does not (e.g., Cleary et al., 2010), and even in situations where the recognition-without-identification effect is reversed (Cleary et al., 2014).

Because the TOT bias pattern was an incidental finding, Cleary et al.'s (2014) review of it as a consistent pattern across recognition-without-identification studies represents the first work devoted solely to this effect. Though many explanations are possible for the association between TOT states and study-status

likelihood ratings for unretrieved targets, Cleary et al. argue that participants infer from the presence of a TOT state an increased likelihood that the unretrievable target must have been studied. If so, this raises the possibility that TOT states are used to make other inferences (beyond regarding whether the unretrieved target was studied), using something that we refer to hereafter as the *TOT heuristic*.

The present study examined whether people infer from the presence of a TOT state itself likely attributes of the unretrieved information. This TOT heuristic hypothesis is important to investigate because most theoretical explanations of TOT experiences include a role of conscious access to attributes regarding the unretrieved word, yet the possibility that participants might sometimes infer those attributes from the TOT state itself (as opposed to actually accessing them) has not been considered. Use of such a TOT heuristic would mean that future research on to what people have access during TOT states will need to account for what is really being accessed versus what is merely being inferred from the presence of a TOT state.

We investigated the specific hypothesis that participants infer from the presence of a TOT state characteristics about the unretrieved target word that are thought to be more fluent. Our reasoning was as follows. Although a TOT state might seem to reflect a state of *disfluency* (because of the inaccessibility of the sought-after word), participants may assume that an unretrieved word for which a TOT state is present is in a *heightened state* of accessibility relative to an unretrieved word for which a TOT state is not present. In fact, this is the very definition of a TOT experience—feeling as if the word is on the verge of access (e.g., Schwartz, 2001). If participants assume that a TOT state indicates a heightened state of accessibility for an unretrieved word relative to a non-TOT state, this could explain why participants infer a greater likelihood of a target's having been studied recently (as recent presentation might be expected to lead to heightened accessibility). If the TOT state *feels* like (or is interpreted as) a heightened state of accessibility for the currently unretrievable target word relative to when a TOT for it does not occur, then participants may assume from the TOT state that the target has qualities or characteristics that are associated with greater accessibility. From this, we hypothesized that participants will attribute from the presence of a TOT state an increased likelihood of fluent attributes of the unretrieved word.

What attributes make a word seem more fluent or accessible? One factor is the clarity of its font (e.g., Whittlesea, Jacoby, & Girard, 1990). For this reason, in Experiment 1a, we examined the hypothesis that participants would judge an unretrieved target as more likely to have earlier appeared in a darker, clearer font if in a TOT state than if not. We were especially interested in such judgments for nonstudied items, where there was no particular font associated with the target. Another factor may be font size, with larger font sizes seeming more fluent (Mueller, Dunlosky, Tauber, & Rhodes, 2014; Rhodes & Castel, 2008). In Experiment 1b, we examined the hypothesis that participants would judge an unretrieved target as more likely to have earlier appeared in a larger font if in a TOT state than if not. In Experiment 2, we examined judgments of word frequency, a known indicator of fluency or accessibility (e.g., McClelland & Rumelhart, 1981). Although TOTs are thought to be more likely for low than high frequency words (e.g., Burke, MacKay, Worthley, & Wade, 1991), if participants assume

that a TOT state indicates a heightened state of accessibility for an unretrieved word relative to when an unretrieved word elicits no TOT, participants may then infer from a TOT state a greater likelihood that the unretrieved word is of higher frequency relative to when a TOT does not occur.

Experiments 1a and 1b

Method

Participants. Colorado State University students participated in exchange for credit toward a course. Forty participated in Experiment 1a, and 56 in Experiment 1b.

Materials. Stimuli were a pool of 80 general knowledge questions and their answers selected from Nelson and Narens' (1980) norms and used in prior research (Cleary et al., 2014). For each participant, 40 of the 80 answers were presented on a study list. In Experiment 1a, 20 of the study answers were presented in a dark black font (the high font clarity condition) and 20 were presented in a light gray font (the low font clarity condition), against a white background. The font color was set using the E-prime software's black and silver font color settings, respectively. Although there were only three conditions (studied in dark font, studied in light font, and nonstudied), we kept the ratio of studied to nonstudied items equal, as in prior research (e.g., Cleary & Specker, 2007; Cleary, 2006; Cleary et al., 2010; Cleary & Reyes, 2009). Therefore, to simplify our counterbalancing, four versions of the experiment were created to rotate the answer through the conditions of studied versus nonstudied and dark versus light font across participants.

In Experiment 1b, the same method was applied to font size instead of clarity. All fonts were presented in black on a white background, and 20 of the 40 study words were presented in 48 point (large) font while 20 were presented in 18 point (small) font (following from Rhodes & Castel, 2008).

Procedure. The procedure was similar to that used by Cleary (2006). Participants were instructed that they would view a list of words on the computer (they were told nothing specific about the font) and that afterward, they would be asked a series of questions. Instructions regarding the test were withheld until after the study list. The study list of 40 words (20 dark, 20 light in Experiment 1a; 20 48-point, 20 18-point in Experiment 1b) appeared individually in a random order for 1 s each with a 1 s interstimulus interval (a duration chosen to reduce effective encoding strategies so that not all studied items would be retrievable at test).

Following the study list, participants were given a description of the test and the prompts that would appear with each question. As the question remained on the screen, all dialog box prompts pertaining to it were given sequentially before proceeding to the next test question. Participants first attempted to answer the question using the dialog box. Then, another dialog box prompted them to indicate if they were experiencing a TOT state for the answer, which was defined as in Cleary (2006). Participants were then prompted to give a rating, regardless of whether the question had been answered or the answer had been studied. In Experiment 1a, this rating indicated the likelihood that the answer had appeared in darker, clearer font versus a lighter, less clear font on the earlier-presented list (0 = definitely lighter, less clear; 10 = definitely darker, more clear). In Experiment 1b, it indicated the likelihood

that the answer had appeared in a larger versus a smaller font (0 = definitely smaller; 10 = definitely larger). The next dialog box prompted a second attempt at identifying the answer; here, participants were encouraged to guess. The final dialog box prompted for partial information about the target, asking "Please type in any partial information you can think of about the word. (Examples: First letter, how it sounds, syllables, etc.)."

Results and Discussion

It is important to first consider how often participants could correctly answer the general knowledge questions. As in prior research (Cleary & Specker, 2007; Cleary, 2006; Cleary & Reyes, 2009), in Experiment 1a, participants correctly answered a lower proportion of questions whose answers were not studied ($M = .39$, $SD = .15$) than questions whose answers were studied, whether they were studied in dark font ($M = .49$, $SD = .21$), $t(39) = 4.28$, $SE = .02$, Cohen's $d = .69$, $p < .001$, or light font ($M = .48$, $SD = .17$), $t(39) = 4.06$, $SE = .02$, Cohen's $d = .65$, $p < .001$. Studying the answers in dark instead of light font did not make them more retrievable when given their questions at test, $t(39) = 0.47$, $SE = .03$, $p = .64$. The same held for Experiment 1b. Participants correctly answered a lower proportion of questions whose answers were not studied ($M = .39$, $SD = .13$) than questions whose answers were, whether they were studied in large font ($M = .51$, $SD = .17$), $t(55) = 6.78$, $SE = .02$, Cohen's $d = .91$, $p < .001$, or small font ($M = .50$, $SD = .17$), $t(55) = 5.86$, $SE = .02$, Cohen's $d = .79$, $p < .001$. Studying the answers in large instead of small font did not make them more retrievable at test, $t(55) = 0.21$, $SE = .03$, $p = .83$.

The rates at which participants provided partial information about the target answer (e.g., first letter, sound) were near zero. The means were: Experiment 1a, studied, dark font ($M = .02$, $SD = .03$), studied, light font ($M = .02$, $SD = .03$), and nonstudied ($M = .04$, $SD = .04$); in Experiment 1b, studied, large font ($M = .016$, $SD = .027$), studied, small font ($M = .015$, $SD = .025$), and nonstudied ($M = .024$, $SD = .023$). Because partial identification rates were at floor, they will not be discussed further.

The data of interest for Experiment 1a were the font clarity ratings given during retrieval failure (when neither the answer nor any partial information about it could be retrieved). Our hypothesis was that font clarity ratings would be higher during TOT than non-TOT states. As in prior work (Cleary & Specker, 2007; Cleary, 2006; Cleary & Reyes, 2009) not all participants reported a TOT state in all conditions; this caused eight to be lost from this analysis in Experiment 1a (seven in Experiment 1b). As shown in Figure 1, font clarity ratings were higher during reported TOT than non-TOT states. (The mean number of TOT states was 4.56 and 4.03 in the studied and nonstudied TOT conditions, respectively.) This main effect of TOT state was the only significant effect to emerge from a 2×2 TOT-state (TOT vs. non-TOT) $\times$ Study-status (target word studied vs. target word nonstudied) repeated measures analysis of variance (ANOVA) performed on the font clarity ratings given during retrieval failure, $F(1, 31) = 18.74$, $MSE = 4.32$, $p\eta^2 = .38$, $p < .001$ (other $Fs < 1.0$). In short, participants thought it more likely that an unretrieved word was clearer upon last reading it when in a TOT state for it than when not.
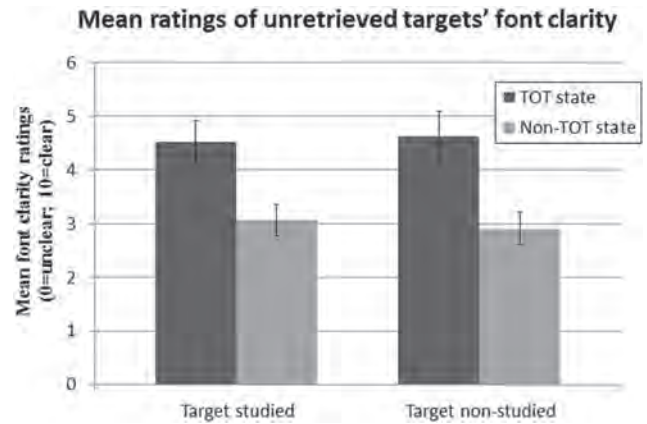

Figure 1. Mean judgments of the font clarity of unretrieved target words during their retrieval failure (including failure to retrieve partial target information) as a function of reported tip-of-the-tongue (TOT) states. Higher ratings indicate a judged greater likelihood that the unretrieved target was previously presented in a darker, clearer font. Each error bar represents the standard error of the mean.

The same was found for judgments of font size in Experiment 1b. As shown in Figure 2, font size ratings given during retrieval failure were higher during TOT than non-TOT states. (The mean number of TOT states was 4.69 and 6.65 in the studied and nonstudied TOT conditions, respectively.) This main effect of TOT state was the only significant effect to emerge from a 2×2 TOT-state (TOT vs. non-TOT) $\times$ Study-status (target word studied vs. target word nonstudied) repeated measures ANOVA performed on the font size ratings given during retrieval failure, $F(1, 48) = 11.12$, $MSE = 3.32$, $p\eta^2 = .19$, $p = .002$ (other $Fs < 1.0$). In short, participants thought it more likely that an unretrieved word was larger upon last reading it when in a TOT state for it than when not.

One possibility is that if a study word is processed more fluently in the first place, participants are more likely to retrieve partial visual information about it later during retrieval failure. If partial visual information is itself used to make an attribution of being in a TOT state, this could lead to a greater likelihood of a TOT experience in the fluent font condition, explaining the association between TOT states and font clarity or size estimates during retrieval failure.

Three aspects of our data rule out this explanation. First, more perceptible answer fonts did not increase TOT states for those answers later. The probability of a TOT state was not higher for unretrieved target words that were studied in dark font ($M = .24$, $SD = .18$) than light font ($M = .23$, $SD = .20$), $t(31) = 0.36$, $SE = .03$, $p = .72$. The probability of a TOT state was also not significantly higher for unretrieved targets that had been presented in large font ($M = .27$, $SD = .20$) than small font ($M = .24$, $SD = .23$), $t(48) = 0.78$, $SE = .03$, $p = .44$.

Second, when unable to retrieve a studied answer, participants were also unable to retrieve its font clarity level or relative size. Font clarity ratings given to unanswered questions whose answers were studied were not higher when the answers were studied in dark font ($M = 3.14$, $SD = 1.82$) than light font ($M = 3.45$, $SD = 1.74$), $t(31) = 1.33$, $SE = .24$, $p = .20$. Similarly, font size ratings
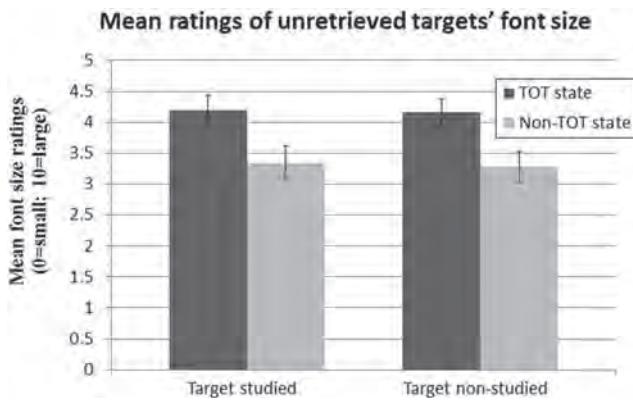
Figure 2. Mean judgments of the font size of unretrieved target words during their retrieval failure (including failure to retrieve partial target information) as a function of reported TOT states. Higher ratings indicate a judged greater likelihood that the unretrieved target word was previously presented in a larger font. Each error bar represents the standard error of the mean.

were not significantly higher when the answers were studied in large font ($M = 3.62$, $SD = 1.53$) than small font ($M = 3.40$, $SD = 1.64$) font, $t(48) = 1.39$, $SE = .16$, $p = .17$.

Finally, among test trials for which the unretrieved answer had not even been studied, participants still gave higher font clarity ratings when reporting a TOT state than when not, $t(31) = 4.04$, $SE = .42$, Cohen's $d = .73$, $p < .001$ (see the right-hand side of Figure 1), and higher font size ratings when reporting a TOT state than when not, $t(48) = 2.88$, $SE = .31$, Cohen's $d = .42$, $p = .006$ (see the right-hand side of Figure 2).

Some might wonder if participants somehow recognize when an unretrieved word was not on the study list, then give low ratings on such trials as a way of indicating that those items are "less clear" in memory. The ability to recognize whether an unretrieved word was studied or not is the aforementioned recognition-without-identification effect (e.g., Cleary & Specker, 2007; Cleary, 2006; Cleary & Reyes, 2009; Cleary et al., 2014). Given that recognition-without-identification is a well-established finding, it is plausible that participants might use their sense of whether an unretrieved target was studied to give font clarity or size judgments. However, the old-new discrimination that characterizes the recognition-without-identification effect did not occur in the present study; neither the font clarity nor font size ratings showed the target old-new discriminability that characterizes the recognition-without-identification effect (Figures 1 and 2). In short, participants were not basing their font clarity or size judgments for unretrieved targets on recognition of the study-status of those targets.

Unlike for unretrieved answers, for *retrieved* answers participants used a heuristic whereby if they recognized that the answer was not studied (perhaps upon its retrieval it had no corresponding visual episodic representation), they tended to infer that it was probably less clear (or smaller). Font clarity ratings for retrieved nonstudied targets were significantly lower ($M = 5.17$, $SD = 2.23$) than for retrieved studied targets, whether studied in darker font ($M = 7.02$, $SD = 1.74$), $t(31) = 4.72$, $SE = .39$, Cohen's $d = .85$, $p < .001$, or lighter font ($M = 6.65$, $SD = 2.03$), $t(31) = 3.72$,

$SE = .40$, Cohen's $d = .67$, $p = .001$. Font size ratings were significantly lower for retrieved nonstudied targets ($M = 4.43$, $SD = 1.69$) than for retrieved targets studied in the larger font ($M = 6.60$, $SD = 1.78$), $t(48) = 7.76$, $SE = .28$, Cohen's $d = 1.12$, $p < .001$; however, ratings did not differ between retrieved targets studied in the smaller font ($M = 4.43$, $SD = 1.88$) and nonstudied retrieved targets, $t(48) = 0.003$, $SE = .29$, $p = .997$. Among studied targets that were retrieved, participants' font clarity ratings did not differ significantly for targets studied in darker versus lighter fonts, $t(31) = 1.52$, $SE = .24$, $p = .14$. However, participants did give higher font size ratings to retrieved targets that were studied in larger font than to those studied in smaller font, $t(48) = 7.09$, $SE = .31$, Cohen's $d = 1.02$, $p < .001$.

The means also suggest that target retrieval success versus failure was a piece of information used in deciding on the rating (i.e., an inference that because the target word does not come to mind, it must be less clear in memory). As evidenced above, the mean font clarity and font size ratings for retrieved targets were higher than the midpoint of the scale, whereas the means for unretrieved targets tended to fall on the lower end of the scale. This pattern is typical in research comparing study-status judgments given during retrieval success versus failure (e.g., Cleary, 2006); here, it is occurring with clarity and size judgments.

Experiment 2

Experiment 2 examined whether the TOT heuristic demonstrated in Experiments 1a and 1b would extend to word frequency judgments for unretrieved targets. Word frequency is an indicator of a word's fluency; more frequently occurring words are easier to access than less frequently occurring words (e.g., Hertwig, Herzog, Schooler, & Reimer, 2008; McClelland & Rumelhart, 1981). Though low frequency words may be more likely to elicit TOTs than high frequency words (e.g., Burke et al., 1991), if participants are inclined to infer from a TOT state a heightened state of accessibility for the unretrieved target relative to when no TOT state is present, they may infer from the presence of a TOT state a greater likelihood of the unretrieved target being of higher frequency.

Method

Participants. Forty Colorado State University students participated in exchange for credit toward a course; one was lost for not doing the task.

Materials. Stimuli were the same as in Experiments 1a and 1b with the exception that all study words were presented in 18 point black font. Target words were counterbalanced across participants for study-status assignment.

The target answers were submitted to the English Lexicon Project Web Site (Balota et al., 2007) to determine their word frequency indices. The majority (71/80) of the target answers' HAL frequency indices were available. Of these, the HAL frequency ($M = 730.63$, $SD = 541.82$) range was 7–1966; the log HAL frequency ($M = 6.22$, $SD = 1.06$) range was 1.95–7.58, Thus, our overall pool of target answers was generally low in word frequency, which is not surprising, given that they are from a pool intended to elicit TOT experiences.

Procedure. The procedure was the same as in Experiments 1a and 1b except that the ratings pertaining to each unretrieved target

at test were judgments of the likelihood of a higher versus lower frequency target word (0 = definitely less frequent, 10 = definitely more frequent).

Results and Discussion

Although our pool of targets was generally low in word frequency, frequency still had an impact on target retrievability as well as on participants' judgments of relative frequency for retrieved targets. We performed a median split to divide the target words into relative frequency categories of high ($M = 6.98$, $SD = 0.39$) and low ($M = 5.44$, $SD = 0.97$) HAL log frequency (the median item was placed into the category to which it was closest—the high category). The probability of fully retrieving the answer was higher for high frequency ($M = .44$, $SD = .14$) than for low frequency ($M = .33$, $SD = .15$) targets, $t(38) = 8.21$, $SE = .01$, Cohen's $d = 1.33$, $p < .001$. Participants also discerned relative word frequency among retrieved targets; they gave higher frequency ratings to high ($M = 6.25$, $SD = 2.41$) than to low ($M = 5.97$, $SD = 2.29$) frequency retrieved targets, $t(38) = 2.76$, $SE = .10$, Cohen's $d = .45$, $p < .01$.

This was not the case for unretrieved targets: Participants did not give higher frequency ratings to unretrieved targets from the high frequency category ($M = 3.71$, $SD = 1.71$) than to those from the low frequency category ($M = 3.71$, $SD = 1.63$), $t(38) = 0.02$, $SE = .12$, $p = .99$. In short, participants were unable to detect relative word frequency for targets that they failed to identify. Their frequency ratings also did not demonstrate the target old-new discriminability that characterizes the recognition-without-identification effect, as they did not give higher frequency ratings to studied ($M = 3.52$, $SD = 1.57$) than to nonstudied ($M = 3.70$, $SD = 1.65$) targets, $t(38) = 1.58$, $SE = .12$, $p = .12$.

Turning to our primary question: Evidence of a TOT heuristic was found. As shown in Figure 3, participants judged an unretrieved target as more likely to be of higher frequency when in a TOT state than when not (note that four participants did not report a TOT state in every condition, and thus were lost from this analysis). A 2×2 TOT-state (TOT vs. non-TOT) $\times$ Study-status (target word studied vs. target word nonstudied) repeated measures ANOVA performed on the word frequency ratings given during retrieval failure revealed a main effect of TOT state, $F(1, 34) = 39.19$, $MSE = 3.20$, $p\eta^2 = .54$, $<.001$ (other Fs < 1.0). (The mean number of TOT states reported was 5.09 for the TOT studied condition and 6.66 for TOT nonstudied condition.) This TOT heuristic is interesting given that the probability of reporting a TOT state for the unretrieved target words did not differ between the high ($M = .26$, $SD = .18$) and low frequency ($M = .26$, $SD = .15$) categories, $t(34) = 0.24$, $SE = .02$, $p = .82$. Overall, these findings suggest that participants assumed that the presence of a TOT state indicated a greater likelihood that the unretrieved target word was a higher frequency word, even though this was not so. This is consistent with the idea that participants view a TOT state as indicating a heightened state of accessibility for the unretrieved target relative to when a TOT state for it does not occur.

General Discussion

People often erroneously use currently available information as a heuristic for making unrelated judgments (e.g., Rhodes & Castel, 2008; Schwarz & Clore, 1983; Tversky & Kahneman, 1974; Xuan, Zhang, He, & Chen, 2007). We investigated a new heuristic that we call the TOT heuristic, whereby people use the presence of a current TOT state to make inferences regarding the characteristics of the unretrieved information. The experiments reported here suggest that people infer from the presence of a TOT state characteristics of the unretrieved information that are consistent with higher fluency or accessibility. When in a TOT state, people judged an unretrieved target as more likely to have previously appeared in a darker, clearer font (Experiment 1a) or a larger font (Experiment 1b); they also judged the unretrieved target as more likely to be of higher word frequency (Experiment 2). The full set of results reported here suggest that the association between reported TOT states and these judgments was not the result of more fluent or accessible memory representations underlying TOT states than non-TOT states. Instead, participants appear to attribute the TOT state itself to an increased likelihood that the unretrieved word had more fluent or accessible traits. In short, TOT states themselves seem to confer a sense of perceptibility or fluency on inaccessible words, rather than vice versa.

During retrieval failure, why should a TOT state lead to the sense of a more fluently accessible word in memory relative to a non-TOT state? It is important to consider that the very definition of a TOT state is that the person feels on the verge of accessing a currently unretrievable word (e.g., Schwartz, 2001); this implies a sense of heightened accessibility for unretrieved words that elicit TOT states relative to unretrieved words that do not. From this perspective, it makes sense that people might assume from the presence of a TOT state that the unretrieved word has qualities that are consistent with heightened accessibility relative to when an unretrieved word does not elicit a TOT state.

Though our paradigm used a study/test-list methodology, the TOT heuristic reported here likely extends to real-world situations. Judgments of word frequency, for example, do not require a study list. In fact, the study list had little impact in the present study. Thus, the TOT heuristic would likely extend to a situation in which no study list preceded the general knowledge questions (partici-

Mean ratings of unretrieved targets' relative frequency
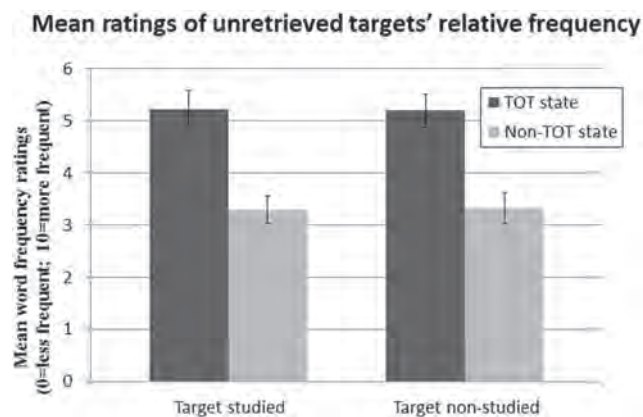
Figure 3. Mean judgments of relative word frequency of unretrieved target words during their retrieval failure (including failure to retrieve partial target information) as a function of reported TOT states. Higher ratings indicate a judged greater likelihood that the unretrieved target word was a higher frequency word. Each error bar represents the standard error of the mean.

pants would likely still show higher judgments of frequency during TOT than non-TOT states). Regarding perceptual clarity judgments, consider a situation in which a witness to a crime is pressed to remember the word printed on the side of a van that was involved. The person cannot remember the word, but is asked to indicate anything else memorable about it: Was the word light or dark? Was it large or small? The TOT heuristic might lead the person to make incorrect inferences about the appearance, size, quality, or other characteristics of the unretrieved information.

The fact that people use TOT states to make inferences about the characteristics of unretrieved information has important implications for the study of (and theoretical understanding of) TOT states. Widely held TOT theories generally assume that TOT states result largely from attributions that people make based on other available information, such as retrieval of some of the unretrieved word's attributes (see Schwartz, 2002; Schwartz & Metcalfe, 2011, for a review). It has not been previously considered in the literature that the presence of a TOT state itself might be used to infer characteristics or qualities regarding the unretrieved information. Our demonstration of a robust TOT heuristic in making inferences about the characteristics of unretrieved target words underscores the need for teasing apart instances of actual access to a word's attributes and instances of mere inference of those attributes from the TOT state itself.

For example, in research claiming that participants have access to partial target information during TOT states, it will be important to demonstrate that the partial access is driving the TOT state and not vice versa (that the TOT state is actually driving the report of partial access via the TOT heuristic). It is possible that there are cases in which TOT states themselves are used to infer certain high frequency or high fluency word attributes that may indeed occur with a high probability in the world and make it *seem* as if partial target word access is occurring when it is not—it is merely an inference being made. For example, if participants are likely to infer high frequency letters or phonemes from the presence of a TOT, this may lead to a higher probability of being correct some of the time than if the guessing was truly random; this could inflate the apparent degree to which participants actually have direct partial access to target attributes during a TOT state.

As a specific example, many more English words start with the letter "t" than start with the letter "j" (Project Gutenberg data). Thus, it is possible that when asked to guess the first letter of an inaccessible word when in a TOT state, participants may be more likely to choose high frequency first letters like "t" (because of the fluency or accessibility) than lower frequency first letters like "j." Because many more words actually do start with "t" than with "j," participants have a higher probability of being right when guessing "t" than when guessing "j." Thus, the mere fact that participants may select the correct first letter more often during TOT states than during non-TOT states is not sufficient evidence of partial access—it may be that participants are using the TOT state to make the inference regarding the first letter, choosing letters that are more fluent or accessible, and thus, more probable. In short, the present findings highlight the importance of disentangling any type of direct partial target access during reported TOT states from the use of a TOT heuristic to infer those attributes.

Future research should examine whether participants use the TOT heuristic to infer other qualities and characteristics of the unretrieved target, such as the number of words in an unretrieved

name (e.g., Hanley & Chapman, 2008), a word's length, or the likelihood that the word starts with a more versus a less frequent first letter or sound. Given the proposed role of partial attribute access in other metacognitive states, such as feelings-of-knowing (e.g., Nomi & Cleary, 2012), future studies should also examine whether similar heuristics are used with such other states. In short, future research should not only further examine the TOT heuristic, but also what other types of metacognitive states might also be used to make similar inferences.

References

Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. A., Kessler, B., Loftis, B., . . . Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39,* 445–459. http://dx.doi.org/10.3758/BF03193014

Brown, A. S. (1991). A review of the tip-of-the-tongue experience. *Psychological Bulletin, 109,* 204–223. http://dx.doi.org/10.1037/0033-2909.109.2.204

Brown, A. S. (2012). *Tip of the tongue states.* New York, NY: Psychology Press.

Burke, D. M., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30,* 542–579. http://dx.doi.org/10.1016/0749-596X(91)90026-G

Cleary, A. M., & Specker, L. E. (2007). Recognition without face identification. *Memory & Cognition, 35,* 1610–1619. http://dx.doi.org/10.3758/BF03193495

Cleary, A. M. (2006). Relating familiarity-based recognition and the tip-of-the-tongue phenomenon: Detecting a word's recency in the absence of access to the word. *Memory & Cognition, 34,* 804–816. http://dx.doi.org/10.3758/BF03193428

Cleary, A. M., Konkel, K. E., Nomi, J. S., & McCabe, D. P. (2010). Odor recognition without identification. *Memory & Cognition, 38,* 452–460. http://dx.doi.org/10.3758/MC.38.4.452

Cleary, A. M., & Reyes, N. L. (2009). Scene recognition without identification. *Acta Psychologica, 131,* 53–62. http://dx.doi.org/10.1016/j.actpsy.2009.02.006

Cleary, A. M., Staley, S. R., & Klein, K. R. (2014). The effect of tip-of-the-tongue states on other cognitive judgments. In B. L. Schwartz & A. S. Brown (Eds.), *Tip-of-the-tongue states and related phenomena* (pp. 75–94). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139547383.005

Hanley, J. R., & Chapman, E. (2008). Partial knowledge in a tip-of-the-tongue state about two- and three-word proper names. *Psychonomic Bulletin & Review, 15,* 156–160. http://dx.doi.org/10.3758/PBR.15.1.156

Hertwig, R., Herzog, S. M., Schooler, L. J., & Reimer, T. (2008). Fluency heuristic: A model of how the mind exploits a by-product of information retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 34,* 1191–1206. http://dx.doi.org/10.1037/a0013025

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. An account of basic findings. *Psychological Review, 88,* 375–407. http://dx.doi.org/10.1037/0033-295X.88.5.375

Mueller, M. L., Dunlosky, J., Tauber, S. K., & Rhodes, M. G. (2014). The font-size effect on judgments of learning: Does it exemplify fluency effects or reflect people's beliefs about memory? *Journal of Memory and Language, 70,* 1–12. http://dx.doi.org/10.1016/j.jml.2013.09.007

Nelson, T. O., & Narens, L. (1980). Norms of 300 general-information questions: Accuracy of recall, latency of recall, and feeling-of-knowing ratings. *Journal of Verbal Learning and Verbal Behavior, 19,* 338–368. http://dx.doi.org/10.1016/S0022-5371(80)90266-2

Nomi, J. S., & Cleary, A. M. (2012). Judgments for inaccessible targets: Comparing recognition without identification and the feeling of know-

ing. *Memory & Cognition, 40,* 1178–1188. http://dx.doi.org/10.3758/s13421-012-0222-4

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137,* 615–625. http://dx.doi.org/10.1037/a0013684

Schwartz, B. L. (2001). The relation of tip-of-the-tongue states and retrieval time. *Memory & Cognition, 29,* 117–126. http://dx.doi.org/10.3758/BF03195746

Schwartz, B. L. (2002). *Tip-of-the-tongue states: Phenomenology, mechanism, and lexical retrieval*. Mahwah, NJ: Erlbaum.

Schwartz, B. L., & Metcalfe, J. (2011). Tip-of-the-tongue (TOT) states: Retrieval, behavior, and experience. *Memory & Cognition, 39,* 737–749. http://dx.doi.org/10.3758/s13421-010-0066-8

Schwarz, N., & Clore, G. L. (1983). Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology, 45,* 513–523. http://dx.doi.org/10.1037/0022-3514.45.3.513

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131. http://dx.doi.org/10.1126/science.185.4157.1124

Whittlesea, B. W. A., Jacoby, L. L., & Girard, K. (1990). Illusions of immediate memory: Evidence for an attributional basis for feelings of familiarity and perceptual quality. *Journal of Memory and Language, 29,* 716–732. http://dx.doi.org/10.1016/0749-596X(90)90045-2

Xuan, B., Zhang, D., He, S., & Chen, X. (2007). Larger stimuli are judged to last longer. *Journal of Vision, 7,* 1–5. http://dx.doi.org/10.1167/7.10.2

Searching for Explanations: How the Internet Inflates Estimates of Internal Knowledge

Matthew Fisher, Mariel K. Goddu, and Frank C. Keil
Yale University

As the Internet has become a nearly ubiquitous resource for acquiring knowledge about the world, questions have arisen about its potential effects on cognition. Here we show that searching the Internet for explanatory knowledge creates an illusion whereby people mistake access to information for their own personal understanding of the information. Evidence from 9 experiments shows that searching for information online leads to an increase in self-assessed knowledge as people mistakenly think they have more knowledge "in the head," even seeing their own brains as more active as depicted by functional MRI (fMRI) images.

Keywords: transactive memory, explanation, knowledge

Just as a walking stick or a baseball glove can supplement the functioning of the body, cognitive tools, computational instruments, and external information sources can supplement the functioning of the mind. The mind can often increase efficiency and power by utilizing outside sources; for tasks like memory, it can rely on cognitive prostheses, such as a diary or a photo album. These external archives can become necessary components of an interdependent memory system (Harris, 1978).

The mind can also become dependent on other minds. When others serve as externalized repositories of information, transactive memory systems can emerge (Wegner, 1987). In these systems, information is distributed across a group such that individuals are responsible for knowing a specified area of expertise. For instance, one person could be responsible for knowing where to find food while another knows how to prepare it. Members of the systems must also track where the rest of the knowledge is stored. Thus, these systems consist of two key elements: internal memory ("What do I know?") and external memory ("Who knows what?") (Hollingshead, 1998; 2001). By reducing redundancy, transactive memory systems work to encode, store, and retrieve information more effectively than could be done by any individual.

Transactive memory systems explain how intimate couples (Wegner, Giuliano, & Hertel, 1985) and familiar groups (Kozlowski & Ilgen, 2006; Peltokorpi, 2008) divide cognitive labor and perform efficiently. These systems can form even with complete strangers, as stereotypes can serve as "defaults" or proxies for another person's expertise (Wegner, Erber, & Raymond, 1991). Better performing memory systems can emerge through communication strategies that allocate domains of knowledge to individuals in the network. Increased group coordination leads to better problem solving than in comparable groups of strangers (Hollingshead, 1998; Littlepage, Robison, & Reddington, 1997). This communication can take place through explicit negotiation (e.g., "you remember the first 3 digits, I will remember the last 4"), but often occurs implicitly. As a relationship develops, members of the system with higher relative self-disclosed expertise will become responsible for knowledge in that domain. Similarly, an individual with access to unique information will become responsible for that information (Wegner, 1987). When groups have not developed these dependencies, decision-making in real-world interactions can be worse than individuals' decisions (Hill, 1982).

Transactive memory may have origins in children's early emerging abilities to navigate the social world and access knowledge in others' minds. External sources of knowledge, especially parents, teachers, and other social partners, play an integral role in children's conceptual development (Gelman, 2009). Information learned from others also exerts a powerful influence over children's notions of what to accept as true—for example, the existence of germs or Santa Claus (Harris, Pasquini, Duke, Asscher, & Pons, 2006). From an early age, children show an emerging but initially limited ability to navigate the terrain of distributed knowledge (Keil, Stein, Webb, Billings, & Rozenblit, 2008). With age they become aware of the breadth, depth, and epistemic limitations inherent to particular kinds of expertise (Danovitch & Keil, 2004). These types of early emerging sensitivities to the content and limitations of other minds may underlie adult transactive memory.

A growing body of theoretical and empirical work suggests that transactive memory systems can be technological as well as social. Though these systems are typically thought to be composed of human minds, our reliance on technology, like the Internet, may form a system bearing many similarities to knowledge dependencies in the social world. The Internet is the largest repository of human knowledge and makes vast amounts of interconnected information easily available to human minds. People quickly be-

come accustomed to outsourcing cognitive tasks to the Internet. They remember where to find information and rely on the Internet to store the actual information (Sparrow, Liu, & Wegner, 2011). This evidence suggests that the Internet can become a part of transactive memory; people rely on information they know they can find online and thus track external memory (who knows the answer), but do not retain internal memory (the actual answer).

The Internet has been described as a "supernormal stimulus" in that its breadth and immediacy far surpass any naturally occurring transactive partner to which our minds might have adapted (Ward, 2013a). Even if the Internet lacks the agency of human transactive memory partners, it shares many of their features and may thus be easily treated as their cognitive equivalent. Compared with a human transactive memory partner, the Internet is more accessible, has more expertise, and can provide access to more information than an entire human transactive memory network. These features leave Internet users with very little responsibility for internal knowledge and may even reduce the extent to which users rely on social others in traditional, interpersonal transactive memory systems.

In a sense, a transactive memory partnership with the Internet is totally one-sided: the Internet stores all the knowledge, and the human is never queried for knowledge. Furthermore, there is no need to negotiate responsibility because the Internet is the expert in all domains. However, to access knowledge in the transactive memory system, the Internet user must navigate the Internet's information in much the same way that one transactive memory partner might know about and query the knowledge contained in another's mind. This interactive aspect of accessing knowledge on the Internet distinguishes it from the way our minds access other information sources. With its unique, supernormal characteristics that allow us to access it much the same way we access human minds, the Internet might be more similar to an ideal memory partner than a mere external storage device. In short, the cognitive systems may well be in place for users to treat the Internet as functionally equivalent to an all-knowing expert in a transactive memory system.

The particular features of the Internet may make it difficult for users to clearly differentiate internally and externally stored information. In most cases of information search, the boundary between information stored "in the head" and information out in the world is quite clear. When we do not know something ourselves, we must take the time and effort to query another source for the answer. If we go to the library to find a fact or call a friend to recall a memory, it is quite clear that the information we seek is not accessible within our own minds. When we go to the Internet in search of an answer, it seems quite clear that we are we consciously seeking outside knowledge. In contrast to other external sources, however, the Internet often provides much more immediate and reliable access to a broad array of expert information. Might the Internet's unique accessibility, speed, and expertise cause us to lose track of our reliance upon it, distorting how we view our own abilities?

One consequence of an inability to monitor one's reliance on the Internet may be that users become miscalibrated regarding their personal knowledge. Self-assessments can be highly inaccurate, often occurring as inflated self-ratings of competence, with most people seeing themselves as above average (Alicke, Klotz, Breitenbecher, Yurak, & Vredenburg, 1995; Dunning, 2005; Pronin,

2009). For example, people overestimate their own ability to offer a quality explanation even in familiar domains (Alter, Oppenheimer, & Zemla, 2010; Fernbach, Rogers, Fox, & Sloman, 2013; Fisher & Keil, 2014; Rozenblit & Keil, 2002). Similar illusions of competence may emerge as individuals become immersed in transactive memory networks. They may overestimate the amount of information contained in their network, producing a "feeling of knowing," even when the content is inaccessible (Hart, 1965; Koriat & Levy-Sadot, 2001). In other words, they may conflate the knowledge for which their partner is responsible with the knowledge that they themselves possess (Wegner, 1987). And in the case of the Internet, an especially immediate and ubiquitous memory partner, there may be especially large knowledge overestimations. As people underestimate how much they are relying on the Internet, success at finding information on the Internet may be conflated with personally mastered information, leading Internet users to erroneously include knowledge stored outside their own heads as their own. That is, when participants access outside knowledge sources, they may become systematically miscalibrated regarding the extent to which they rely on their transactive memory partner. It is not that they misattribute the source of their knowledge, they could know full well where it came from, but rather they may inflate the sense of how much of the sum total of knowledge is stored internally.

We present evidence from nine experiments that searching the Internet leads people to conflate information that can be found online with knowledge "in the head." One's self-assessed ability to answer questions increased after searching for explanations online in a previous, unrelated task (Experiment 1a and b), an effect that held even after controlling for time, content, and features of the search process (Experiments 1c). The effect derives from a true misattribution of the sources of knowledge, not a change in understanding of what counts as internal knowledge (Experiment 2a and b) and is not driven by a "halo effect" or general overconfidence (Experiment 3). We provide evidence that this effect occurs specifically because information online can so easily be accessed through search (Experiment 4a–c).

Experiment 1a

Experiment 1a used a between-subjects design to test whether searching the Internet for explanations leads to higher subsequent ratings for the ability to answer entirely different questions in unrelated domains. In one condition, participants used the Internet to find the answers to common explanatory knowledge questions; then, in the second phase, they evaluated their ability to explain the answers to unrelated sets of questions in various domains of knowledge. In the other condition, participants were asked *not* to use the Internet in the initial induction portion of the study and then, in the second phase, assessed their ability to explain the same unrelated questions seen by the participants who had used the Internet.

Method

Participants. Two hundred two participants (119 men, 83 women, $M_{Age} = 32.59$, $SD = 12.01$) from the United States completed the study through Amazon's Mechanical Turk (Rand, 2012). Based on pilot testing, it was determined that a sample size

of 75–100 participants per condition would be required to detect an effect. Once the requested amount of participants completed the experiment, data collection ended. Five participants were eliminated for failing to follow the instructions to look up answers online; failure to follow instructions was assessed via participants' answers to the following question at the end of the survey: "For how many of the trivia questions at the beginning of this survey did you use the Internet to find the answer? Please answer honestly, this will aid us in our research." Participants in the no Internet condition who chose any number greater than zero were eliminated, and participants in the Internet condition who chose any number fewer than four were eliminated. Participants did not complete multiple experiments; each experiment contains a unique naïve sample. Informed consent was obtained from all participants in all experiments.

Procedure and design. Experiment 1a consisted of two conditions, each with two components: induction and self-assessment. During the induction phase, participants were either instructed to use the Internet to find explanations to common questions (Internet condition) or were instructed not to use the Internet to find the answers to those same questions (no Internet condition). During the second, entirely separate self-assessment phase, participants in both conditions were asked to evaluate how well they could explain the answers to groups of questions in a variety of domains. These questions were entirely unrelated to the questions in the induction phase.

In the induction phase, participants in the Internet condition saw a random subset of four out of six questions about explanatory knowledge such as, "How does a zipper work?" and were asked to search the Internet to "confirm the details of the explanation" (see Appendix A for the full set of questions and Appendix B for the exact instructions). The question contained the phrase "confirm the details" because the explanations were common enough that most people could offer some account without looking up the comprehensive answer. The idea was that participants should have some sense of the answers they were searching for, such that they might more readily and consistently inflate their internal knowledge with the knowledge they were accessing. For this reason, all induction questions were selected from a group of Google autocompleted queries beginning with "Why" and "How"; the questions were selected through piloting with both Internet and no Internet condition instructions to avoid possible ceiling and floor effects. After finding a good explanation for each of the induction questions in this first phase of the experiment, participants in the Internet condition reported the URL of the "most helpful website" and rated their ability to explain the answer to the question on a 1 (*very poorly*) to 7 (*very well*) Likert scale. Participants in the no Internet condition viewed the same random subset of questions and were asked to rate their ability to explain the answers to the questions "without using any outside sources." The purpose of asking participants to rate how well they could explain the answers to the induction questions in this first phase was to track likely differences between confidence in no Internet users and Internet users, who might feel more sure of the answers after looking them up.

During the second phase, the self-assessment phase, all participants rated their ability to answer questions about knowledge in six domains unrelated to the questions posed in the induction phase: weather, science, American history, food, health treatments, and the human body. For each set, participants considered four

questions, for example, "Why are there more Atlantic hurricanes in August and September?", "How do tornadoes form?", "Why are cloudy nights warmer?" (see Appendix C for complete list of questions for each domain set). Participants were asked, "How well could you answer detailed questions about [topic] similar to these?" on a 1 (*very poorly*) to 7 (*very well*) Likert scale.

Results

Participants who had looked up explanations on the Internet in the induction phase rated themselves as being able to give significantly better explanations to the questions in the unrelated domains during the self-assessment phase ($M = 3.61$, $SD = 1.27$, 95% confidence interval [CI] = [3.40, 3.91]) than those who had not used the Internet ($M = 3.07$, $SD = 1.06$, 95% CI = [2.88, 3.27]), $t(195) = 3.24$, $p = .001$, Cohen's $d = 0.50$ (Figure 1). The effect was observable across all six domains for which participants were asked to assess their knowledge.

However, participants in the Internet condition spent longer in the induction phase ($M = 214.40$ s, $SD = 129.93$) than participants in the no Internet condition ($M = 26.26$ s, $SD = 26.26$), $t(195) = 14.65$, $p < .001$. Extended reflection on the initial questions may have increased explanatory confidence, accounting for the difference between conditions. Furthermore, participants in the Internet condition rated themselves as having a better ability to explain the items in the induction phase ($M = 5.00$, $SD = 1.42$) than those in the no Internet condition ($M = 3.34$, $SD = 1.19$), $t(195) = 8.90$, $p < .001$.

In addition, the results of this experiment failed to address the possibility that Internet use was not inflating Internet condition participants' confidence in their knowledge, but rather that the No Internet participants' self-assessed knowledge ratings were *deflated* from baseline by lack of Internet use. Experiment 1b was designed to test whether the Internet participants' self-assessed knowledge ratings were in fact rising from a baseline.

Experiment 1b

Experiment 1b used a nearly identical design to Experiment 1a; the key difference was that this experiment added a self-assessment phase *prior* to the induction phase for both the Internet and the no Internet conditions. This additional knowledge self-assessment phase was identical to the second phase of Experiment 1a: It asked participants to evaluate their knowledge about different domains with representative questions through the question, "How well could you answer detailed questions about [topic] similar to these?" Participants provided ratings on a 1 (*very poorly*) to 7 (*very well*) Likert scale. Because this preinduction self-assessment phase was also identical to the third phase of this new Experiment 1b, its addition was intended to allow for direct comparison between pre- and postinduction self-assessed knowledge ratings, testing whether the observed effect (the difference in postinduction self-assessed knowledge ratings between the Internet and no Internet conditions) occurred because Internet use inflated users' confidence from baseline or because a lack of Internet use *deflated* confidence from baseline in the no Internet condition. Because Internet use is quite widespread in the United States, with more than 71.1% of households reporting accessing the Internet in 2011, we wanted to determine the directionality of

the effect (File, 2013). This was perhaps especially important given that our participants were Amazon Mechanical Turk workers who are presumably heavier Internet users than average.

Method

Participants. One hundred fifty-two participants (76 men, 76 women, M_{Age} = 32.14, SD = 10.24) from the United States completed the study through Amazon's Mechanical Turk. Ten participants were eliminated for not following instructions in the induction phase to look up answers online, as judged by their responses to the "Internet check" question at the end ("For how many of the trivia questions at the beginning of this survey did you alter the search provided in the link to find the answer? Please answer honestly, this will aid us in our research").

Procedure and design. Experiment 1b used the same design as Experiment 1a but also included an additional, preinduction self-assessment. During the preinduction self-assessment phase, participants in both the Internet and the no Internet conditions were asked to evaluate how well they could explain the answers to questions about specific subject matter (these questions were identical to the self-assessment questions in Experiment 1a and may be viewed in full in Appendix C). Next, during the induction phase, and exactly as in Experiment 1a, participants either used the Internet to find explanations to common questions or were shown the same questions with instructions not to use the Internet to find the answers. Finally, in the postinduction self-assessment phase, participants responded to the same set of questions from the preinduction self-assessment phase.

Results

There was no difference in the preinduction self-assessment baseline ratings between the Internet (M = 3.21, SD = 1.16, 95% CI = [3.08, 3.34]) and the no Internet condition (M = 3.21, SD = 1.33, 95% CI = [3.04, 3.37]), $t(140)$ = −0.04, p = .99. Replicating the results from Experiment 1a, participants who looked for explanations on the Internet during the induction phase rated themselves as being able to be give significantly better explanations in the unrelated domains during the postinduction self-assessment phase (M = 3.63, SD = 1.52, 95% CI = [3.44, 3.81]) than participants who had not used the Internet (M = 3.15, SD = 1.21, 95% CI = [2.96, 3.29]), $t(140)$ = −2.1, p < .05, Cohen's d = 0.35 (Figure 1). The effect was observable across all six domains for which participants were asked to assess their knowledge, consistent with the account that the baseline of self-assessed knowledge is systematically inflated because of Internet use.

Just as in Experiment 1a, however, participants in the Internet condition spent longer in the induction phase (M = 196.24 seconds, SD = 166.31) than participants in the no Internet condition (M = 28.75 seconds, SD = 21.67), $t(140)$ = 7.17, p < .001. And just as in Experiment 1a, participants in the Internet condition rated themselves as having a better ability to explain the items in the induction phase (M = 4.18, SD = 1.58) than those in the no Internet condition (M = 3.56, SD = 1.34), $t(140)$ = 2.50, p < .05. Experiment 1c addressed these confounds by equating both time spent in the induction task and amount of learning (i.e., self-assessed ability to explain induction questions) during the induction task across conditions.

Experiment 1c

The results of Experiment 1a and b provided initial evidence for an effect whereby searching the Internet for explanations results in an increase from baseline in self-assessed knowledge for unrelated domains. However, in those experiments, both time spent in the induction phase and information learned during the induction phase may have accounted for the difference between the Internet and no Internet conditions. Experiment 1c sought to rule out these alternative explanations.

Experiment 1c was designed to match both the amount of time spent and the content (i.e., the explanations viewed in the induction phase) across the Internet and no Internet conditions. The design of Experiment 1c also addressed a third possibility: that features of autonomous searching behaviors (including source scrutiny, a sense of self-directed learning, etc.) could explain the difference in self-assessed knowledge ratings between the conditions.

Method

Participants. Two hundred four participants (120 men, 84 women, M_{Age} = 32.85, SD = 10.29) from the United States completed the study through Amazon's Mechanical Turk. Nine participants were eliminated for failing to look up the answers to the questions in the Internet condition, as assessed by the Internet check question at the end of the survey.

Procedure and design. The design for Experiment 1c was identical to that of Experiment 1a, with three changes. First, participants in the Internet condition were provided with specific instructions for how to find each explanation to the induction questions, thereby constraining their searching to prespecified sources. For example, participants in the Internet condition were asked, "Why are there dimples on a golf ball?" and instructed, "Please search for the scientificamerican.com page with this information." We specifically selected these web sources because they contained primarily textual content and included little or no graphics.

Participants in the no Internet condition were provided with the *exact text* from the same websites for which Internet participants were instructed to search. (In contrast, in Experiments 1a and b, the no Internet participants received no information at all about the questions in the induction phase.) Because the websites from which the explanations were drawn were specially selected for their heavy textual content, this ensured that participants across conditions viewed the same explanatory content, controlling for the amount of new information participants accessed during the induction phase across conditions.

Last, because the results of a separate pilot study showed that participants took an average of 12.6 s to find the webpage listed in the instructions, Experiment 1c also introduced a 12.6-s delay before explanations were displayed for participants in the no Internet condition, thereby controlling for time spent in the induction phase.

Participants in the Internet condition were excluded either if they failed to provide the URL of the intended page or if they provided any URL different from the intended page for any question in the induction phase.

Results

Participants in the Internet condition spent the same number of time on each explanation ($M = 68.00$ s) as participants in the no Internet condition ($M = 73.36$s), $t(193) = -1.01, p = .31$.

Once again, participants in the Internet condition provided higher self-assessments of knowledge postinduction ($M = 3.78$, $SD = 1.19$, 95% CI = [3.54, 4.03]) than participants in the no Internet condition ($M = 3.07, SD = 1.12$, 95% CI = [2.85, 3.28]), $t(193) = 4.30, p < .001$, Cohen's $d = 0.63$ (Figure 1).

Though the explanatory content viewed in the induction phase was exactly matched across conditions, when asked how well they could explain the questions in the induction phase, participants in the Internet condition gave higher ratings ($M = 5.07$ $SD = 0.99$) than participants in the no Internet condition ($M = 4.33$, $SD = 1.33$), $t(193) = -4.33, p < .001$. A linear regression model, controlling for self-rated ability to explain the questions in the induction phase, B = .039, β = .41, $p < .001$, showed experimental condition to still be a significant predictor of knowledge self-assessments, B = .042, β = .17, $p < .01$. Even though participants in the Internet condition rated their ability to explain the induction questions higher than those in the no Internet condition, this difference did not explain the difference in knowledge ratings in the self-assessment phase.

Notably, in addition to equating for time spent and exact content accessed during the induction phase, the results of Experiment 1c suggest that features of autonomous searching such as evaluating, comparing, or choosing between multiple sources of information cannot explain the effect because participants in the Internet condition were told exactly where to go to retrieve their explanatory information.

Discussion

Across three studies, Experiment 1 demonstrated that searching for explanations online increases self-assessed knowledge in unrelated domains. The effect is observed even when time spent in the induction phase is the same for participants in both Internet and no Internet conditions, when the content viewed across conditions is identical, and when Internet condition participants' autonomous search behavior is restricted through the assignment of a particular web source.

Experiment 2a

Participants in the Internet and no Internet conditions in Experiments 1a–c could have interpreted the dependent measure differently. That is, the phrasing of the dependent measure ("How well could you explain the answers to questions similar to these about [topic]?") may have left the meaning of "you" open to interpretation. If participants in the Internet condition were considering both the knowledge in their heads *and* online information when assessing their ability to answer questions in various domains, then it would be entirely unsurprising that they deemed themselves more knowledgeable. Experiments 2a and b were designed to resolve this ambiguity inherent to the phrasing of the dependent measure in Experiments 1a–c, thus allowing for more accurate interpretations of findings from those experiments.

Method

Participants. Two hundred three participants (99 men, 104 women, $M_{Age} = 33.24, SD = 11.35$) from the United States completed the study through Amazon's Mechanical Turk. Eleven participants were eliminated for either not looking up the answers to the questions in the Internet condition or using the Internet to look up answers to the questions in the no Internet condition as determined by answers to the end-of-survey Internet check question.

Procedure and design. In Experiment 2a, a new dependent measure replaced those used in the self-assessment phase of Experiment 1. Instead of asking participants to rate how well they could answer questions about topics using a Likert scale ranging from 1 (*very poorly*) to 7 (*very well*), participants were shown a scale consisting of seven functional MRI (fMRI) images of varying levels of activation, as illustrated by colored regions of increasing size (Figure 2). Participants were told, "Scientists have shown that increased activity in certain brain regions corresponds with higher quality explanations." This dependent variable was designed to unambiguously emphasize one's brain as the location of personally held knowledge. Participants were then asked to select the image that would correspond with their brain activity when they answered the self-assessed knowledge questions in each of the six domains. To further ensure that participants accurately interpreted this new dependent measure as pertaining to their own, independently held knowledge, at the end of the experiment participants explained all of the factors they considered when making judgments about their brain activity via free response. The procedure was otherwise identical to Experiment 1a.

Results

Replicating the effect found in Experiment 1, participants in the Internet condition chose the images with more brain activity ($M = 4.66, SD = .99$, 95% CI = [4.40, 4.83]) than those in the no Internet condition ($M = 4.12, SD = 1.13$, 95% CI = [3.94, 4.39]), $t(190) = 3.52, p = .001$, Cohen's $d = 0.43$ (Figure 1).

Two independent raters coded free responses about the factors participants considered when making judgments about their brain activity and found that Internet participants were not considering knowledge online when making their ratings in the self-assessment phase. When asked what they did consider, 41% spontaneously mentioned their current knowledge, 37% the complexity of their explanation, 30% the complexity of the question, 13% the amount of thinking required, 17% other explanations, and only 2% cited access to other sources. Interrater reliability was high (κ = .83, $p < .001$) with disagreements resolved through discussion. Removing participants who considered accessing outside sources as a relevant factor had no effect on the significance of the results.

Experiment 2b

The findings of Experiment 2a suggested that participants in both conditions had interpreted the dependent measure in the self-assessment phase as intended—that is, as pertaining solely to the knowledge they held in their heads, rather than to their own knowledge plus knowledge accessible from outside sources. Experiment 2b addressed the possibility of a misinterpretation of the
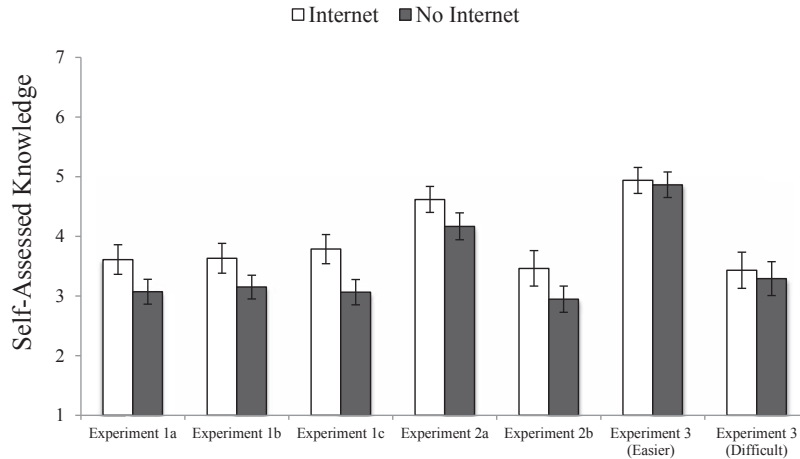
□ Internet ■ No Internet

Figure 1. Differences in self-assessed knowledge between the Internet and no Internet conditions for Experiments 1–3. Error bars indicate mean ± 95% confidence interval (CI).

dependent measures even more directly with instructions clarifying that the ratings in the self-assessment phase should reflect the participant's current knowledge "without any outside sources."

Method

Participants. One hundred ninety-nine participants (127 men, 72 women, $M_{Age} = 31.42$, $SD = 10.92$) from the United States completed the study through Amazon's Mechanical Turk. Twelve participants were eliminated for either not looking up the answers to the questions in the Internet condition or using the Internet to look up answers to the questions in the no Internet condition as determined by the end-of-survey Internet check question.

Procedure and design. The procedure and design of Experiment 2b were identical to those of Experiment 1a with one difference: Each question from the self-assessment phase of the experiment asked participants explicitly how well they could answer questions about this topic "without any outside sources." In previous experiments, this phrase had appeared only in the instructions for the induction phase of the no Internet condition, not in the self-assessment phase questions. The addition of this phrase was intended to explicitly restrict participants' judgments about the boundaries of knowledge to include only their own internal knowledge.

Results

Participants in the Internet condition rated their ability to answer questions without using outside sources higher ($M = 3.41$, $SD =$

1.47, 95% CI = [3.10, 3.72]) than participants in the no Internet condition ($M = 2.94$, $SD = 1.16$, 95% CI = [2.72, 3.16]), $t(193) =$ 4.30, $p < .001$, Cohen's $d = 0.36$ (Figure 1).

Discussion

The findings from Experiments 2a and b provide direct evidence that participants interpreted the self-assessed knowledge questions similarly across the Internet and no Internet conditions. In other words, the results of Experiment 2 suggest that participants in the Internet condition of these experiments and Experiment 1 did not consider knowledge available online when rating their knowledge in new domains during the self-assessment phase of the experiment.

Experiment 3

A possible explanation for the observed effect is that using the Internet to access explanations simply increases confidence in one's knowledge or abilities more generally. Experiment 3 was designed to explore the possibility of such a "halo effect" by replacing the topics in the self-assessment phase with explanatory knowledge topics that are similar yet cannot be accessed using the Internet. Detailed autobiographical knowledge is one of the few forms of knowledge that cannot be found online yet closely mirrors the kind of explanatory questions used in the previous experiments. If a difference between the self-assessed knowledge ratings for the Internet and no Internet conditions still exists for these questions, it would suggest that general overconfidence from In-
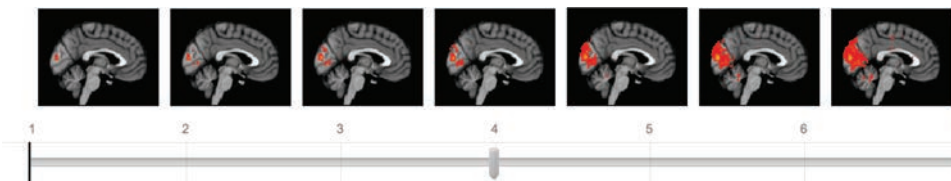

Figure 2. Measure of self-reported brain activity in Experiment 2a. See the online article for the color version of this figure.

ternet use could account for the results of Experiment 1–2. However, no difference between the conditions would be evidence for a boundary condition of the phenomenon, indicating the previous results are not explained by a "halo effect."

Method

Participants. Three hundred two participants (194 men, 108 women, M_{Age} = 31.32, SD = 9.86) from the United States completed the study through Amazon's Mechanical Turk. Twenty-two participants were eliminated for either not looking up the answers to the questions in the Internet condition or using the Internet to look up answers to the questions in the no Internet condition as determined by responses to the end-of-survey Internet check question.

Procedure and design. In Experiment 3, we changed the type of topics presented during the self-assessment phase of the experiment. Instead of asking about explanatory questions that can be answered using the Internet, we asked participants about autobiographical explanatory knowledge for which the Internet would be of no help. The six knowledge topics were personal history, personal future, relationships, local culture, personal habits, and emotions. For example, questions about relationships were, "Why are you so close with your best friend?"; "How are you similar to your mother?"; and, "How could you become friendlier with your next door neighbor?" (see Appendix D for the full list of autobiographical questions).

In pilot testing, participants viewed these autobiographical knowledge self-assessment questions as significantly easier than the explanatory self-assessment topics from Experiments 1–2 (presumably because they were much more familiar topics); so, we included a second set of questions which were rated to be equally as difficult as the explanatory knowledge questions from Experiments 1–2. The difficult autobiographical questions were grouped into the same categories as the easier autobiographical questions and were chosen from the results of pilot tests in which self-assessed knowledge ratings were similar to the self-assessment questions used in Experiments 1–2 (see Appendix E for the full list of the difficult autobiographical questions).

Experiment 3 thus used a 2 (Internet vs. no Internet) × 2 (Easier vs. Difficult) between-subjects design. Just as in previous experiments, after the induction phase participants viewed six unrelated knowledge topics (with three representative sample questions each) and were then asked, "How well could you explain the answer to questions about [*topic*] similar to these?" They provided their responses on a 1 (*very poorly*) to 7 (*very well*) Likert scale.

Results

A one-way analysis of variance (ANOVA) showed that self-assessed knowledge ratings for autobiographical questions were the same after accessing the Internet (M = 4.24, SD = 1.33) compared with not accessing the Internet (M = 4.04, SD = 1.38), $F(1, 276)$ = 1.37, p = .30 (Figure 1). As expected, participants gave higher ratings for the easier autobiographical questions (M = 4.91, SD = 0.96) compared with the difficult autobiographical questions (M = 3.35, SD = 1.25), $F(1, 276)$ = 134.21, p < .001. There was no interaction, indicating that at both levels of difficulty, using the Internet did not boost self-assessed knowledge for questions that could not be found online.

Discussion

Experiment 3 suggests that accessing the Internet does not lead to a general overconfidence, but rather to a more specific illusion of knowledge that occurs only in domains where the Internet would be of use. If induction time and induction content (Experiment 1), imprecise interpretations of personally held knowledge (Experiment 2), and general overconfidence effects (Experiment 3) cannot explain the observed inflation of self-assessed knowledge, what else might? Experiment 4 explores whether the process of querying through Internet search might be the underlying mechanism explaining the effect.

Experiment 4a

The findings from Experiments 1–3 raise important questions about the locus of this effect. Experiment 4a provides evidence that actively posing queries through Internet search engines is the specific mechanism by which Internet usage causes an increase in self-assessed knowledge. Experiment 4a was designed to investigate (a) whether the effect persists when the act of searching is removed from accessing explanations on the Internet and (b) whether the effect persists even when less popular search engines are used for searching. If there is no effect when the act of searching is removed from accessing explanations yet remains even when Internet users couple with unfamiliar "partners," together these results would be strong evidence that active searching is the element of Internet access that drives the observed effect in which participants inflate their self-assessed knowledge after Internet use.

Method

Participants. One hundred fifty-seven participants (77 men, 80 women; M_{Age} = 31.94, SD = 11.34) from the United States completed the study through Amazon's Mechanical Turk. Nine participants were eliminated for failing to follow instructions by using a search engine they were not instructed to use; this was assessed using an end-of-survey Internet check question that explicitly asked participants whether they had used a search engine other than the one they had been assigned ("For how many of the trivia questions at the beginning of this survey did you alter the search provided in the link to find the answer? Please answer honestly, this will aid us in our research").

Procedure and design. Experiment 4a contained two conditions, each a variation on the design of Experiment 1c. The first condition, the no search condition, was designed to test whether there is an increase in knowledge self-assessments when the search component is entirely removed from the induction task for participants in the Internet condition. In this condition, participants saw an experimental setup that was identical to that of the Internet condition of Experiment 1c; however, instead of searching for the answers to induction questions online themselves, participants in this condition were provided with a link that took them directly to the website with the explanation.

Second, the other search engines condition was designed to test whether the effect might only occur using a search engine with which one has successfully queried for knowledge in the past. If an association between a particular search engine and successfully

accessing knowledge drives the increase in self-assessed knowledge in the Internet conditions of the previous experiments, then using less popular search engines should yield a weaker effect (or perhaps none at all). However, if actively querying an information-rich source via search engine drives the effect, then participants should increase their knowledge self-assessments after using *any* search engine. The other search engine condition was identical to the Internet condition of Experiment 1c, except that participants were instructed to look up the answers to the induction questions using one of the following 5 search engines (varying in popularity): duckduckgo.com, blekko.com, ixquick.com, search.yahoo.com, ask.com.

Results

Participants who used links to access information instead of searching provided lower self-knowledge ratings ($M = 3.20$, $SD = 0.99$, 95% CI = [3.09, 3.31]) than participants who used other search engines ($M = 3.63$, $SD = 1.27$, 95% CI = [3.49, 3.77]), $t(146) = -2.28$, $p = .03$, Cohen's $d = 0.37$. A one-way ANOVA showed no difference in self-assessed knowledge across the different website that participants accessed, $F(4, 77) = .63$, $p = .64$.

Experiment 4b

The findings of Experiment 4a suggest that active search is necessary in order for Internet usage to result in inflated estimates of self-assessed knowledge. To further explore the account that searching is the specific mechanism by which the effect occurs, Experiment 4b investigated whether the effect still holds when searching the Internet yields unhelpful information. Does self-assessed knowledge increase even if the search engine does not provide a satisfactory answer?

Method

Participants. One hundred fifty-one participants (106 men, 45 women, $M_{Age} = 29.79$, $SD = 7.94$) from the United States completed the study through Amazon's Mechanical Turk. Six participants were eliminated for not following instructions by removing the Google search filters; this was assessed using an end-of-survey Internet check question that explicitly asked participants whether they had removed the Google search filters ("For how many of the trivia questions at the beginning of this survey did you alter the search provided in the link to find the answer? Please answer honestly, this will aid us in our research").

Procedure and design. The design of Experiment 4b's two conditions was identical to that used in the Internet condition induction in Experiment 1c except for one difference: in place of the original explanatory induction questions, each of the two conditions of Experiment 4b used a new set of induction questions. These two new sets of induction questions were matched in structure and content to each other; the only difference between them was that one set contained questions that a top Google search result could answer comprehensively (the answer condition), while the other set consisted of questions with answers that could not be found using Google (no answer condition). For example, participants in the Answer condition would be asked "Why is ancient Egyptian history more peaceful than Mesopotamian history?",

which returns an article that clearly answers the question, and participants in the no answer condition would be asked "Why is ancient Kushite history more peaceful than Greek history?", which is parallel in content and structure yet does not have an answer easily found online (see Appendix F for the full set of questions). Just as in the induction phase of Experiment 1c, participants saw a random subset of 4 induction questions and were instructed to "search the Internet to confirm the details" of their explanations to these questions. After the induction phase, they completed the same general knowledge self-assessment questions used in previous experiments.

Results

Knowledge ratings did not differ between participants in the answer condition ($M = 4.00$, $SD = 1.19$, 95% CI = [3.74, 4.26]) and those in the no answer condition ($M = 4.11$, $SD = 1.22$, 95% CI = [3.81, 4.41]), $t(143) = -0.55$, $p = .58$. To draw meaningful comparisons between these results and those of previous experiments, we combined the results of relevant earlier experiments. Experiments 1a–c and 2b provided four successful demonstrations of the effect (though Experiment 2a also successfully replicated the effect, the change in the dependent measure to fMRI pictures shifted ratings higher compared with the other experiments). Using the data from the no Internet condition of these four previous studies for comparison with the results of Experiment 4b, we can determine whether search activity, even if unsuccessful, leads to increased ratings of knowledge.

Pooling across the no Internet conditions from the previous studies, we found that participants in the answer condition of Experiment 4b increased their knowledge ratings compared with the aggregate no Internet baseline formed by combining Experiments 1a–c and Experiment 2b ($M = 3.05$, $SD = 1.13$, 95% CI = [2.94, 3.16]), $t(476) = -6.81$, $p < .001$. The results of this comparison hold if the answer condition ratings are compared individually to the no Internet condition from Study 1a, $t(153) = -4.43$, $p < .001$, Study 1b, $t(188) = -5.66$, $p < .001$, Study 1c, $t(184) = -5.46$, $p < .001$, and Study 2b, $t(185) = -6.15$, $p < .001$. Surprisingly, even despite unsuccessful search efforts, participants in the no answer condition *also* increased their self-assessed knowledge compared with the aggregated no Internet ratings, $t(461) = -6.94$, $p < .001$. This result also holds when the no answer condition is compared with each of the previous no Internet conditions individually, Study 1a, $t(138) = -4.69$, $p < .001$, Study 1b, $t(173) = -5.91$, $p < .001$, Study 1c, $t(169) = -5.70$, $p < .001$, and Study 2b, $t(170) = -6.32$, $p < .001$. This is strong support for searching as the mechanism that gives rise to illusions of knowledge from Internet use.

Experiment 4c

The findings from Experiment 4b provide initial evidence that search activity, even when unsuccessful because of hard-to-find relevant results, drives the observed effect. Experiment 4c further explores the extent to which the retrieval of search results causes Internet users to inflate their self-assessed knowledge in that it uses an even stronger test. Experiment 4c asks whether the illusion persists even when searching returns only irrelevant results or no results at all. If search *success* is causing participants to inflate their self-assessed knowledge, then participants who access irrelevant results or zero

results will have lower ratings of self-assessed knowledge than participants in the analogous Experiments 1a–c and 2b who successfully access relevant search results. However, if search activity alone, regardless of search success, is driving the inflation of self-assessed knowledge, then participants who search unsuccessfully will rate themselves higher than participants in the no Internet conditions from previous experiments.

Method

Participants.   One hundred thirty-eight participants (men = 86, women = 52; M_{Age} = 31.26, SD = 10.39) from the United States completed the study through Amazon's Mechanical Turk. Seven participants were eliminated for removing the filters placed on the search; this was assessed using an end-of-survey Internet check question that explicitly asked participants whether they had removed the search filters they had been assigned.

Procedure and design.   Experiment 4c consisted of two conditions that were each variations on the design of the Internet condition of Experiment 1a. These conditions were together designed to investigate whether impeding the effectiveness of participants' search activity by filtering search results affected participants' subsequent self-assessed knowledge ratings.

In the filtered results condition, participants were instructed to search for the explanations to the induction questions using a filtered Google search that provided only the most recently posted results (i.e., within the past week). These recent results from the filtered Google search did not provide direct answers to the induction questions. In the no results condition, participants were instructed to search for the answers to the induction questions using a Google filter that blocked *all* results, with the Google results page displaying a message to participants that their search "did not match any documents."

Results

Participants in the filtered results did not differ in their self-assessed knowledge (M = 3.57, SD = 1.27, 95% CI = [3.27, 3.87]) compared with those in the no results condition (M = 3.75, SD = 1.17, 95% CI = [3.46, 4.04]), $t(129) = -.82, p = .42$. Again pooling together ratings from participants in the no Internet conditions of Experiment 1a–c and 2b to form a baseline for comparison, we found that participants in the filtered results condition rated their knowledge higher (M = 3.57, SD = 1.27) than participants who had not searched online for answers to the induction questions (M = 3.05, SD = 1.13, 95% CI = [2.94, 3.16]), $t(465) = -3.49, p = .001$. The results holds if the filtered results ratings are compared individually to the no Internet condition from Study 1a, $t(142) = -2.06, p < .05$, Study 1b, $t(177) = -2.84, p < .01$, Study 1c, $t(173) = -2.74, p < .01$, and Study 2b, $t(174) = -3.43, p = .001$.

Strikingly, the ratings from the no results condition, in which participants' searching activities returned *zero* search results at all, were also higher (M = 3.75, SD = 1.17) than the aggregate no Internet condition (M = 3.05, SD = 1.13, 95% CI = [2.94, 3.16]), $t(458) = -4.50, p < .001$. When compared individually, the no results ratings were also higher than the no Internet condition from Study 1a, $t(135) = -2.93, p < .01$, Study 1b, $t(170) = -3.86, p < .001$, Study 1c, $t(166) = -3.71, p < .001$, and Study 2b, $t(167) = -4.37, p < .001$.

Discussion

The illusion of knowledge from Internet use appears to be driven by the act of searching. The effect does not depend on previous success on a specific search engine, but rather generalizes to less popular search engines as well (Experiment 4a). It persists when the queries posed to the search engine are not answered (Experiment 4b) and remains even in cases where the search query fails to provide relevant answers or even any results at all (Experiment 4c). Even when stripped of such potentially integral features, Internet searching still results in increases in self-assessed knowledge. This suggests that the illusion is driven by the act of searching itself.

General Discussion

Searching for answers online leads to an illusion such that externally accessible information is conflated with knowledge "in the head" (Experiment 1a and b). This holds true even when controlling for time, content, and search autonomy during the task (Experiment 1c). Furthermore, participants who used the Internet to access explanations expected to have increased brain activity, corresponding to higher quality explanations, while answering unrelated questions (Experiment 2a). This effect is not driven by a misinterpretation of the dependent measure (Experiment 2b) or general overconfidence (Experiment 3) and is driven by querying Internet search engines (Experiment 4a-c).[1]

In many ways, our minds treat the Internet as a transactive memory partner, broadening the scope of knowledge to which we have access. The results of these experiments suggest that searching the Internet may cause a systematic failure to recognize the extent to which we rely on outsourced knowledge. Searching for explanations on the Internet inflates self-assessed knowledge in unrelated domains. Our results provide further evidence for the growing body of research suggesting that the Internet may function as a transactive memory partner (Sparrow, Liu, & Wegner, 2011).

People tend to inaccurately recall the original source of their internal memories (Johnson, 1997; Johnson, Hashtroudi, & Lindsay, 1993). In this regard, our findings might be initially unsurprising: When searching online, people misattribute the source of the specific answers they find because they think the answer was stored in their own mind instead of on the Internet. However, in the current set of studies, we first asked people one set of questions in the induction phase and then asked them an entirely separate set of questions in different domains of knowledge. This design rules out the explanation that participants merely failed to monitor the fact that the Internet was the source of their knowledge. Rather, our results suggest that what participants failed to accurately monitor was the proportion of internal and external memory comprising the sum total of accessible knowledge, mistaking outsourced knowledge for internal knowledge. People neglect the extent to which they would rely on their partner in the transactive memory system to access explanatory knowledge.

This illusion of knowledge might well be found for sources other than the Internet: for example, an expert librarian may experience a similar illusion when accessing a reference Rolodex.

_____
[1] See Table 1 for summary of experimental findings.

Table 1
Summary of Experimental Results

Experiment	Method	Results (self-assessed knowledge ratings)	Conclusions
1a	Internet condition uses Internet to look up explanations to common questions; no Internet condition does not.	Internet > no Internet	Internet condition gives higher self-knowledge ratings than no Internet condition
1b	Same as Experiment 1a, but all participants make self-assessed knowledge ratings both before and after induction phase.	Preinduction, no difference between Internet and no Internet	Searching the Internet increases self-assessed knowledge from baseline
1c	Internet condition searches constrained to specific sources; no Internet condition sees identical explanations	Controlling for induction phase ratings, Internet > no Internet	Time, content, and autonomous search activity do not explain the effect
2a	Same as 1a, but DV for self-assessed knowledge questions = fMRI "brain activity"	Internet > no Internet	Participants are not misinterpreting the dependent measure
2b	Same as 1a, but DV for self-assessed knowledge questions = "on your own, with no outside sources"	Internet > no Internet	Participants are not misinterpreting the dependent measure
3	Same as 1a, but questions for self-assessed knowledge phase are autobiographical explanatory questions	No difference between Internet and No Internet conditions	Effect not explained by general overconfidence
4a	Link condition clicked on a link to explanation instead of searching; Other search engines condition used 5 alternative engines for searching	Other search engines > Link	Effect driven by active search independent of search engine
4b	Answer condition searched for induction questions easily found on Internet; no Answer condition searched for matched-content questions not answered in any search result	Both answer and no answer conditions > no Internet baseline	Effect holds even without direct answers to search query
4c	Recent results condition searched for induction explanations in Google search returning irrelevant recent results only; zero results condition returned zero search results.	Both recent results and no results conditions > no Internet baseline	Effect holds even without any results for search query

An individual in a highly integrated social environment (Hutchins, 1995) may conflate knowledge "in the head" with knowledge stored in other human sources, such as fellow members of a cockpit crew. While such effects may be possible, the rise of the Internet has surely broadened the scope of this effect. Before the Internet, there was no similarly massive, external knowledge database. People relied on less immediate and accessible inanimate stores of external knowledge, such as books— or, they relied on other minds in transactive memory systems. In contrast with other sources and cognitive tools for informational access, the Internet is nearly always accessible, can be searched efficiently, and provides immediate feedback. For these reasons, the Internet might become even more easily integrated with the human mind than other external sources of knowledge and perhaps even more so than human transactive memory partners, promoting much stronger illusions of knowledge.

Recent evidence suggests similar illusions occur when users search for fact-based information online (Ward, 2013b). After using Google to retrieve answers to questions, people seem to believe they came up with these answers on their own; they show an increase in "cognitive self-esteem," a measure of confidence in one's own ability to think about and remember information, and predict higher performance on a subsequent trivia quiz to be taken without access to the Internet. These fact-based effects are dependent on the reliability and the familiarity of the search engine, suggesting the processes by which the Internet affects cognition function differently across types of knowledge. These differences across informational contexts highlight the need for further research on the effects of different forms of online information search.

Confusion of accessible knowledge with one's personal knowledge may not be a gradual result of cultural immersion. Instead, it may be an early emerging tendency that remains even as children learn to access to the division of cognitive labor in the world around them. Tasks like learning the meanings of new words may be facilitated by a tendency for children to believe that they "knew it all along" (Kominsky & Keil, 2014; Mills & Keil, 2004; Taylor, Esbensen, & Bennett, 1994). Such misattributions may endow children with an adaptive confidence that their understandings are well grounded. The Internet may exaggerate this bias even in adults, leading to failures in recognizing the limits of internal explanatory knowledge.

The participants in our studies completed the experiments online and presumably use Internet search engines frequently. Why might we still observe an effect if the participants are already closely connected with the Internet as a transactive memory partner? It may be that chronic (overall frequency of use) and recent (experimental induction) search both influence knowledge self-assessments. In the area of social priming, similar effects have been found. Chronic and recent influences combine additively, so experimental exposures can be an effective way of mimicking chronic states (Bargh, Bond, Lombardi, & Tota, 1986; Higgins & Bargh, 1987). In political psychology, for example, where self-interest is assumed to drive political and economic choices, when participants are primed with self-interest its influence is even stronger (Young, Thomsen, Borgida, Sullivan, & Aldrich, 1991). In the case of the Internet, frequent use may boost ratings of self-assessed knowledge, but in-the-moment online search independently increases ratings as well.

There are clearly benefits to the freely accessible information on the Internet; however, there may be costs inherent to the strategy of accessing that information. The boundary between personal and interpersonal knowledge is becoming increasingly blurred (Clark & Chalmers, 1998). As technology makes information ever more easily available and accessible through searching, the ability to assess one's internal "unplugged" knowledge will only become more difficult. Erroneously situating external knowledge within their own heads, people may unwittingly exaggerate how much intellectual work they can do in situations where they are truly on their own.

References

Alicke, M. D., Klotz, M. L., Breitenbecher, D. L., Yurak, T. J., & Vredenburg, D. S. (1995). Personal contact, individuation, and the better-than-average effect. *Journal of Personality and Social Psychology, 68,* 804–825. http://dx.doi.org/10.1037/0022-3514.68.5.804

Alter, A. L., Oppenheimer, D. M., & Zemla, J. C. (2010). Missing the trees for the forest: A construal level account of the illusion of explanatory depth. *Journal of Personality and Social Psychology, 99,* 436–451. http://dx.doi.org/10.1037/a0020218

Bargh, J. A., Bond, R. N., Lombardi, W. J., & Tota, M. E. (1986). The additive nature of chronic and temporary sources of construct accessibility. *Journal of Personality and Social Psychology, 50,* 869–878. http://dx.doi.org/10.1037/0022-3514.50.5.869

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58,* 7–19. http://dx.doi.org/10.1093/analys/58.1.7

Danovitch, J. H., & Keil, F. C. (2004). Should you ask a fisherman or a biologist? Developmental shifts in ways of clustering knowledge. *Child Development, 75,* 918–931. http://dx.doi.org/10.1111/j.1467-8624.2004 .00714.x

Dunning, D. (2005). *Self-insight: Roadblocks and detours on the path to knowing thyself.* New York, NY: Psychology Press. http://dx.doi.org/ 10.4324/9780203337998

Fernbach, P. M., Rogers, T., Fox, C. R., & Sloman, S. A. (2013). Political extremism is supported by an illusion of understanding. *Psychological Science, 24,* 939–946. http://dx.doi.org/10.1177/0956797612464058

File, T. (2013). Computer and Internet use in the United States. *Population Characteristics,* 1–14.

Fisher, M., & Keil, F. C. (2014). The illusion of argument justification. *Journal of Experimental Psychology: General, 143,* 425–433. http://dx .doi.org/10.1037/a0032234

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology, 60,* 115–140. http://dx.doi.org/ 10.1146/annurev.psych.59.103006.093659

Harris, J. E. (1978). External memory aids. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 172–180). London, England: Academic Press.

Harris, P. L., Pasquini, E. S., Duke, S., Asscher, J. J., & Pons, F. (2006). Germs and angels: The role of testimony in young children's ontology. *Developmental Science, 9,* 76–96. http://dx.doi.org/10.1111/j.1467-7687.2005.00465.x

Hart, J. T. (1965). Memory and the feeling-of-knowing experience. *Journal of Educational Psychology, 56,* 208.

Higgins, E. T., & Bargh, J. A. (1987). Social cognition and social perception. *Annual Review of Psychology, 38,* 369–425. http://dx.doi.org/ 10.1146/annurev.ps.38.020187.002101

Hill, G. W. (1982). Group versus individual performance: Are N + 1 heads better than one? *Psychological Bulletin, 91,* 517–539. http://dx.doi.org/ 10.1037/0033-2909.91.3.517

Hollingshead, A. B. (1998). Communication, learning, and retrieval in transactive memory systems. *Journal of Experimental Social Psychology, 34,* 423–442. http://dx.doi.org/10.1006/jesp.1998.1358

Hollingshead, A. B. (2001). Cognitive interdependence and convergent expectations in transactive memory. *Journal of Personality and Social Psychology, 81,* 1080–1089. http://dx.doi.org/10.1037/0022-3514.81.6.1080

Hutchins, E. (1995). *Cognition in the wild.* Cambridge, MA: MIT Press.

Johnson, M. K. (1997). Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences, 352,* 1733–1745. http://dx.doi.org/10.1098/rstb.1997.0156

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114,* 3–28. http://dx.doi.org/10.1037/0033-2909.114.1.3

Keil, F. C., Stein, C., Webb, L., Billings, V. D., & Rozenblit, L. (2008). Discerning the division of cognitive labor: An emerging understanding of how knowledge is clustered in other minds. *Cognitive Science, 32,* 259–300. http://dx.doi.org/10.1080/03640210701863339

Kominsky, J. F., & Keil, F. C. (2014). Overestimation of knowledge about word meanings: The "misplaced meaning" effect. *Cognitive Science, 38,* 1604–1633. http://dx.doi.org/10.1111/cogs.12122

Koriat, A., & Levy-Sadot, R. (2001). The combined contributions of the cue-familiarity and accessibility heuristics to feelings of knowing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27,* 34–53. http://dx.doi.org/10.1037/0278-7393.27.1.34

Kozlowski, S. W. J., & Ilgen, D. R. (2006). Enhancing the effectiveness of work groups and teams. *Psychological Science in the Public Interest, 7,* 77–124.

Littlepage, G., Robison, W., & Reddington, K. (1997). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes, 69,* 133–147. http://dx.doi.org/10.1006/obhd.1997.2677

Mills, C. M., & Keil, F. C. (2004). Knowing the limits of one's understanding: The development of an awareness of an illusion of explanatory depth. *Journal of Experimental Child Psychology, 87,* 1–32. http://dx .doi.org/10.1016/j.jecp.2003.09.003

Peltokorpi, V. (2008). Transactive memory systems. *Review of General Psychology, 12,* 378–394. http://dx.doi.org/10.1037/1089-2680.12.4.378

Pronin, E. (2009). The introspection illusion. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (pp. 1–67). Burlington, VT: Academic Press.

Rand, D. G. (2012). The promise of Mechanical Turk: How online labor markets can help theorists run behavioral experiments. *Journal of Theoretical Biology, 299,* 172–179. http://dx.doi.org/10.1016/j.jtbi.2011.03 .004

Rozenblit, L., & Keil, F. (2002). The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26,* 521–562. http:// dx.doi.org/10.1207/s15516709cog2605_1

Sparrow, B., Liu, J., & Wegner, D. M. (2011). Google effects on memory: Cognitive consequences of having information at our fingertips. *Science, 333,* 776–778. http://dx.doi.org/10.1126/science.1207745

Taylor, M., Esbensen, B. M., & Bennett, R. T. (1994). Children's understanding of knowledge acquisition: The tendency for children to report that they have always known what they have just learned. *Child Development, 65,* 1581–1604. http://dx.doi.org/10.2307/1131282

Ward, A. F. (2013a). Supernormal: How the Internet is changing our memories and our minds. *Psychological Inquiry, 24,* 341–348. http://dx .doi.org/10.1080/1047840X.2013.850148

Ward, A. F. (2013b). *One with the Cloud: Why people mistake the Internet's knowledge for their own* (Unpublished doctoral dissertation). Cambridge, MA: Harvard University.

Wegner, D. M. (1987). Transactive memory: A contemporary analysis of the group mind. In B. Mullen & G. R. Goethals (Eds.), *Theories of group behavior* (pp. 185–208). New York, NY: Springer-Verlag. http://dx.doi .org/10.1007/978-1-4612-4634-3_9

Wegner, D. M., Erber, R., & Raymond, P. (1991). Transactive memory in close relationships. *Journal of Personality and Social Psychology, 61,* 923–929. http://dx.doi.org/10.1037/0022-3514.61.6.923

Wegner, D. M., Giuliano, T., & Hertel, P. T. (1985). Cognitive interdependence in close relationships. In W. J. Ickes (Ed.), *Compatible and incompatible relationships* (pp. 253–276). New York: Springer-Verlag. http://dx.doi.org/10.1007/978-1-4612-5044-9_12

Young, J., Thomsen, C. J., Borgida, E., Sullivan, J. L., & Aldrich, J. H. (1991). When self-interest makes a difference: The role of construct accessibility in political reasoning. *Journal of Experimental Social Psychology, 27,* 271–296. http://dx.doi.org/10.1016/0022-1031(91)90016-Y

Appendix A

Questions Used in the Induction Phase

Why are there leap years?
Why are there more women than men?
Why are there phases of the moon?
Why are there time zones?

How does a zipper work?
Why are there dimples on a golf ball?
Why are there jokers in a deck of cards?
How is glass made?

Appendix B

Induction Phase Instructions From Experiment 1a

Internet condition:

We are interested in how well people can explain the answers to common questions. Please search the Internet to confirm the details of the explanation, and then evaluate. Please copy and paste the URL of the most helpful website in the space provided.

No Internet condition:

In this task, you will be asked a series of questions. We are interested in how well people can explain the answers to common questions. Please evaluate your understanding, using no outside sources.

Appendix C

Topics and Questions Used as the Dependent Measure in Experiments 1–2, 4–5a-b

Weather

Consider the following questions about weather:

1. Why are there more Atlantic hurricanes in August and September?

2. How do tornadoes form?

3. Why are cloudy nights warmer?

American History

Consider the following questions about American history:

1. Why did the Civil War begin?

2. How were the first labor unions formed?

3. Why did Nixon resign?

Science

Consider the following questions about science:

1. How do scientists determine the dates of fossils?

2. How do scientists know that the universe is expanding?

3. Why can't x-rays penetrate lead?

Food

Consider the following questions about food:

1. What is gluten?

2. Why does Swiss cheese have holes?

3. How is vinegar made?

(Appendices continue)

Anatomy and Physiology

Consider the following questions about anatomy and physiology:

1. Why do people laugh?

2. How does the heart pump blood?

3. Why do men go bald?

Health Issues

Consider the following questions about health issues:

1. Why are so many people allergic to peanuts?

2. Why can't HIV be transmitted through saliva?

3. Why can't you drink on antibiotics?

Appendix D

Topics and Easier Questions Used as the Dependent Measure in Experiment 3

Personal History

Consider the following questions about your personal history:

1. How did you choose your current career?

2. Why did you choose to live where you currently live?

3. How did you decide what to study during high school?

Relationships

Consider the following questions about relationships:

1. Why are you close with your best friend?

2. How are you similar to your mother?

3. How could you become friendlier with your next door neighbor?

Local Culture

Consider the following questions about the local culture where you live:

1. How does the way people dress in your town reflect their socioeconomic status?

2. How is your town different from other parts of the country?

3. How do the restaurants near where you live reflect your state's culture?

Personal Habits

Consider the following questions about personal habits:

1. How do you choose what music to listen to?

2. How do you decide what to wear on important days?

3. How do you decide what to do on the weekend?

Future

Consider the following questions about the future:

1. How will you feel when you become elderly?

2. How will you try to succeed next week?

3. How will your life satisfaction be one year from now?

Emotions

Consider the following questions about emotions:

1. Why do you become annoyed by some things that don't seem to bother others?

2. Why do you become frustrated?

3. What causes you to feel most alive?

(Appendices continue)

Appendix E

Topics and Difficult Questions Used as the Dependent Measure in Experiment 3

Personal History

1. What is the relationship between the classes you chose during freshman year of high school and your current career?

2. How did the number of windows in your current living space influence your feelings of social connectedness after you moved in?

3. How did your learning style in your high school freshman year math class affect your later interest in miniature golf?

Relationships

1. How does your best friend influence your protein intake?

2. What are the origins of the difference in the degree to which you and your mother enjoy the genre of 'Mystery'?

3. How could you discover enough about your next door neighbor's sense of humor enough to reliably predict when he or she will laugh?

Local Culture

1. How does the menu organization at individually owned restaurants in your town compare with the menu organization at chain restaurants in your town?

2. How is your town's or county's governing body different from where your relatives live?

3. In your area, how are people's hairstyles correlated with their religious beliefs?

Personal Habits

1. How do songs in the key of D affect your mood the next day?

2. How do car advertisements affect the clothes you wear on formal occasions?

3. How does the way you make weekend plans reflect the way your father made weekend plans as a child?

Future

1. In what ways will being elderly be similar to the time times of physical discomfort you have already experienced?

2. How will the number of phone calls you make at your job affect the ways in which you try to succeed next week?

3. One year from now, how will your attention to detail affect your life satisfaction?

Emotions

1. How does being annoyed affect how likely you are to attend a sporting event next year?

2. How are your current feelings of frustration related to your first memory?

3. How do lunar patterns affect your emotional well-being?

Appendix F

Questions Used in the Induction Phase of Experiment 4b

Questions with answers online:

Why is ancient Egyptian history more peaceful than Mesopotamian history?

How does the location of Cameroon affect the health of its inhabitants? How do mountains affect the weather? How did the Erie Canal affect New York City?

Questions without answers online:

Why is ancient Kushite history more peaceful than Greek history?

How does the location of Pierre, South Dakota, affect the health of its inhabitants?

How do wheat fields affect the weather?

How did the Erie Canal affect Tioga County?

BRIEF REPORT

Stress Increases Cue-Triggered "Wanting" for Sweet Reward in Humans

Eva Pool, Tobias Brosch, Sylvain Delplanque,
and David Sander
University of Geneva

Stress can increase reward pursuits: This has traditionally been seen as an attempt to relieve negative affect through the hedonic properties of a reward. However, reward pursuit is not always proportional to the pleasure experienced, because reward processing involves distinct components, including the motivation to obtain a reward (i.e., wanting) and the hedonic pleasure during the reward consumption (i.e., liking). Research conducted on rodents demonstrates that stress might directly amplify the cue-triggered wanting, suggesting that under stress wanting can be independent from liking. Here, we aimed to test whether a similar mechanism exists in humans. We used analog of a Pavlovian-Instrumental Transfer test (PIT) with an olfactory reward to measure the cue triggered wanting for a reward but also the sensory hedonic liking felt during the consumption of the same reward. The analog of a PIT procedure, in which participants learned to associate a neutral image and an instrumental action with a chocolate odor, was combined with either a stress-inducing or stress-free behavioral procedure. Results showed that compared with participants in the stress-free condition, those in the stress condition mobilized more effort in instrumental action when the reward-associated cue was displayed, even though they did not report the reward as being more pleasurable. These findings suggest that, in humans, stress selectively increases cue-triggered wanting, independently of the hedonic properties of the reward. Such a mechanism supports the novel explanation proposed by animal research as to why stress often produces cue-triggered bursts of binge eating, relapses in drug addiction, or gambling.

Keywords: stress, incentive salience, wanting, liking, human Pavlovian-Instrumental Transfer

Supplemental materials: http://dx.doi.org/10.1037/xan0000052.supp

Have you ever eaten more high-calorie foods during a stressful period? As documented by a consistent corpus of literature (e.g., O'Connor & Conner, 2011), stress cannot only increase the consumption of high-calorie foods, but it can also increase the use of other kinds of rewards, such as drugs (Sinha, 2001) or sexual stimuli (Chumbley et al., 2014). Although these effects of stress have been proven to have a large impact on public health problems (e.g., addiction relapses or binge eating; Lo Sauro, Ravaldi, Cabras, Faravelli, & Ricca, 2008), the underlying psychological mechanisms remain poorly understood.

It has been proposed that rewards are used to reduce the negative effects of stress, which are compensated by the hedonic pleasure triggered by their consumption (Koob & Le Moal, 2001). According to this proposal, stress increases the pursuit of rewards, as consumption is made even more pleasurable by relieving the negative effects of stress.

The incentive salience theory proposes an alternative mechanism in which the key principle is independent of the hedonic properties of the reward (Berridge & Robinson, 1998). According to this theory, the pursuit of a reward is not always directly proportional to the pleasure experienced, because reward processing involves distinct components, including the motivation to obtain a reward (i.e., wanting) and the hedonic pleasure during the reward consumption (i.e., liking), which are usually correlated but can be dissociated under particular circumstances. Experiments conducted on rodents showed that direct manipulation of mesolimbic dopamine increases effort mobilization after the presentation of a reward-associated cue (i.e., wanting), without simultaneously increasing hedonic pleasure (i.e., liking) during reward consumption (Peciña, Cagniard, Berridge, Aldridge, & Zhuang, 2003;

Wyvell & Berridge, 2000). This supports the idea that wanting and liking rely on two distinct neuronal networks that can be activated independently of each other (Berridge & Robinson, 1998).

Manipulating the glucocorticoid system in a population of rodents, Peciña, Schulkin, and Berridge (2006) provided evidence suggesting that the stress-induced increase of reward pursuits might be driven by a selective increase of wanting. The glucocorticoid system is known to be involved in the mediation of physiological and behavioral responses to stress (Herman et al., 2003). To investigate its effects on wanting, Peciña and coworkers (2006) used a three-phase paradigm called the Pavlovian-Instrumental Transfer test (PIT; Lovibond, 1983). First, during instrumental conditioning, a behavioral response (e.g., pressing a lever) was associated with the food reward. Second, during Pavlovian conditioning, neutral stimuli (i.e., sounds) were associated with the absence or presence of a food reward (negatively or positively conditioned stimulus: conditional stimulus CS− or CS+). During the transfer test, the Pavlovian stimuli (CS+, CS−) were presented and their influence on instrumental action was measured: The increase in action energization after CS+ presentation was taken to reflect cue-induced wanting. After Pavlovian and instrumental conditioning, but before the transfer test, Peciña and colleagues (2006) manipulated the glucocorticoid system by microinjecting cortisol-releasing factor in the nucleus accumbens. Rodents in which the glucocorticoid system was stimulated showed a larger cue-induced wanting compared with rodents who received a placebo treatment. The transfer test was administered under extinction, thus rodents never experienced any aversive state reduction triggered by the hedonic properties of the reward. The investigators thereby demonstrated that, similar to dopamine manipulation, manipulation of stress-related systems increased wanting, independently of the relieving liking dimension of the reward. Although these findings provide a novel explanation for the stress-induced increase of reward pursuits, to the best of our knowledge, they have never been demonstrated in humans.

This study aimed to investigate whether stress influences wanting in humans by using methods and concept operationalizations comparable with relevant animal research. We used an analog of a PIT adapted to a human population (Talmi, Seymour, Dayan, & Dolan, 2008): During instrumental conditioning, an instrumental action (i.e., pressing a handgrip) was associated with chocolate odor (unconditioned stimulus, US). During an analog of a Pavlovian conditioning task geometric figures were associated with the presence (CS+) or absence (CS− and baseline) of chocolate odor (see Pool, Brosch, Delplanque, & Sander, 2014), and during the transfer test, the effort mobilized on the handgrip was measured during the presentation of the Pavlovian stimuli (CS+, CS−, and baseline). After the instrumental and the analogous Pavlovian conditioning but before the transfer test, participants underwent either a stress-inducing task or a control stress-free task (Schwabe, Haddad, & Schachinger, 2008). Participants in the stress and stress-free conditions then performed the transfer test and evaluated how much they liked smelling the chocolate odor.

Based on the animal literature, our prediction was that the cue-induced wanting would be larger in the stress condition than in the stress-free condition, reflected by a larger increase in the effort mobilized during the presentation of CS+ compared with other CSs. We expected this increase of cue-induced wanting for chocolate odor even in the absence of an increase of hedonic pleasure during chocolate odor presentation.

Method

Participants

Forty-one participants who liked chocolate were recruited at the University of Geneva. They were asked not to eat, practice sport, or drink coffee 4 hr before the experimental session and received 30 Swiss francs for their participation. Five participants were later excluded: 2 for technical problems and 3 for being under psychotropic treatment. The 36 (19 men) remaining participants (24.15 ± 3.05 years old) had no reported olfactory trouble.

Materials

Stimuli. The Pavlovian stimuli consisted of three geometric complex figures typically used in human conditioning paradigms (Gottfried, O'Doherty, & Dolan, 2003; O'Doherty et al., 2004; Valentin, Dickinson, & O'Doherty, 2007) that in a pilot study ($n = 26$) were rated as similarly neutral on a pleasantness scale (see Pool et al., 2014). They were displayed in the center of the computer screen with a visual angle of 8°. The Pavlovian identities of three images used as CS+, CS−, and baseline were counterbalanced across participants. The US consisted of a chocolate odor (20% dissolved in propylene glycol; Firmenich, SA, Geneva, Switzerland), which was released for 1.5 s by using a computer-controlled olfactometer with an airflow fixed at 1.5 L/min delivering the olfactory stimulation rapidly, without thermal and tactile confounds via a nasal cannula (see Ischer et al., 2014).

Instrumental apparatus. The mobilized effort was measured through an isometric handgrip (TDS121C) connected to the MP150 Biopac Systems (Santa Barbara, CA) with a 1,000 Hz sampling rate. The dynamic value of the signal was read by MATLAB (version 8.0) and used to provide participants with visual online feedback (Psychtoolbox 3.0; for the visual interface implemented in MATLAB) that reflected the force exerted on the handgrip. This visual feedback was illustrated through the "mercury" of a thermometer-like image displayed on the left side of the screen (30° visual angle) that moved up and down according to the mobilized effort (see Figure 1a and 1c). The mercury of the thermometer-like display reached the top if the handgrip was squeezed with at least 50% or 70% (criterion varied every 1 s) of the participants' maximal force.

Procedure

First, participants completed the instrumental and the analogous Pavlovian conditioning. After the instrumental and the analogous Pavlovian conditioning but before the transfer test, 18 participants then performed the stress-inducing task, whereas the other group performed a stress-free task. Subsequently, they took a 10-min break, and then performed the PIT test (adapted from Talmi et al., 2008; see Table 1). Finally, they evaluated the perceived pleasantness of the odors.

Instrumental conditioning. Participants learned to squeeze a handgrip to trigger the release of chocolate odor. There were 24 trials each comprised of a 12-s "task-on" period followed immediately by a "task-off" period of 4–12 s (8 s average).

During the task-on periods, a geometric image and a thermometer were displayed in the center and on the left side of the screen,
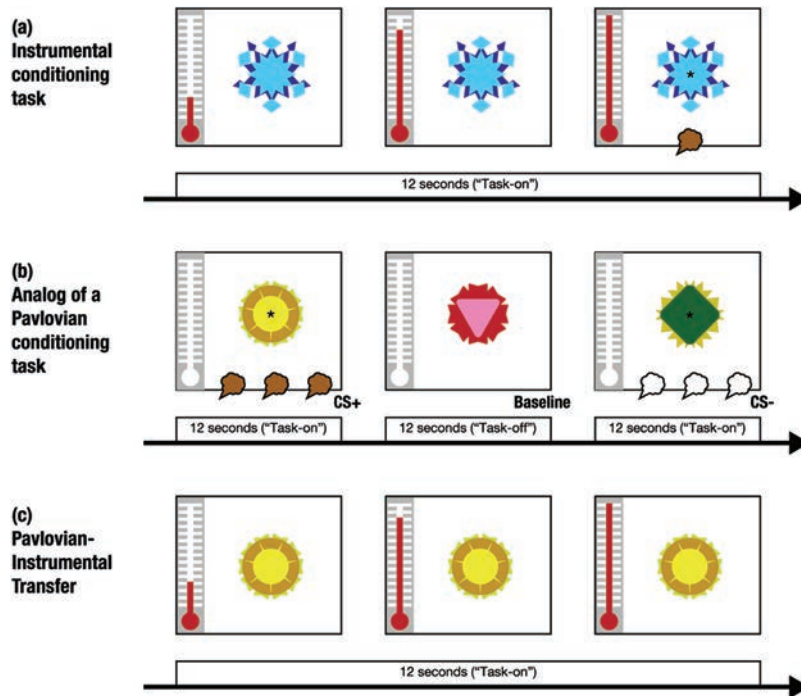
Figure 1. The analog of a human Pavlovian-Instrumental Transfer (PIT) paradigm adapted from Talmi et al. (2008). During instrumental conditioning, (a) participants learned to squeeze a handgrip to trigger the release of a rewarding chocolate odor. During the analogous Pavlovian conditioning, (b) participants were exposed to repeated pairings of the positive conditioned stimulus (CS+) with the rewarding chocolate odor and the negative conditioned stimulus (CS−) with the odorless air. When the CS+ or CS− was displayed, a target appeared in the center of the image and participants had to press a key that triggered odor release. The baseline was displayed without any target, and no odor was released. The PIT test (c) was administered under extinction, the CS+, the CS−, and the baseline were displayed in random order (here a CS+ trial is illustrated), and participants could squeeze the handgrip if they wished to do so. See the online article for the color version of this figure.

respectively. The fluid movement of the thermometer-like display's mercury provided online visual feedback of the effort participants exerted on the handgrip (see Figure 1a). Participants were asked to keep their gaze on the central geometric image and to squeeze the handgrip, thereby bringing the mercury of the thermometer- like display up to the maximum and then down again, without paying attention to the speed of compressing the grips. They were told that during the 12-s presentation of the thermometer-like display, there were three "special 1-s windows" and that if they happened to squeeze the handgrip during one of

these time windows, they would trigger the release of chocolate odor. Finally, they were told that they were free to choose when to squeeze on the grip and were encouraged to use their intuition. In reality, only two special 1-s windows were randomly selected in each task-on period to be rewarded with chocolate odor, and if participants happened to squeeze the handgrip with at least 50% or 70% of their maximal force during these time windows, a sniffing signal (a black asterisk; 2° visual angle) was displayed at the center of the geometric image and the chocolate odor was delivered. During the task-off periods, a fixation cross (2° visual angle) was

Table 1

Illustration of the Analog of a Human Pavlovian Instrumental Transfer Test Combined With Stress Induction

Phase 1 Instrumental conditioning	Phase 2 Analogous Pavlovian conditioning	Stress induction	Phase 3 Transfer test in extinction
24 trials R → O1	18 trials CS+ → O1 18 trials CS− → Ø	SECPT or WWT	6 trials CS+ → R? 6 trials CS− → R? 6 trials baseline → R?

Note. The instrumental action (R) consists of pressing a handgrip dynamometer. The conditioned stimuli (CS) consist of geometric images. The CS+ is associated with a chocolate odor (O1), whereas the CS− is associated with the absence of a chocolate odor (Ø). Participants in the stress group performed the socially evaluated cold pressor test (SECPT; Schwabe et al., 2008), thereby inducing stress. Participants in the stress-free group performed a warm water test (WWT; Schwabe et al., 2008). The analog of the human Pavlovian instrumental transfer procedure has been adapted from Talmi et al. (2008) for our research question.

displayed at the center of the screen and participants were asked to keep their gaze on the fixation cross and to relax their hand to recalibrate the baseline force.

Analogous Pavlovian conditioning. The procedure that we previously elaborated (Pool et al., 2014) was applied here. Briefly, three initially neutral images were attributed the Pavlovian roles of "baseline," "CS+," and "CS−." There were 36 trials composed of a 12-s task-on period during which the CS+ or the CS− was displayed on a computer screen, followed by a 12-s task-off period during which a baseline image was displayed.

During the task-on periods, a target appeared every 4 s at the center of the CS image three times per period. Participants had to press the "A" key as fast as possible after they perceived the target that was presented for a maximum of 1 s. Each time the CS+ image was displayed and the participant pressed the key, a chocolate odor was released; when the CS− image was displayed, odorless air was released. Participants were informed that the kind of odor released depended only on the CS image and not on the key-pressing task. They were told that the key-pressing task was a measure of their sustained attention independent of the odor-image contingencies (see also Talmi et al., 2008). To further emphasize this aspect, participants were also informed that the odor would be released after a 1-s interval after target onset if they had not responded until then. However, the odor was released faster when participants pressed on the keyboard than when participants did not press on the keyboard. Because of this instrumental component the conditioning task can be considered a hybrid of Pavlovian and discriminative instrumental learning, rather than a pure Pavlovian learning. During the task-off periods, the baseline image was displayed without any target, and no odor was released (see Figure 1b).

After the analogous Pavlovian conditioning, participants evaluated the pleasantness of the images used as CS+, CS−, and baseline on a visual analog scale (from extremely unpleasant to extremely pleasant) presented at the center of the computer screen with a visual angle of 23°. The order of the images was randomized across participants.

Stress manipulation. After the pleasantness ratings of the images used as CS+, CS−, and baseline, participants belonging to the stress group ($n = 18$; 8 men) performed a socially evaluated cold pressor test (as described by Schwabe et al., 2008). They were asked to immerse their nondominant hand in cold water (0–2 °C) for as long as possible (they could remove the hand at their discretion, but the test was ended by the experimenter after 3 min). They were told that they were being videotaped to analyze their facial expression, and the experimenter observed them the entire time. The stress-free group ($n = 18$; 10 men) was instructed to put their hand in warm water (35–37 °C) for 3 min without being observed (warm water test). Immediately after the cold pressor task, using pencil and paper participants evaluated on a scale how pleasant, stressful, and painful the task was for them, from 0 (*not at all*) to 10 (*extremely*). Ten minutes before and 30 min after the task, samples of saliva were collected through Salivette (Sarstedt AG & co, Nümbrecht, Germany), as a manipulation check.

PIT test. After the stress induction task participants took a 10-min break and then they received the same instructions as in instrumental conditioning. First they completed 12 trials identical to those in instrumental conditioning (two special 1-s windows were rewarded) followed by 12 trials administered under partial

extinction (one special 1-s window was rewarded). Immediately afterward, they performed 18 transfer test trials administered under extinction (no time window was rewarded). In the transfer test, one of the Pavlovian stimuli (CS+, CS−, or baseline) replaced the instrumental geometric image during the entire trial (see Figure 1c). The presentation order on the transfer tests was randomized across the three stimuli (CS+, CS−, and baseline). There were two cycles of testing. In each cycle, each cue was presented three times consecutively, so that each Pavlovian stimuli was presented six times for a total of 18 transfer trials.

Odor evaluation. Immediately after the PIT, participants evaluated the pleasantness (from extremely unpleasant to extremely pleasant), the familiarity (from not familiar at all to extremely familiar), the edibility (from not edible to extremely edible), and the intensity (from not perceived to extremely strong) of the chocolate odor and the odorless air on visual analog scales displayed on a computer screen.

Results

Instrumental Conditioning

A repeated-measures analysis of variance (ANOVA) applied to the number of squeezes surpassing 50% of each participant's maximal force (Talmi et al., 2008) over 24 trials revealed a marginal effect of trial, $F(23, 805) = 1.50, p = .06, \eta^2 = .04, 95\%$ confidence intervals (CIs) = .00, .05 suggesting that participants learned that squeezing the handgrip triggered the release of the rewarding chocolate odor. Figure 2A shows that participants readily learned to squeeze after five trials; a linear contrast showed that the squeeze frequency increased linearly during these first five trials, $t(35) = 2.99, p < .01, d = .32, 95\%$ CIs = .09, .53. This increase was not significant in Trials 6–15 or in Trials 15–24 ($ps > .6$).

To control that participants assigned to stress and stress-free groups did not statistically differ in their Pavlovian learning, we applied a 5 (trials: 1, 2, 3, 4, or 5) × 2 (group: stress or stress-free) mixed repeated-measures ANOVA to the number of squeezes surpassing the criterion of 50% of the participants' maximal force. The analysis revealed a main effect of trial $F(4, 136) = 5.23, p < .01, \eta^2 = .13, 95\%$ CIs = .01, .54, without a significant interaction between image and group ($p = .69$); therefore, suggesting that the increase of squeeze frequency over time was similar in participants that would have later been assigned to the stress and the stress-free group. This analysis also revealed a descriptive difference in the average squeeze frequency between the two groups, participants that would have later been assigned to the stress-free group squeezed on average 1.88 more than the participants that would have later been assigned to the stress group. This difference was not significant ($p = .13$), but it was large. Therefore, we decided to control for these pre-existing differences in all the statistical tests assessing the effect of stress by comparing the squeeze frequency between the two groups after the stress induction task.

Analogous of Pavlovian Conditioning

Successful Pavlovian contingency learning was revealed by both the reaction times (RTs) of the key-pressing task and the
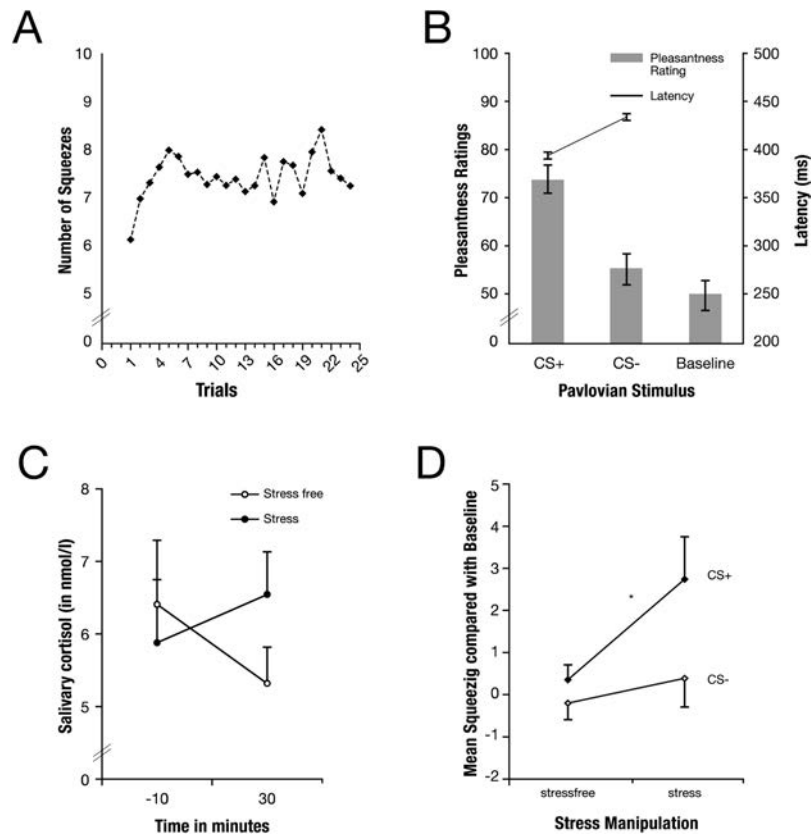
Figure 2. (A) Instrumental conditioning. The number of times participants (*n* = 36) squeeze the handgrip is displayed as a function of trials over time. (B) The analogous Pavlovian conditioning. The bars (left axis) illustrate the pleasantness rating of the images used as Pavlovian stimuli after conditioning, and the line plot (right axis) illustrates the latency to detect the cue during the presentation of the positive conditioned stimulus (CS+) or the negative conditioned stimulus (CS−). Error bars (±1 *SEM*) are adapted for within-subject design (Cousineau, 2005). (C) Salivary cortisol (nanomoles per liter) 10 min before and 30 min after the stress-inducing and the control task. Error bars represent *SEM*. (D) The PIT in the stress-free and the stress groups (*n* = 18 in each group). The increase of the number of squeezes compared with the baseline is displayed as a function of the CS+ and the CS−. Error bars represent *SEM*.

likability of the CSs (see Figure 2B). For the key-pressing task, we analyzed RTs on the first target of the on-task period. All responses that were more than 3 *SD*s from the mean (>2% of the trials) or absent (>5.5% of the trials) were removed. A paired *t* test showed that participants were faster when the CS+ was displayed than when the CS− was displayed, $t(35) = 3.60, p < .01, d = .37$, 95% CIs = .15, .39.

A repeated-measures ANOVA applied to the likability ratings of the three Pavlovian images (CS+, CS−, or baseline) revealed a significant effect, $F(2, 70) = 10.89, p < .01, \eta^2 = .23$, 95% CIs = .14, .55; participants liked the CS+ image more than the CS−, $t(35) = 3.47, p < .01, d = .73$, 95% CIs = .27, 1.17, which did not statistically differ from the baseline in likability ratings (*p* = .35).

To control that participants assigned to the stress and stress-free groups did not statistically differ in their Pavlovian learning, we applied a 2 (image: CS+ or CS−) × 2 (group: stress or stress-free) mixed repeated-measures ANOVA to the RTs. The analysis revealed a main effect of image, $F(1, 34) = 12.66, p = .01, \eta^2 = $

.27, 95% CIs = .05, .48, without interaction between image and group (*p* = .67); therefore, suggesting that the difference between CS+ and CS− was similar in participants that would later be assigned to the stress and the stress-free group. A 3 (image: CS+, CS−, baseline) × 2 (group: stress or stress-free) mixed repeated-measures ANOVA was applied to the likability ratings. The analysis revealed a main effect of image, $F(2, 68) = 10.83, p < .01, \eta^2 = .24$, 95% CIs = .12, .57, without a significant interaction between the effect of image and group (*p* = .44), thereby suggesting that the effect of image on the likability rating was similar in participants that would later be assigned to the stress and the stress-free groups.

Stress Manipulation

On average, participants who performed the socially evaluated cold pressor test (stress group) kept their hand in the cold water for 80.61 s (*SEM* = 14.75) and they reported a higher level of stress (*M* = 5.22, *SEM* = 0.67) and pain (*M* = 6.33, *SEM* = 0.67),

$ts(34) = 5.95, 9.42, ps < .01, ds = 2.04; 3.23, 95\%$ CIs $= 1.16,$
2.77; 2.13, 4.11 compared with the participants who performed the
warm water test (stress-free group; $Ms = 0.77, 0.33, SEMs = 0.33,$
0.16, respectively). Participants in the stress group also reported a
lower level of pleasure ($M = 2.11, SEM = 0.46$) than participants
in the stress-free group ($M = 6.99, SEM = 0.58; t(34) = 6.421,$
$p < .001\ d = 2.202, 95\%$ CIs $= 1.23, 2.88$). Moreover, the prepost
variation of cortisol induced by the task was marginally larger in
the stress group compared with the stress-free group, $t(34) = 2.20,$
$p = .051\ d = .75, 95\%$ CIs $= .05, 1.40$ (see Figure 2C).

PIT test

A 3 (image: CS+, CS−, and baseline) × 6 (extinction trial) ×
2 (group: stress or stress-free) mixed repeated-measures ANOVA
was applied to the number of squeezes surpassing 50% of the
participants' maximal force. Because there was a large difference
in the average squeeze frequency during the instrumental condi-
tioning between participants attributed to the stress and the stress-
free groups, the average squeeze frequency during the instrumental
conditioning was modeled as a covariate in all the between groups
tests. The analysis revealed a main effect of image, $F(2, 68) =$
$6.61, p < .01, \eta^2 = .16, 95\%$ CIs $= .04, .47$, indicating that the
squeeze frequency increased during the CS+ compared with the
CS−, $t(35) = 2.79, p < .01, d = .41, 95\%$ CIs $= .11, .72$, which
did not statistically differ from the baseline ($p = .83$). Moreover,
the analysis revealed a main effect of trial number, $F(5, 170) =$
$2.71, p = .02, \eta^2 = .07, 95\%$ CIs $= .01, .44$, showing that the
squeeze frequency globally decreased over time. Most important
for our hypothesis, the analysis showed a two-way interaction
between image and group, $F(2, 66) = 5.911, p < .01, \eta^2 = .15,$
95% CIs $= .04, .46$, revealing that the increase in number of
squeezes toward the CS+ (compared with the CS−) was larger in

the stress group compared with the stress-free group, $t(34) = 2.32,$
$p = .02, d = .77, 95\%$ CIs $= .09, 1.44$ (see Figure 3A and B and
Figure 4A). Furthermore, this analysis did not reveal a main effect
of group ($p > .1$), suggesting that the stress and the stress-free
groups did not statistically differ in the overall squeeze frequency.

To further investigate whether the effect of stress was specific to
the CS+, we computed the relative score for the CS+ and the CS−,
by subtracting the squeeze frequency during the presentation of the
baseline image from the squeeze frequency during the presentation of
the CS (i.e., [CS+− baseline] and [CS−− baseline]). We then
conducted two planned contrasts. These contrasts revealed that the
relative increase in number of squeezes toward the CS+ was signif-
icantly larger in the stress group compared with the stress-free group,
$t(34) = 2.39, p = .03, d = 1.04, 95\%$ CIs $= .11, 1.47$, but the relative
number of squeezes toward the CS− did not significantly differ
between the two groups ($p > .4$; see Figure 2D).

Odor Evaluation

Paired t tests revealed that the chocolate odor was evaluated as
being more edible, $t(34) = 9.38, p < .01, d = 1.27, 95\%$ CIs $=$
$.63, 1.43$; more intense, $t(34) = 10.88, p < .001, d = 2.06, 95\%$
CIs $= 1.42, 2.61$; more familiar, $t(34) = 5.28, p < .01, d = .70,$
95% CIs $= .38, .99$; and more pleasant, $t(34) = 6.19, p < .01, d =$
$1.60, 95\%$ CIs $= .63, 1.43$, than the odorless air. The perceived
pleasantness of the chocolate odor was not significantly different
in the stress and the stress-free group, $t(34) = .19, p = .85$ (see
Figure 4B). Moreover, the two groups did not significantly differ
on the perception of the other dimensions of odor (all $ps > .4$), nor
in the differential perceived pleasantness of the chocolate odor
compared with the odorless air ($p > .7$)

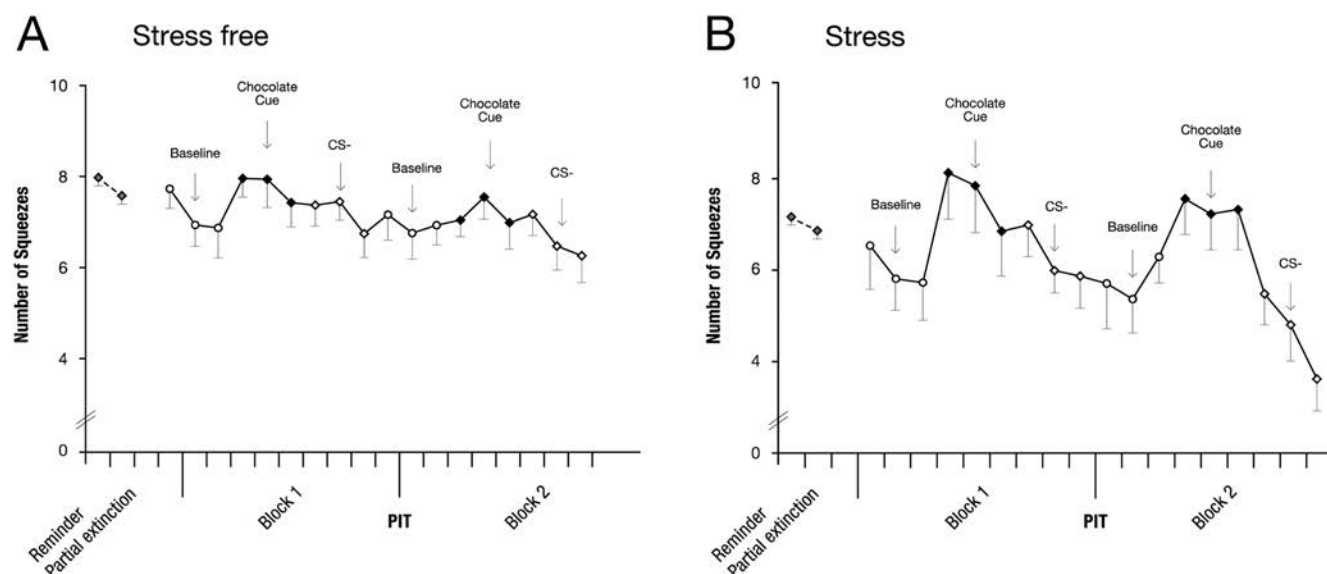*Figure 3.* PIT in the stress-free (A) and the stress (B) group. The number of times participants ($n = 18$ in each
group) squeezed the handgrip is displayed as a function of the conditioned stimuli (CSs) perceived by the
participants during the block administered under extinction. Each CS was presented three times in a row during
one block and the presentation order of the CSs in each block was randomized.
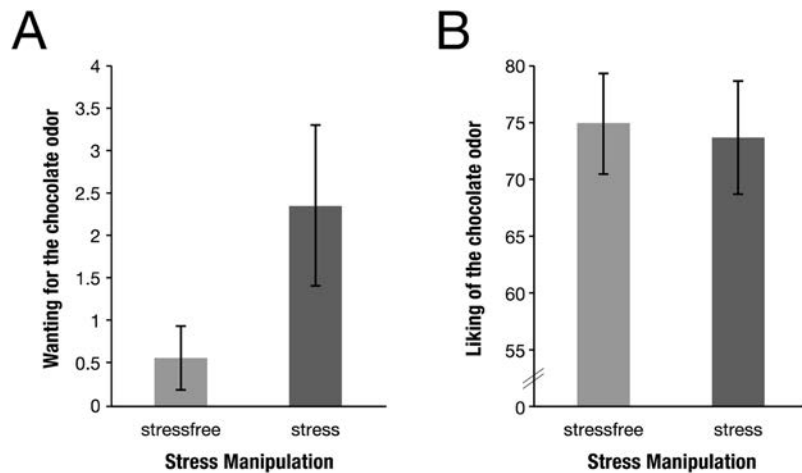
Figure 4. (A) Wanting for chocolate odor (increase in the number of squeezes during presentation of the CS+ compared with the CS−). (B) Liking of chocolate odor (rating on a scale from 0 to 100). Error bars represent *SEM*.

Discussion

This study aimed to investigate whether, as predicted by the incentive salience theory, stress increases wanting in humans. We used paradigms and concept operationalizations that were as similar and comparable as possible to those used in previous research conducted on rodents. We adapted an analog of a human PIT, which originally used a monetary reward (Talmi et al., 2008), by instead using an olfactory reward (i.e., chocolate odor) to assess the effort mobilized to obtain it (i.e., wanting) and the hedonic pleasure during its consumption (i.e., liking); we administered this paradigm in stress and stress-free conditions. Our findings demonstrate, in humans, that the cue-triggered wanting observed in the analog of a PIT was amplified in the stress condition compared with the stress-free condition. Moreover, results showed that global effects in the analog of a PIT with an olfactory reward are similar to those in the analog of a PIT with a monetary reward: The effort mobilized in instrumental action is influenced by the presentation of Pavlovian stimuli. The effort invested in instrumental action was larger during the presentation of the stimulus that was previously associated with chocolate odor, as compared with the presentation of the stimuli that were not.

Note that the analogous Pavlovian conditioning we used involved an instrumental component (i.e., pressing a key to discover whether the image was associated with a chocolate odor or not). Colwill and Rescorla (1988) demonstrated that transfer effect is significantly bigger for discriminative stimuli (i.e., stimuli predicting a privileged time interval during which an instrumental action leads to a reward) compared with Pavlovian stimuli. Nonetheless, in contrast to the discriminative instrumental learning, in our Pavlovian procedure the instrumental action had limited predictive value: the chocolate odor was delivered based on the CS image, even when the instrumental action was not performed.

Comparable with that in rodents (Peciña et al., 2006), this wanting amplification in humans is critically dependent on the interaction between the state of the individual (i.e., stress) and the presence of a reward-associated cue in the environment (see Figure

S1 in Supplemental Materials for a visual comparison). Stress did not globally increase effort mobilization, therefore, ruling out possible motor confounds as well as the possibility that the increase of effort induced by stress relied on a general state of arousal. Descriptively, the stress-free group squeezed globally more than the stress group during the PIT; however, this descriptive difference was not statistically significant and it was already present during the instrumental conditioning was administered before the stress manipulation. Therefore, differences in the global squeeze frequency were probably because of a simple sampling effect rather than an effect of stress (see supplementary analysis for more details). Consistent with the animal literature (e.g., Peciña et al., 2006), the stress amplification seemed to be specific to the perception of the reward-associated cue: during the presentation of the reward-associated cue, the stress group mobilized more effort than the stress-free group, but during the presentation of the stimulus that was not associated with the reward the stress-free and the stress group did not seem to behave differently. Moreover, the stress induction procedure was administered after the Pavlovian and instrumental conditioning, thereby excluding possible confounds related to learning processes (see, e.g., Allman, DeLeon, Cataldo, Holland, & Johnson, 2010, for a similar procedure).

Our results also showed that, although participants mobilized more effort to smell the chocolate odor when under stress, they did not report the odor as being more pleasurable. This finding supports the incentive salience theory, which postulates that wanting and liking represent two different components of reward processing that can be activated independently of each other under particular circumstances (Berridge & Robinson, 1998, 2003). It also supports the idea that the increase of reward pursuits induced by stress might not be driven by a top-down attempt to relieve the negative effects of stress through reward consumption (Koob & Le Moal, 2001), but may instead be driven by a direct bottom-up effect of stress on cue-triggered wanting (Peciña et al., 2006). Although the present finding provides evidence showing that stress increases cue-triggered wanting and not self-reported liking, the

mechanism underlying this phenomenon remains to be explored. Recent evidence from animal (Cabib & Puglisi-Allegra, 2012) and human (e.g., Lewis, Porcelli, & Delgado, 2014) literature suggests that stress increases the dopaminergic activity in the ventral striatum. According to the incentive salience theory, the dopaminergic activity in this region is selectively involved in wanting (and not in liking; Berridge & Robinson, 2003); therefore, one could speculate that stress directly and selectively activate the mesolimbic brain network involved in wanting.

In contrast to previous rodent experiments, in our study, we did not manipulate brain activity; instead, we used a more ecological procedure to induce stress behaviorally. This suggests that the selective activation of a reward component can occur not only when the brain is directly manipulated (Havermans, 2011, 2012), but can also occur in more ecological settings such as behavioral manipulation. Note, however, that in our study, differences in the increase of cortisol between the two groups was smaller compared with those in other experiments that used the same procedure of stress induction (Schwabe et al., 2008). This might be because we included female participants, thereby increasing the noise of the cortisol measure, variation being because of menstrual cycle effects on cortisol responses (Kirschbaum, Pirke, & Hellhammer, 1993; see also Figure S2 in Supplemental Materials).

The stress effect on "wanting" was significant but rather moderate. The moderate size of the effect is not surprising given that findings on rodents have similar effect sizes (Peciña et al., 2006) and that in an experimental setting we could only induce a mild stress. Therefore, future studies are necessary to further investigate the effect of more intense everyday life stressors on human wanting and liking.

In conclusion, the present study supports the conceptualization of a hedonic-independent mechanism underlying the increase of reward pursuits induced by stress (Peciña et al., 2006) by showing that, compared with a stress-free situation, in a stressful situation participants are willing to work more to obtain a reward, even though they do not report liking it more. This selective increase of cue-triggered wanting from stress might be crucial for modeling the effects of stress on binge eating, relapses in addiction, and gambling.

References

Allman, M. J., DeLeon, I. G., Cataldo, M. F., Holland, P. C., & Johnson, A. W. (2010). Learning processes affecting human decision making: An assessment of reinforcer-selective Pavlovian-to-instrumental transfer following reinforcer devaluation. *Journal of Experimental Psychology: Animal Behavior Processes, 36,* 402–408. http://dx.doi.org/10.1037/a0017876

Berridge, K. C., & Robinson, T. E. (1998). What is the role of dopamine in reward: Hedonic impact, reward learning, or incentive salience? *Brain Research Reviews, 28,* 309–369. http://dx.doi.org/10.1016/S0165-0173(98)00019-8

Berridge, K. C., & Robinson, T. E. (2003). Parsing reward. *Trends in Neurosciences, 26,* 507–513. http://dx.doi.org/10.1016/S0166-2236(03)00233-9

Cabib, S., & Puglisi-Allegra, S. (2012). The mesoaccumbens dopamine in coping with stress. *Neuroscience and Biobehavioral Reviews, 36,* 79–89. http://dx.doi.org/10.1016/j.neubiorev.2011.04.012

Chumbley, J. R., Hulme, O., Köchli, H., Russell, E., Van Uum, S. A. Pizzagalli, D., & Fehr, E. (2014). Stress and reward: Long term cortisol

exposure predicts the strength of sexual preference. *Physiology & Behavior, 131,* 33–40. http://dx.doi.org/10.1016/j.physbeh.2014.04.013

Colwill, R. M., & Rescorla, R. A. (1988). Association between the discriminative stimulus and the reinforcer in instrumental learning. *Journal of Experimental Psychology: Animal Behavior Processes, 14,* 155–164. http://dx.doi.org/10.1037/0097-7403.14.2.155

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1,* 42–45.

Gottfried, J. A., O'Doherty, J., & Dolan, R. J. (2003). Encoding predictive reward value in human amygdala and orbitofrontal cortex. *Science, 301,* 1104–1107. http://dx.doi.org/10.1126/science.1087919

Havermans, R. C. (2011). "You Say it's Liking, I Say it's Wanting . . .". On the difficulty of disentangling food reward in man. *Appetite, 57,* 286–294. http://dx.doi.org/10.1016/j.appet.2011.05.310

Havermans, R. C. (2012). How to tell where 'liking' ends and 'wanting' begins. *Appetite, 58,* 252–255. http://dx.doi.org/10.1016/j.appet.2011.10.013

Herman, J. P., Figueiredo, H., Mueller, N. K., Ulrich-Lai, Y., Ostrander, M. M., Choi, D. C., & Cullinan, W. E. (2003). Central mechanisms of stress integration: Hierarchical circuitry controlling hypothalamo-pituitary-adrenocortical responsiveness. *Frontiers in Neuroendocrinology, 24,* 151–180. http://dx.doi.org/10.1016/j.yfrne.2003.07.001

Ischer, M., Baron, N., Mermoud, C., Cayeux, I., Porcherot, C., Sander, D., & Delplanque, S. (2014). How incorporation of scents could enhance immersive virtual experiences. [Advance online publication]. *Frontiers in Psychology, 5,* 736. http://dx.doi.org/10.3389/fpsyg.2014.00736

Kirschbaum, C., Pirke, K. M., & Hellhammer, D. H. (1993). The 'Trier Social Stress Test'—a tool for investigating psychobiological stress responses in a laboratory setting. *Neuropsychobiology, 28,* 76–81. http://dx.doi.org/10.1159/000119004

Koob, G. F., & Le Moal, M. (2001). Drug addiction, dysregulation of reward, and allostasis. *Neuropsychopharmacology, 24,* 97–129. http://dx.doi.org/10.1016/S0893-133X(00)00195-0

Lewis, A. H., Porcelli, A. J., & Delgado, M. R. (2014). The effects of acute stress exposure on striatal activity during Pavlovian conditioning with monetary gains and losses. [Advance online publication]. *Frontiers in Behavioral Neuroscience, 8,* 179. http://dx.doi.org/10.3389/fnbeh.2014.00179

Lo Sauro, C., Ravaldi, C., Cabras, P. L., Faravelli, C., & Ricca, V. (2008). Stress, hypothalamic-pituitary-adrenal axis and eating disorders. *Neuropsychobiology, 57,* 95–115. http://dx.doi.org/10.1159/000138912

Lovibond, P. F. (1983). Facilitation of instrumental behavior by a Pavlovian appetitive conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes, 9,* 225–247. http://dx.doi.org/10.1037/0097-7403.9.3.225

O'Connor, D. B., & Conner, M. (2011). Effects of stress on eating behavior. In R. J. Contrada & A. Baum (Eds.), *The handbook of stress science: Biology, psychology, and health* (pp. 275–286). New York, NY: Springer.

O'Doherty, J., Dayan, P., Schultz, J., Deichmann, R., Friston, K., & Dolan, R. J. (2004). Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science, 304,* 452–454. http://dx.doi.org/10.1126/science.1094285

Peciña, S., Cagniard, B., Berridge, K. C., Aldridge, J. W., & Zhuang, X. (2003). Hyperdopaminergic mutant mice have higher "wanting" but not "liking" for sweet rewards. *The Journal of Neuroscience, 23,* 9395–9402.

Peciña, S., Schulkin, J., & Berridge, K. C. (2006). Nucleus accumbens corticotropin-releasing factor increases cue-triggered motivation for sucrose reward: Paradoxical positive incentive effects in stress? [Advance online publication]. *BMC Biology, 4,* 8. http://dx.doi.org/10.1186/1741-7007-4-8

Pool, E., Brosch, T., Delplanque, S., & Sander, D. (2014). Where is the chocolate? Rapid spatial orienting toward stimuli associated with primary rewards. *Cognition, 130,* 348–359. http://dx.doi.org/10.1016/j .cognition.2013.12.002

Schwabe, L., Haddad, L., & Schachinger, H. (2008). HPA axis activation by a socially evaluated cold-pressor test. *Psychoneuroendocrinology, 33,* 890–895. http://dx.doi.org/10.1016/j.psyneuen.2008.03.001

Sinha, R. (2001). How does stress increase risk of drug abuse and relapse? *Psychopharmacology, 158,* 343–359. http://dx.doi.org/10.1007/ s002130100917

Talmi, D., Seymour, B., Dayan, P., & Dolan, R. J. (2008). Human Pavlovian-instrumental transfer. *The Journal of Neuroscience, 28,* 360– 368. http://dx.doi.org/10.1523/JNEUROSCI.4028-07.2008

Valentin, V. V., Dickinson, A., & O'Doherty, J. P. (2007). Determining the neural substrates of goal-directed learning in the human brain. *The Journal of Neuroscience, 27,* 4019–4026. http://dx.doi.org/10.1523/ JNEUROSCI.0564-07.2007

Wyvell, C. L., & Berridge, K. C. (2000). Intra-accumbens amphetamine increases the conditioned incentive salience of sucrose reward: Enhancement of reward "wanting" without enhanced "liking" or response reinforcement. *The Journal of Neuroscience, 20,* 8122–8130.

Memory as a Hologram: An Analysis of Learning and Recall

Donald R. J. Franklin and D. J. K. Mewhort
Queen's University

We present a holographic theory of human memory. According to the theory, a subject's vocabulary resides in a dynamic distributed representation—a hologram. Studying or recalling a word alters both the existing representation of that word in the hologram and all words associated with it. Recall is always prompted by a recall cue (either a start instruction or the word just recalled). Order of report is a joint function of the item and associative information residing in the hologram at the time the report is made. We apply the model to archival data involving simple free recall, learning in multitrial free recall, simple serial recall, and learning in multitrial serial recall. The model captures accuracy and order of report in both free and serial recall. It also captures learning and subjective organisation in multitrial free recall. We offer the model as an alternative to the short- and long-term account of memory postulated in the modal model.

Keywords: hologram, memory, learning, recall

If you were asked whether you know the meaning of the word "apple," you are likely to be able to affirm your knowledge in less than a second. Likewise, if you were asked whether you know a fact that you do not know (e.g., Margaret Trudeau's maiden name), you are likely to confirm that you do not know within a comparable time (see Glucksberg & McCloskey, 1981). In both examples, you have to search what you do know. If search involved an item-by-item process, quick confirmation of both what you know and what you do not know is hard to explain—if search were serial, instead of reading the remainder of this sentence, you would still be lost in thought, seeking Margaret Trudeau's maiden name. Given that you are not lost in an exhaustive item-by-item search, you must have used a parallel search, likely with a content-addressing mechanism.

In this article, we present an account of memory as a hologram—a method of data storage that supports a content-addressable search. To present the model, we will focus on learning and ordered recall. How we order recall is a fundamental problem because behaviour is extended in time. Because verbal report is necessarily a serial process, subjects must order their responses, even though multiple potential responses are available in memory.

Analysis of how subjects order behaviour has been a longstanding topic of controversy (e.g., Lashley, 1951). Order of recall must depend on the words stored in memory, on associations among them, and on information about the words and the context in which they were stored. Particularly controversial has been the balance of two contributing sources of information: item-to-item associations and item-to-context associations. The latter refers to information about the words, such as their position in a spatial or temporal stream (e.g., Dennis, 2009).

Ebbinghaus's (1885/1913) account was based on information taken from the list; to order report, it depended on associative links among items, particularly item-to-item links. Young (1962) used a transfer paradigm to argue against the associative-link idea; a rebuttal by Johnson (1975) showed that Young's analysis was fatally flawed. Fifteen years later, Lewandowsky and Murdock (1989) revived interest in the associative-link idea. In their account, memory was a holographic store. At the start of each trial, memory was empty, and as subjects studied a list of words, the hologram stored the studied items and the pairwise associations among adjacent items. During recall, subjects used the associative chain to drive report. Mewhort and Popham (1991) used the same associative-chain ideas to simulate report of tachistoscopically presented letter strings under conditions of masking and letter spacing; that is, they exploited pairwise associations to handle left-to-right scanning (see Mewhort & Campbell, 1981).

In the last two decades, however, a wealth of data has surfaced that challenge the item-to-item chaining mechanism. Instead, current theorists focus on mechanisms based on context, specifically on item-to-context associations. As we will document later, our account does not use context-to-item associations (either temporal or spatial) to account for simple list-learning paradigms, but we acknowledge that subjects use such information in more complex situations. Indeed, one of the desirable characteristics of our holographic model is the ability to combine sources of information, in particular, item-to-item information and item-to-context information.

Bryden's (1967) account is likely the earliest of modern context-based theory. Based on tasks in which space and time were put in conflict (e.g., dichotic-listening or split-span tasks), he proposed that order of report reflected ranking of items in space or by time, and he provided a neurophysiological model to explain the ranking. Unfortunately, although the model could account for either spatial or temporal organisation, it was unable to explain the switch from a temporal to spatial dimension (see also Mewhort, 1973, 1974).

Although more recent models generally ignore examples involving spatial–temporal conflict, they agree that order of report is based on item-to-context information. Unfortunately, the nature of the context is controversial. Alternative suggestions include time, temporal position within a list, list position, and position from the ends of the list (see Botvinick & Plaut, 2006; Brown, Neath, & Chater, 2007; Brown, Preece, & Hulme, 2000; Burgess & Hitch, 1999; Farrell & Lewandowsky, 2002, 2012; Grossberg & Pearson, 2008; Henson, Norris, Page, & Baddeley, 1996; Lewandowsky & Farrell, 2008).

Curiously, because item-to-item associations cannot do the full job, most accounts based on item-to-context information deny that item-to-item associations are useful at all (e.g., Dennis, 2009). Our view is that some item-to-item associations surely exist, and if so, they are likely to be put to use. It is equally curious that current accounts conspicuously ignore subjects' prior knowledge (i.e., their semantic memory or lexicon).

In this article, we propose a holographic account of memory. The account acknowledges subjects' prior knowledge of words by storing vectors that represent words in a holographic lexicon. The lexicon consists of a single vector that is a composite of all words and relations among words. The model postulates that feedback during study and report alters the strength of all items (words and their associates) in semantic memory. Further, recall of a word is based on all information in semantic memory as the word is considered for report. Hence, we agree with theorists who argue that item-to-item associations are insufficient to account for recall order (e.g., Dennis, 2009). Instead of item-to-context associations, order of report in the holographic account is based on the mechanisms of storage and retrieval.

The Theory

The lexicon (also known as semantic memory) is at the heart of the theory. It contains a representation of all words and associations among words in the model's vocabulary. The lexicon is based on the mathematics of light holography and exploits Gabor's (1968, 1969) demonstration that memory systems based on vector convolution mimic a hologram (see also Longuet-Higgins, 1968; Murdock, 1982; Poggio, 1973).

Because it is a long-term memory (LTM), the lexicon would be treated conventionally as a stable store. Discussing the Atkinson-Shiffrin model, for example, Shiffrin (1999, p. 20) noted that "the primary structural distinction in the memory system is between the active memories (all the short-term and sensory stores) and the passive memory (long-term store)." By contrast, we treat the lexicon as a dynamic store: Changing the strength of any word or association increases the strength of similar objects and decreases the strength of dissimilar ones. Our treatment of the lexicon as a dynamic store reflects the properties of the holographic memory system in which we have implemented it.

During the study of a list of words, presentation of an item not only strengthens its representation in the lexicon but also reduces the strength of the preceding studied items. The reduction is an interference effect (overwriting); the interference is a primary contributor to recency effects. In addition, through rehearsal, subjects create interitem associations and add them to the lexicon. Adding items to the lexicon changes the strength of the corresponding representations already present in the lexicon.

The reduction in the strength of preceding studied items is often described in structural terms implying a two-store mix that allows the subject to sample from one store or the other. The sampling idea is motivated by shape of the accuracy profile across the list—the serial-position curve—in particular, the relative size of the primacy and recency components of the accuracy profile. The structural view assumes that subjects report words from the recency end of the list from short-term memory (STM) and then pick up items from the primacy end of the list from a second less labile store (e.g., Craik, 1970; Waugh & Norman, 1965). In effect, the structural view assumes that the reduction in the strength of initially presented items is a step function as the subject switches from one store to the other—a view that has been criticised on both logical and empirical grounds (Gruneburg, 1970; Melton, 1963). Instead of a structural account, we describe the reduction in terms of feedback to the lexicon (overwriting); feedback provides a continuous reduction in the strength of previously studied items. Here, we echo the view urged by Ebbinghaus: "The earlier images are more and more overlaid, so to speak, and covered by the later ones. Therefore, in the case of the earlier images, the possibility of recurrence offers itself more rarely and with greater difficulty" (Ebbinghaus, 1885/1913, Chapter 7, Section 26).

The ability to form item-to-item associations is limited by the effort involved. Rundus and Atkinson's (1970) data illustrate how the limitation works. They asked subjects to rehearse out loud and found the number of rehearsals of each word fell off exponentially with position in the list. Figure 1 shows their data (in closed circles). In addition, we fit the data to a geometric function (the discrete form of an exponential). The fitted geometric function is shown in open circles. As is clear in the figure, the geometric function fits their data extremely well.

During recall, subjects use all information, both item and associative, to select successive responses. Item and associative information are obtained from all lexical entries and, when combined, yield an overall measure of strength for each entry. The entry whose strength is both closest to a criterion and within defined boundaries is selected for report. Recall halts when no item is within the bounds and all restart options have been exhausted. Feedback from report affects all lexical entries.

Because the state of the lexicon controls recall, the theory anticipates trial-to-trial interactions. Task differences affect the way retrieval is implemented. In serial recall, for example, associative information is given priority initially when selecting an item for recall. When recall halts for the first time, however, selection is based on a different criterion—one based on item information.
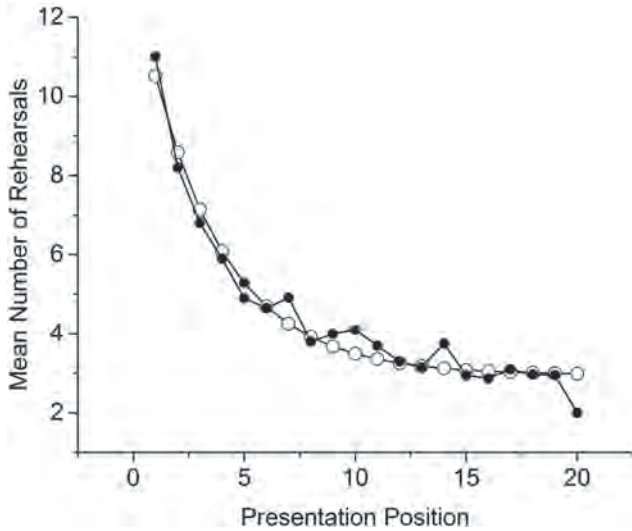
Figure 1. Number of overt rehearsals as a function of presentation position. Data (closed circles) from Rundus and Atkinson (1970) fit to a geometric function (open circles).

Feedback During Study and Retrieval Alters the Lexicon

When a subject studies or reports a word, its strength in the lexicon is changed. Changing one item, or association, also changes the strength of the other material in the lexicon. The way that particular items and associations are treated depends on the procedure involved in the task. For a simple free-recall trial, for example, changes are introduced by virtue of studying and reporting the list of items presented on the trial. For a learning task, by contrast, additional changes are introduced by virtue of the repetition of the materials. Likewise, if an additional constraint is imposed, as in a serial-recall task, the emphasis given to item and associative information is changed accordingly. We will discuss the implementation of each paradigm separately.

Learning depends on both item and associative feedback to the lexicon. When subjects see a list of words for the first time, feedback is based on those items recalled. If the subject sees the same material again, feedback is based additionally on the current strength of the material in memory. As a result, feedback is moderated by information retained from one trial to another.

Implementing the Theory

Representation assumptions. For the simulations that follow, each word is represented by a vector of 2,048 values; each value is derived independently by sampling from a Gaussian distribution with a mean of 0 and variance of 1/2,048. The dimensionality of 2,048 provides the system with appropriate resolution for the vectors.

Vectors created in this fashion are often described as *orthonormal in expectation*. That is, the dot-product of each vector with itself is approximately 1.0, and the dot-product of two different arbitrarily chosen vectors is approximately 0.0. As a result, the dot-product—instead of the normalized dot-product (the vector cosine)—can be used to measure the similarity of any two vectors.

An association between each pair of words is constructed by computing the outer-product of the pair. The outer-product is compressed to a vector of the same dimensionality as the dimensionality of the word vectors using circular convolution ($*$), that is, $\mathbf{z} = \mathbf{x} * \mathbf{y}.$

Figure 2 illustrates the arithmetic of circular convolution: The top panel shows two vectors, $\mathbf{x}$ and $\mathbf{y}$; $\mathbf{x}$ is shown as a standard column vector; $\mathbf{y}$ is transposed, $(\mathbf{y}^{\mathbf{T}})$ to a row vector.

The left side of the middle panel shows the outer-product matrix produced by multiplying the two vectors, that is, $\mathbf{x} \times \mathbf{y}^{\mathbf{T}}$. As shown, the outer-product matrix is constructed by computing three rows (one corresponding to each of the rows in $\mathbf{x}$). The elements in each row are computed by multiplying the row element in $\mathbf{x}$ by each of the columns in $\mathbf{y}$.

The right side of the middle panel shows a Latin square. The values in the square index values from the outer-product matrix that must be summed to form the convolved vector $\mathbf{z}$. The bottom panel sums the relevant values to form each entry of the convolved vector $\mathbf{z}$. Circular convolution is commutative, that is, $(\mathbf{x} * \mathbf{y}) = (\mathbf{y} * \mathbf{x})$. Plate (2003) and Kelly, Blostein, and Mewhort (2013) discuss methods by which it can be made noncommutative.

Circular correlation (#) is an approximate inverse of circular convolution. Figure 3 illustrates the arithmetic: The top panel shows two vectors $\mathbf{x}^{\mathbf{T}}$ and $\mathbf{z}$. Vector $\mathbf{z}$ is the convolution of $\mathbf{x}$ and $\mathbf{y}$ from Figure 2. The middle panel of Figure 3 shows their outer product. The bottom panel of Figure 3 shows a facsimile vector $(\mathbf{y}')$ constructed by forming the correlation $(\mathbf{x} \# \mathbf{z})$. Although $\mathbf{y}'$ is
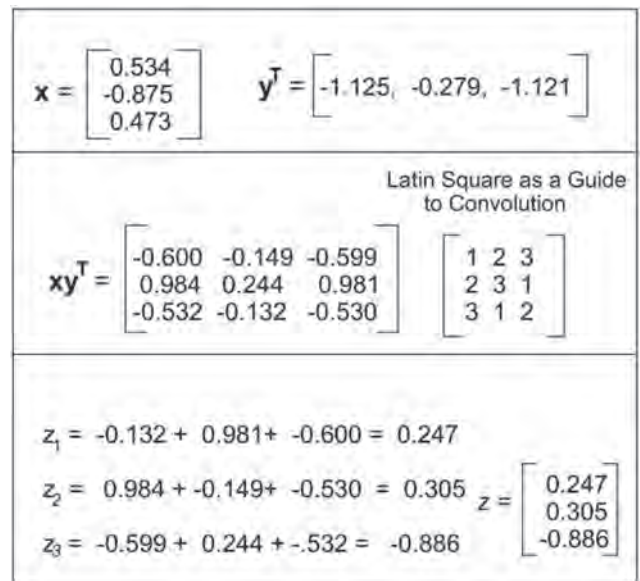
Figure 2. The calculations for circular convolution. The top panel shows two vectors, $\mathbf{x}$ and $\mathbf{y}^{\mathbf{T}}$, to be convolved. The middle panel shows their outer-product matrix and a Latin square indexing the values summed to collapse the matrix. The bottom panel sums the components of the out-product matrix and shows the resulting vector $\mathbf{z}$.

Figure 3. The calculations for circular correlation. The top panel shows two vectors, $\mathbf{x}^T$ and $\mathbf{z}$ (from Figure 2), to be correlated. The middle panel shows their outer-product matrix, and a Latin square indexing the values summed to collapse the matrix. The bottom panel sums the components of the out-product matrix and shows the resulting vector $\mathbf{y}'$.

not identical to $\mathbf{y}$, it resembles $\mathbf{y}$ as indicated by its cosine with $\mathbf{y}$ (0.93).[1]

Our representation assumptions are fundamentally the same as used by Lewandowsky and Murdock (1989; see also Franklin, 2013; Franklin & Mewhort, 2002). Like Lewandowsky and Murdock, we use a hologram for memory. In their simulations, the hologram is empty at the start of each trial—they treat the hologram as a store for *studied* items (see also Murdock, 1982). Unlike Lewandowsky and Murdock, we treat the hologram as a permanent store that holds the *whole of the subject's lexicon*—the subject's vocabulary and interword associations. It is not empty at the start of each study trial.[2]

Johns and Jones (2010) have questioned the use of random vectors as exemplars in models of memory on the grounds that random vectors typically create a population of exemplars that is too Gaussian in shape. The too-Gaussian problem cannot be solved easily because the use of structured vectors raises other complications. Kelly et al. (2013) have suggested that structure in the Latin square (used in convolution and correlation) can be confounded with structure in the vectors themselves. Of course, the confounding does not occur with random vectors because, by definition, random vectors lack structure. Kelly et al.'s work motivated us to stick with tradition and to use random vectors rather than structured vectors (such as those produced by the Bound Encoding of the Aggregate Language Environment (BEAGLE) model or by Latent Semantic Analysis (LSA), see Jones, Kintsch, & Mewhort, 2006; Jones & Mewhort, 2007; Landauer & Dumais, 1997, respectively). Although we use random vectors to represent words, we do not attach semantics to particular exemplars. The features in the

random vectors might represent descriptions of the words or brain states that represent the words. For the purposes of our demonstrations, the referent for the features can be left vague. The interactive nature of the vector-holographic architecture illustrated in the model is its critical contribution.

Some properties of a holographic store. A hologram, like human memory, is a distributed representation that is robust to loss of medium. To illustrate the property, we constructed a hologram by superimposing 2,000 item vectors of dimensionality 2,048 and the 1,999,000 (C[2000, 2]) word-to-word associations, each weighted by a constant (0.01). We added five lexical items selected at random to the hologram at strength 0.8. As a baseline, we measured the strength for each of the five items in the hologram. Armed with the baseline measurement, we replaced 64, 128, 256, 512, or 1,024 elements of the hologram at random with a value of 0.0. Strength is measured using the dot-product of each of the five items with the lexicon. At each step, we assessed the strength of the critical five items in the hologram.

We repeated the procedure (i.e., we built a fresh hologram, selected five items at random, and lesioned the hologram) 100 times to obtain 500 data points for each percentage of memory destroyed. The data, shown in Figure 4, are means of the 500 data points. As shown in Figure 4, as more of the hologram is destroyed, the strength of the representation of the critical five items is reduced, but even when 50% of the values in the hologram have been set to zero, the strength of the critical five items remains high.

Second, holographic memories are dynamic. Adding or strengthening a word stored in memory changes the strength of all words; the change (either an increase or a decrease) depends on the similarity of each item to the item added. To illustrate the dynamic property, we constructed a hologram of 100 items, each of dimensionality 200. All items, weighted by 0.01, were added to the hologram.

Figure 5 shows the strength (measured with the dot-product) of the 100 items. The top panel (left side) shows the strength of the items immediately after all items had been added to the hologram. As is shown, the strengths hover around the nominal strength of 0.01. The top panel (right side) shows the corresponding strengths after the first item was added again with a weight of 0.35. Again, the strengths hover around the nominal 0.01, but there is considerably greater variability. The lower panels show a corresponding increase in variance as the final two items are added to the hologram.

The changes in strength shown in Figure 5 are systematic. Figure 6 replots the change in item strength for all memory items using similarity of all items in the hologram to the just-added item on the abscissa.

As shown in Figure 6, the change in strength of any item depends upon its similarity to the item just readded and on the

---

[1] We have described circular convolution and correlation in terms of the outer-product matrix and the calculations used to compress it. In practice, because it is faster, we compute convolution vectors using an algorithm based on the Fourier transform (the routine CO6EKF from the Numerical Algorithms Group, www.nag.com).

[2] Whether or not memory should be empty on each trial has become a contentious issue (e.g., Murdock & Kahana, 1993 vs. Shiffrin, Ratcliff, Murnane, & Nobel, 1993). Our use of a holographic lexicon assumes that memory always contains knowledge.
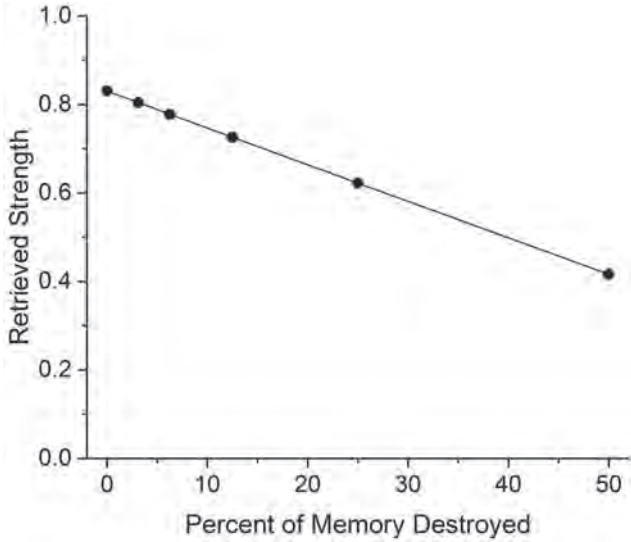
Figure 4. Retrieval strength as a function of the percentage of memory lesioned.

weight at which the item was readded; the slope of the change function equals the weight of the item readded. Figures 5 and 6 illustrate the interactive nature of the holographic store. Changing the strength of an item in memory affects the strength of all other items. The changes resemble the excitatory part of a spreading-activation model (e.g., Collins & Loftus, 1975), without the hydraulics necessary to make such a model work. But, as shown, the changes in the hologram are inhibitory as well.

In the sections to follow, we will apply the model to archival data from free-recall, learning, and serial-recall paradigms to demonstrate that the model can capture the fundamental characteristics of performance in the paradigms. Given that the model is able to capture data from the several paradigms, we offer it as a new approach to the problem of ordering recall.

Free-Recall Paradigms

In a typical free-recall trial, subjects are given a list of words for study and are subsequently asked to recall as many words as possible in any order. When a subject studies a word, its representation is strengthened in the lexicon. The extent of the change depends on the position of the word in the study list. On each trial, the model adds a start item that signals that the next item is the first item to be studied and an end-of-list item (i.e., a recall cue). The list's length is represented by the symbol LL; hence, for the simulation, the study list includes the list proper (i.e., items 1 to LL) along with items 0 and $(LL + 1)$.

Following Rundus and Atkinson (1970), the efficiency with which adjacent items can be associated falls off as more items are studied; the strength of association for successive pairs of studied items follows a decreasing geometric function of list position. To

Figure 5. Retrieved strength of 100 items as some are strengthened in a lexicon.

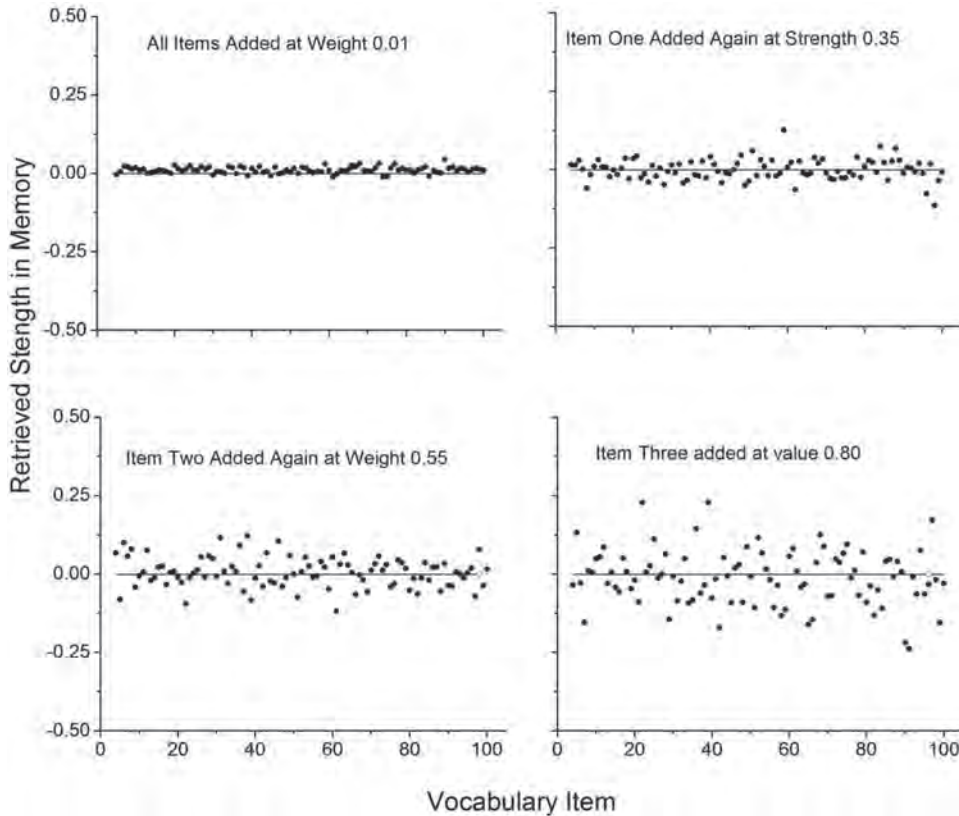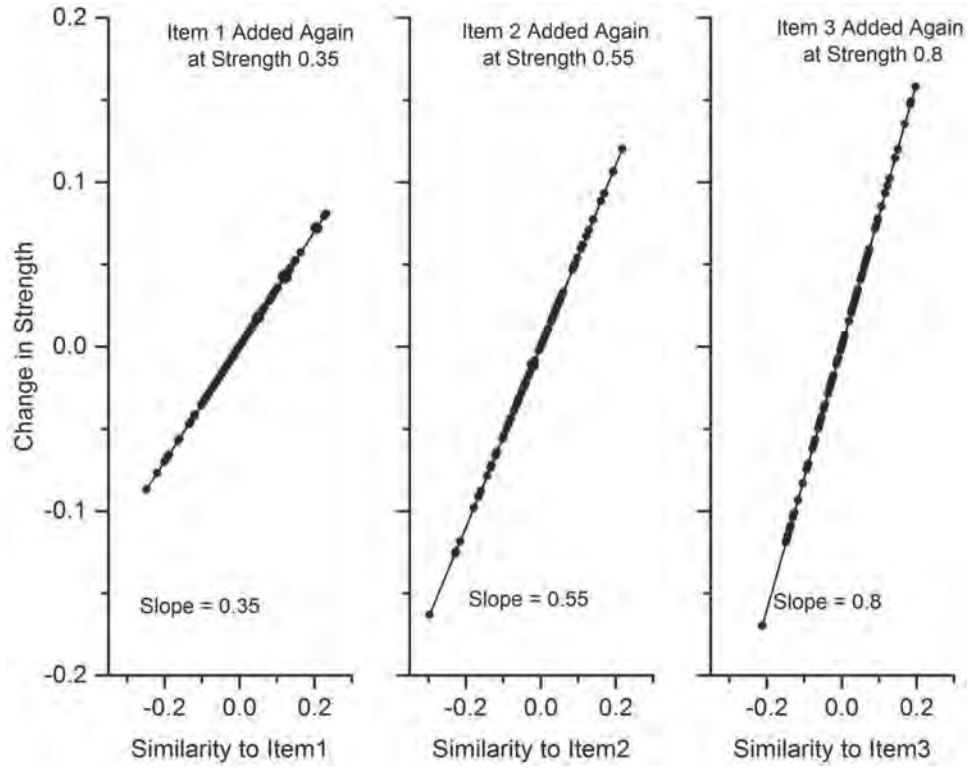*Figure 6.* Change in retrieval strength as a function of similarity of the item added to each newly strengthened item.

implement the decrease in efficiency, we set the strength of the association between adjacent pairs of items using a geometric function starting from a context item that precedes the list. The values of associative strength, ω_i, are computed as follows:

$$\omega_i = \omega_0 \times \lambda^{(i-1)}$$

where ω_0 is the maximum strength of association, λ is a scaling parameter, and i runs from 1 to $(LL + 1)$.

Studying an item not only reinforces its existing strength in the lexicon but also interferes with items studied earlier in the list. One can anticipate each item's strength by using a decreasing geometric series back from the last item presented. The strength, γ_i, of the word in the ith serial position in the study list is

$$\gamma_i = \beta + \gamma_0 \times \theta^P$$

where γ_0 is the maximum value of any studied item, θ is a scaling parameter that represents the proportion of strength retained, β is the minimum item weight (fixed at 0.35), and P is a counter that increases from 0 to (LL) as i decreases from $(LL + 1)$ to 1.

Initial report is prompted by the end-of-list item $(LL + 1)$, and the word reported serves as the cue for the next report. When a word is recalled (either correctly or in error), a proportion of the reported item and a proportion of the association between the reported item and its report cue, is fed back to the lexicon. The feedback is calculated

$$\text{Item}: \left(1 - \left[\mathbf{r} \bullet \mathbf{L}\right]\right) \times \eta,$$

$$\text{Associative}: 1 - \left(\left[\mathbf{p} * \mathbf{r}\right] \bullet \mathbf{L}\right) \times \eta$$

where $\mathbf{r}$ is the recalled word, $\mathbf{L}$ is the lexicon, $\mathbf{p}$ is the report cue, and η is a scaling parameter set to 0.063 (a value derived from pilot studies).

The choice of the end-of-list item to prompt the initial report deserves comment. Helstrup (1984) has shown that subjects in a free-recall paradigm try initially to treat the task as a serial-recall task. As a result, they start report with the first presented item, and performance on the initial items is high. Over the first few practice trials, they soon learn that, for a long list, their serial strategy leaves the final items from the list unreported. Accordingly, they change strategies and begin to recall from the end of the list. With further experience in the task, they make fine adjustments to their start-position strategy. We will illustrate the adjustments using Murdock's (1962) data when we apply the model to the free-recall paradigm.

Of course, with short lists, subjects may have little incentive to change strategies because the final items remain available. Therefore, the subjects stick with a serial strategy. For such cases, the model would use Item 0 as the initial report cue. Ward, Tan, and Grenfell-Essam (2010) and Corballis (1967) have described such cases.

To decide which item to report, the model computes the strength of all items in the lexicon, and the strength of the association of the probe with the lexicon. The decision is based on the sum of item and associative information in the lexicon. Hence, the first step in

retrieval requires calculating the strength with which each vocabulary item is associated with the probe. To compute the associative strength values, we first obtain a composite of the items in the lexicon in which each item is represented roughly proportional to the strength of its association with the probe. To obtain the composite, $\mathbf{f}$, we compute

$$\mathbf{f} = \mathbf{r} \,\#\, \mathbf{L}$$

where $\mathbf{f}$, the composite, contains information (a) about all vocabulary items that have been associated with the probe, and (b) about items that are similar to the vocabulary items that have been associated with the probe. We assess the similarity of the composite to each item in the lexicon (i.e., to each vocabulary item, $\mathbf{v_i}$). To assess the similarity of the $\mathbf{i}$th item to $\mathbf{f}$, we compute $S(\mathbf{v_i})$, a scalar,

$$S(\mathbf{v_i}) = \mathbf{f} \bullet \mathbf{v_i}$$

where $\mathbf{i} = 0$ to 2,000.

A vocabulary item's momentary strength is defined as the sum of its current strength plus the strength contributed by the associative information. Each word's current strength is obtained by computing $\mathbf{v_i} \bullet \mathbf{L}$. The momentary strength for the $\mathbf{i}$th vocabulary item is $S(\mathbf{v_i}) + \mathbf{v_i} \bullet \mathbf{L}$, where, as before, $\mathbf{i} = 0$ to 2,000.

The word with a momentary strength closest to criterion is selected for recall, provided the momentary strength falls within limits. In the current simulations, the criterion was set at 1.0 (the power of the item vectors), with limits set at ± 0.5. Note that the selection rule specifies report of the item closest to 1.0, not the item with the highest momentary strength. Recall halts when no candidate has an appropriate momentary strength.

The selection rule picks the word with strength closest to criterion. The rationale for the rule is that a neural pathway includes neurons with variance in activation. If we think of a vector as representing a neural pathway, the representation and storage mechanisms of the holographic system we describe shares the properties of variance in representation. We think of the criterion as the mean of the distribution of variances so that the rule specifies the item closest to the mean of the pathway.

To show the model in action, we applied it to standard experiments. Our applications are not comprehensive—it is likely impossible to be comprehensive in a single article. Our intent is to show that the model can accommodate classic phenomena, especially those thought to support alternative conceptions and accommodate phenomena designed to illuminate particular interpretations of current interest.

Before introducing the demonstrations, however, we offer a cautionary note. Any manipulation is analytic only within the theoretical framework in which it has been advanced. Most manipulations will constrain performance, but if the theoretical framework is flawed, the constraints will not reflect the issues that they were intended to illuminate, and the manipulations will not have the theoretical force they were designed to have. Hence, our strategy in testing the model is to show that it responds to the constraints in the same way that subjects do, but we reserve our interpretation of the results to issues defined within our model.

For the demonstrations to follow, the lexicon was constructed by summing the word vectors and the pairwise associations among the words. We used a vocabulary of 2,000 words and their 1,999,000 associations (C[2000, 2] = 1,999,000). Each entry to

the lexicon was weighted by a scalar, 0.001. Because subjects have differing vocabularies, we constructed a fresh lexicon for each subject.

Single-Trial Free Recall

A sample free-recall trial. In a free-recall trial, subjects study a list of words and are subsequently asked to recall as many words as possible. Figure 7 tracks a sample trial to illustrate how the recall mechanism works. The ordinate is the momentary strength for each of 85 words taken from a lexicon of 2,000. We restricted the figure to 85 words to avoid excessive clutter. The studied words are shown as the leftmost 12 items in each panel. Each panel (numbered 1 to 6) shows the strength of each of the 85 words on a report-by-report basis. As is shown, Word 12 was the first item to be reported. The next report was Word 7, followed by Word 20 (an intrusion). Reporting continues until none of 2,000 words' momentary activation is close enough to criterion (1 ± 0.5).

Two features of the example deserve emphasis. First, reporting does not stop because a response is an intrusion rather than a studied item. It is possible (but extremely unlikely) for the model to produce a series of intrusions without reporting a studied item. Second, the model has a natural stopping rule: reporting stops when none of the 2,000 words in the lexicon have a momentary strength close enough to the criterion of (1 ± 0.5).

Murdock (1962). In his classic book, *Models of human memory*, Norman (1970) reprinted the serial-position curve from Murdock (1962) as an example of stable data worthy of theoretical attention. Fortunately, a complete record of Murdock's data can be downloaded from the Computational Memory Lab at University of Pennsylvania (http://memory.psych.upenn.edu/DataArchive). We start our analysis by fitting the model to Murdock's data.

Murdock (1962) asked subjects to study lists of 10, 15, 20, 30, or 40 different words drawn from the Toronto Noun Pool. Each word was presented orally at a rate of 1s or 2 s per word. After the list had been presented, the subjects were invited to report the words in any order. As we noted earlier, subjects adjust their report strategy as they practice the task. An analysis of Murdock's data illustrates the adjustments. Figure 8 shows mean output position (position in the report) as a function of presentation position in 10-trial blocks. Initially (top left panel) subjects start report with the last list item and work backward before switching to report of the initial list items. The asymptotic items come later in the report stream. By Trial 35 (bottom left panel), subjects report the fourth-from-last item first and work forward. The change is shown clearly in the hook that appears in the figure at the last four presentation positions. By Trial 70 (bottom right panel), the hook has deepened.

Early authorities suggested that the change in report documented by the hook does not occur with lists presented visually (e.g., Beaman & Morton, 2000; Kahana, 1996), but they either confounded modality with amount of practice or administered too little practice to allow the hook to develop. In a large experiment, Roberts (1972) compared both visual and auditory presentation (across both list length and rate of presentation), and his data exhibit the hook in both modalities. The change in report order is not a modality effect; as we have suggested, it is a strategic shift that reflects subjects' reflection on performance for end-of-list items (although there may be other reasons as well).
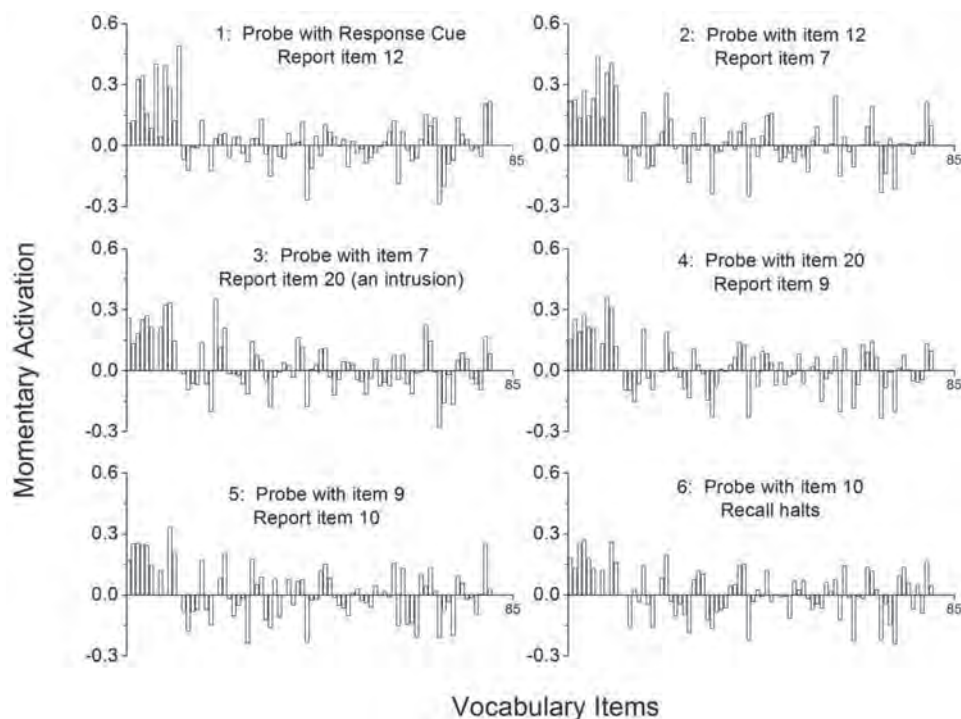
Figure 7. Momentary activation for 80 of 2,000 words across successive reports. The first 12 words were studied; the remaining items were unstudied but were included in the lexicon.

Figure 9 corroborates the change in report order using a different measure, the conditional response probability (CRP) curve, a measure of the probability of recalling item *a* immediately followed by item **b**, conditioned on the recall of **b** (see Kahana, 1996). The CRP is plotted on the ordinate, with the distance between presented items on the abscissa. As subjects' change strategy across trials, the tendency to report in forward order increases.

The change in report order across trials illustrates a problem with averaging data across trials. Models that do not include a mechanism sensitive to the strategic shift in report should not be applied across a shift in strategy. Because the holographic model does not include a shift mechanism, we simulated performance in Murdock's (1962) experiment where the empirical order of report was stable. The decision to use trials with a stable order of report is a conservative one pending empirical results to point the way to the kind of strategic mechanism needed by the model.

We applied the model to the first 10 trials of Murdock's (1962) 20-item list presented at a rate of one word per second. On each trial of the simulation, the model studied 20 words. Recall was prompted using the end-of-list marker (word $LL + 1$). Recall continued until none of the 2,000 vocabulary items had a momentary strength close enough to criterion. To honour the idea that, in a free-recall task, subjects try to use all information at their disposal, when report halted for the first time, the model attempted to restart by using the beginning-of-the-list marker (Word 0) as a probe.

We fit the model to the serial-position curve (smoothed so that the fit was less perturbed by noise) using a downhill simplex optimization algorithm (see Press, Teukolsky, Vetterling, & Flannery, 1992, pp.

402–406). Once we had obtained the fitted parameters, we ran the model 20 times independently (using the fitted parameters) and recorded the range of values produced. The procedure of fitting first and then running 20 times was designed to reveal the variance inherent in the model's behaviour as a result of its use of Gaussian vectors to represent words.

Figure 10 presents the results of the simulation. The filled symbols connected with a solid line shows Murdock's (1962) data. The vertical bars show the range of values produced by the model over the 20 independent runs. Each run included 15 simulated subjects, each with their own lexicon. As is clear in the figure, the model captures the shape of the serial-position curve.

The serial-position curve shows the number of items recalled correctly as a function of presentation position. It is an aggregate function: Because each trial allows only one word at each stimulus position, no subject would produce the serial-position curve like that in Figure 10 on any single trial. Hence, the serial-position curve is an artifact produced by averaging over trials. Nevertheless, the serial-position curve has been an important analysis from the beginning of experimental work in psychology. Nipher (1878), who published the first known scientific experiment in memory, calculated serial-position curves, as did Ebbinghaus (1885/1913). Subsequently, the majority of published list-learning experiments include the serial-position curve as part of the analysis. The historical importance of the serial-position curve, and its popularity as an analytical tool, make it an excellent choice as a measure to assess the model's fitting ability.

The shape of the serial-position curve suggests that subjects prefer to report the most recent items first. That preference is often interpreted in terms of a multiple-store theory (e.g., Craik, 1970;
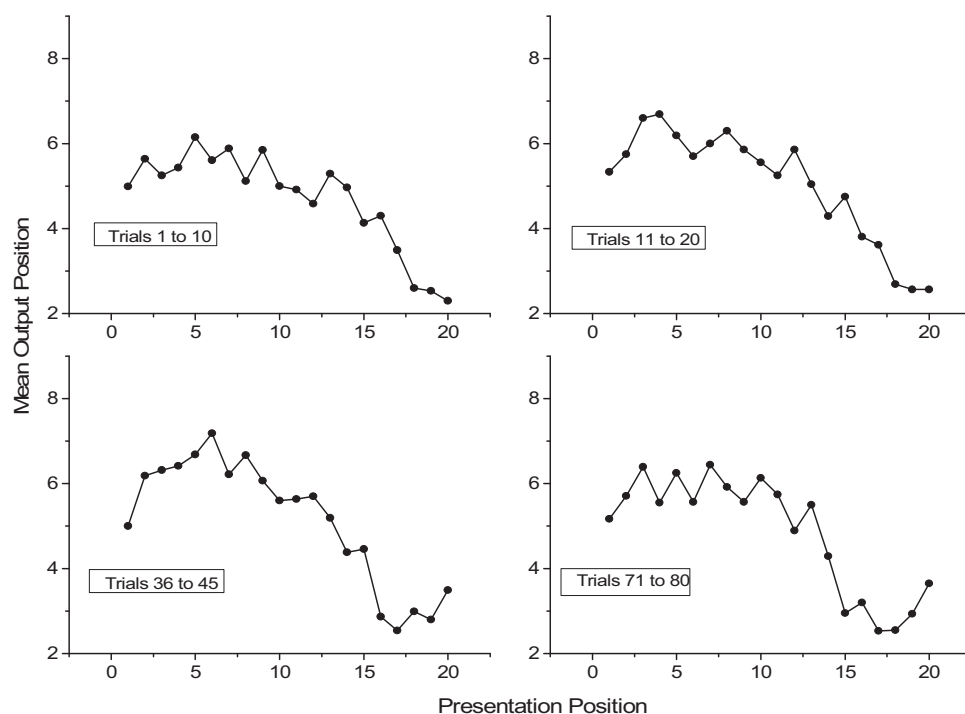
Figure 8. Mean output position as function of presentation position for blocks of 10 trials in ascending order. The data are drawn from Murdock (1962).

Waugh & Norman, 1965). We shall say more about the viability of such theory later. At this point, it is worth noting that our account is based on a single-store—the subject's lexicon.

Because it is an aggregate function, the serial-position curve gives us limited information regarding how recall may have varied across individual trials. To get a more complete account of recall, we need to supplement the basic serial-position curve by examining order of report and other basic data such as the number of items reported per trial.

The next analysis examines the number of words reported correctly per trial. We tallied the number words reported correctly per trial for each of Murdock's (1962) subjects, and calculated a frequency distribution pooled across subjects. For comparison, we computed the corresponding frequency distribution pooled across the 20 runs of the model described in Figure 10.

Figure 11 presents the results of the frequency analysis of words correct per trial for both the simulated data and the human data. Closed circles represent the frequency distribution drawn from the Murdock (1962). Open circles represent the frequency distribution of the simulated data. As is evident in Figure 11, the 15 simulated subjects (pooled across 20 runs) are a close match to human subjects on this measure. Both the model and Murdock's subjects reported approximately the same distribution of words correctly per trial.

It is important to note that the curve in Figure 11 was not a fitted curve. Rather, we used the data produced from the 20 runs using parameters from the simulated serial-position curve. The distribution of words correct per trial comes for free without refitting.

The free-recall paradigm allows subjects to order their responses without constraint, but, of course, subjects do not report in random order. Presumably, the way that they constrain their report reflects

the mechanisms used to order recall. Because of its potential as a window on the nature of the mechanisms, several measures have been advanced to describe order of report.

Figure 12 shows one measure: the position in the output stream as a function of presentation position (the same measure we used in Figure 8). The solid line presents Murdock's (1962) data, and the vertical bars show the range of data from the model from the 20 runs shown in Figures 10 and 11. As is shown in Figure 12, the model's function tracks the pattern of the data. Note that items recalled most often, as shown on the serial-position curve, are recalled earliest in the output stream. Items recalled less often tend to be recalled later. The strong negative correlation between the serial-position curve and the output-position curve was also reported by Deese and Kaufman (1957).

Figure 13 shows the CRP function drawn from Murdock's (1962) data in the left-hand panel. The right-hand panel displays the values drawn from the simulated data. As is clear in the figure, the probability of a forward response is much greater than the probability of a backward response for both the subjects and the model. Like the serial-position curve, the CRP function is a composite. The CRP function is complicated by the fact that it is not only a composite but also a conditional composite. For that reason, statistical analysis on the CRP function is often problematic.

The final measure of order of report eschews conditional analysis in favour of a correlation between input position (the serial-position in which the word was presented) and the output position (the position in which it was reported). The measure has been used in earlier work (Deese & Kaufman, 1957; Mewhort, 1974). Here, we used Kendall's tau for the correlation. Hence, if the report were strictly first to last, the correla-
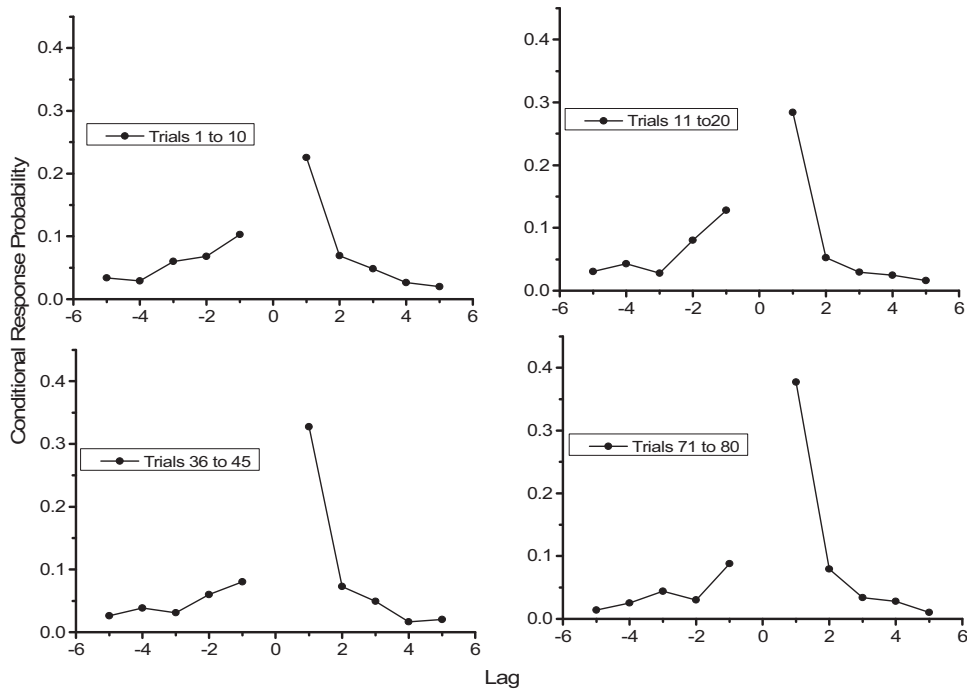
Figure 9. Response probability conditioned on position of the current response for ascending blocks of trials from the same data set used in Figure 8. Forward ordering of report increases with practice in the task.

tion would be $+1$. If it were strictly last to first, the correlation would be -1. Most trials yielded an intermediate value indicating a mixture of orders. If a particular trial involves fewer than two correct responses, tau was set to 0.0.

Figure 10. Accuracy of report as a function of presentation position. The solid line shows Murdock's (1962) data; the vertical bars show the range of simulated values over 20 independent runs of the model.

Figure 14 presents the distribution of correlations for Murdock's (1962) subjects and for the model (across the 20 runs). As is clear in Figure 14, the model captures the distribution well.

Figures 10 through 14 describe the basic characteristics of report in a typical free-recall task. Some of the measures should be interpreted with caution because they are composite scores or are based on conditional analyses. Taken together, however, they provide a clear picture of performance in the task.

It is important to note that we obtained parameters for the model by fitting the serial-position curve shown in Figure 10. The data shown in the remaining figures are based on data from the 20 runs (shown in Figure 10); the 20 runs are, of course, based on the parameters of the original fit. We conclude that the model captures the basic facts of report in free recall—it provides a picture of the topology of report. Table 1 presents the free parameters used in the simulation.

A comment about the parameters is in order. Franklin and Mewhort (2013) documented that the parameters are insensitive to changes in list length provided the word pool, rate of presentation, and subject pool are constant. In addition to the free parameters listed in Table 1, we fixed the bounds around the criterion at 0.5. Increasing the bounds is one way to make the model more flexible; it raises the level of performance while making it easier to fit the model to data.

Both our account and formal implementations of dual-store accounts (e.g., Atkinson & Shiffrin, 1968; Raaijmakers & Shiffrin, 1981) are consistent with the main features of the serial-position data, but the holographic model captures a more complete picture of performance. Neither account is forced by the data. As is often the case, there are at least two theories consistent with the data. We await an empirical test to distinguish the theories.
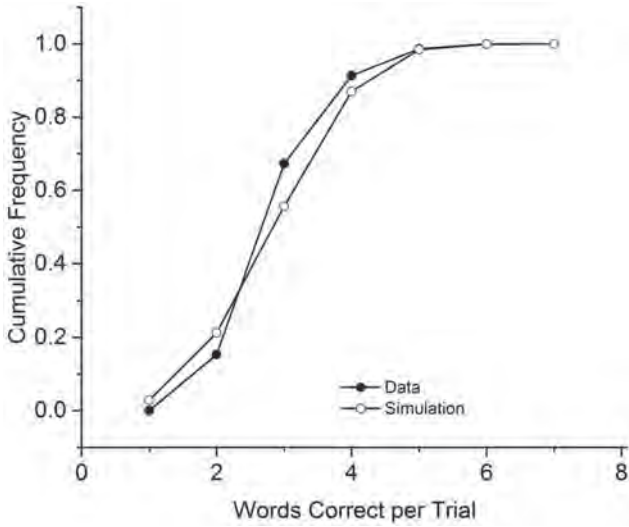
Figure 11. Cumulative proportion of words correct per trial. The closed circles present data from Murdock (1962). The open circles present data from the independent 20 runs of the model derived by fitting the serial position curve in Figure 10.

In the next section, we will apply the model to an extension of the simple free-recall case. The extension was motivated by theoretical concerns of interest at the time. The same concerns do not necessarily motivate our interest in the extension, but for historical
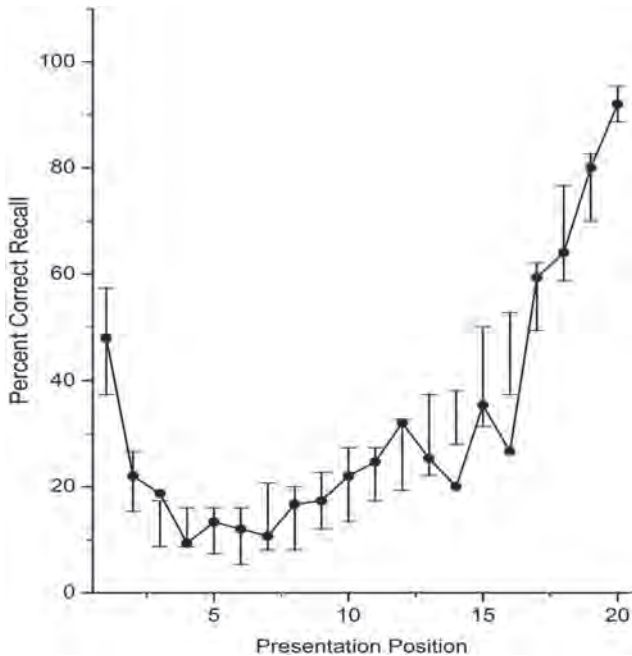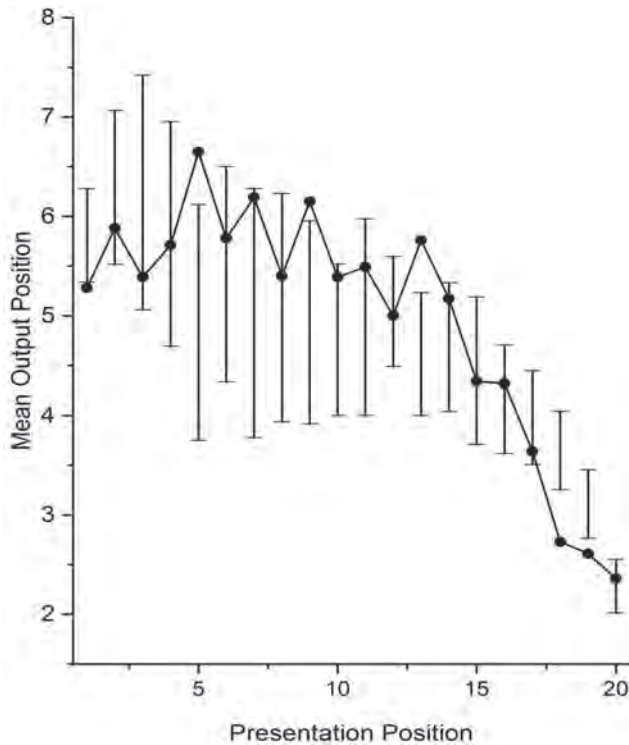

Figure 12. Mean output position as a function of presentation position. The solid line shows Murdock's (1962) data; the vertical bars show the range of 20 independent runs of the model derived by fitting the serial position curve in Figure 10.
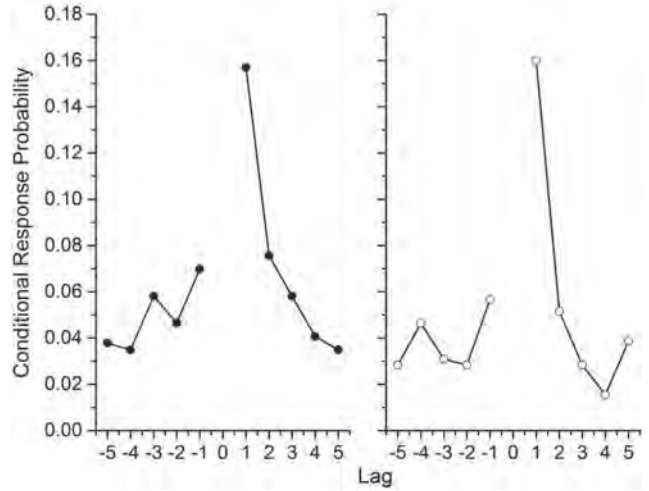
Figure 13. Conditional response probability relative to the position of the current response. The left-hand panel shows Murdock's (1962) data. The right-hand panel represents data drawn from the model averaged across the 20 independent runs derived by fitting the serial position curve in Figure 10.

reasons if none other, it is of interest to see if the model can accommodate the extended situation.

Extensions to Single-Trial Free Recall

Craik (1970) asked subjects to study a series of 15-word lists for immediate report. Recall of each list showed a typical serial-position curve with a small primacy effect and a pronounced recency effect (see Figure 15). After finishing the series of free-recall tests, the subjects were invited to recall all words that they could remember from any of the tests. The serial-position curve
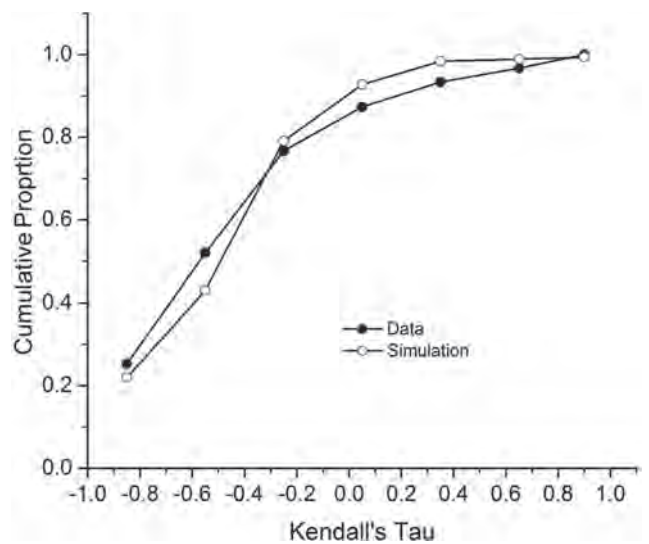
Figure 14. The distribution of tau (order-of-report) scores. The closed circles show the function drawn from Murdock's (1962) data. The open circles represent data drawn from the model derived by fitting the serial position curve in Figure 10.

Table 1
Free Parameters

Simulation	Free parameters[a]					
	ω_0	γ_0	λ	θ	δ	Ǫ
Murdock (1962)	0.140683	0.177325	0.963693	0.804632	—	—
Craik (1970)	0.167997	0.208553	0.887460	0.867606	—	—
Klein et al. (2005) immediate free-recall	0.275530	0.061587	0.980672	0.883129	—	—
Klein et al. (2005) free-recall learning	0.170011	0.162542	0.925955	0.920406	0.03	0.03
Klein et al. (2005) serial-recall learning	0.234218	0.077323	0.970268	0.964993	0.620028	0.000643
Subjective organization	0.190543	0.180603	0.810512	0.805652	0.065028	0.046643

[a] Parameters ω_0, γ_0, λ, and θ are used in single-trial free recall and on the first trial of free-recall learning and serial learning. Subsequent learning trials use only parameters δ and Ǫ.

from the final recall-all test showed a modest primacy effect and a negative recency effect: Report from the final position of the 15-word lists—the positions that had been reported well in immediate recall—were depressed to a low level, hence the designation *negative-recency effect* (see Figure 15, closed symbols). Watkins and Watkins (1974) showed, however, that negative recency becomes a null recency effect if subjects do not know the list length on successive trials. Craik interpreted the results in terms of a two-store structural model, as did Watkins and Watkins.

We ran the model following Craik's (1970) paradigm. We fit the aggregate free-recall data (based on Craik's initial 10 free-recall trials) to obtain parameters for the 15-item lists. Using the parameters obtained from the fit, we administered 10 simple free-recall trials, followed by a test asking for recall of the words from any of the lists. Although list length was constant, the model did not

"know" that fact; we did not try to build in a strategic mechanism to exploit knowledge of list length. Hence, following Watkins and Watkins (1974), our simulation of Craik's data should show a null recency rather than Craik's negative recency effect.

Figure 15 shows Craik's data (left panel) along with the results of the simulation (right panel). As shown in Figure 15 (open symbols), the simulated serial-position curve for the initial simple free-recall trials (averaged over the 10 trials) took the standard form: a small primacy effect and a pronounced recency effect. The simulated serial-position curve for the final recall-all test (open symbols) showed a modest primacy effect and a null recency effect. In short, the simulation captures the key features of Craik's data and is consistent with Watkins and Watkins (1974) results.

Craik's (1970) results are usually interpreted in structural terms. The idea is that reports contributing to the primacy portion of the
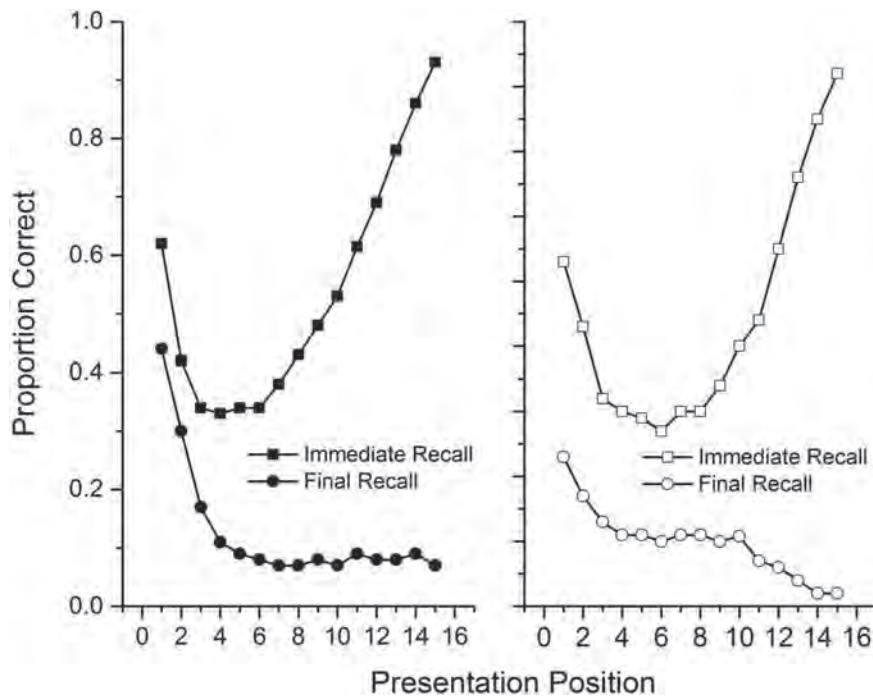

Figure 15. Accuracy as a function of presentation position in Craik's (1970) negative-recency paradigm. The left panel shows Craik's (1970) original data. The right panel shows the model's simulation of Craik's (1970) paradigm.

trial-by-trial serial-position curve are taken from one memory (often called LTM or secondary memory), whereas reports contributing to the recency portion of the curve are taken from a second memory (often called STM or primary memory). From the structural perspective, the negative recency effect occurs because items stored in STM are held only for a brief time, so that when the recall-all test occurs, the words are no longer available.

Glanzer and Cunitz (1966) advanced a similar interpretation for an experiment that used an end-of-list distraction technique to knock words out of STM. They predicted (and found) that the recency component of the serial-position curve would be eliminated. Unfortunately, although their empirical results are consistent with the structural interpretation, they do not force that interpretation. Indeed, Bjork and Whitten (1974; see also Koppenaal & Glanzer, 1990) provided an extensive analysis of the two-memory structural idea and showed that it fails. Neath (1993) manipulated the timing of distraction (both during study and immediately prior to report) and the nature of distraction (i.e., homogeneous or changing). Because of the way distraction affected rehearsal, he also argued against a two-store account. In Bjork and Whitten's words, "The customary two-process theoretical account of immediate free recall is certainly incomplete, if not wrong" (p. 189).

The present simulation reproduces Craik's (1970) contrast between the serial-position curve for immediate recall and the serial-position curve for the report-all condition, but our model does not include the structural distinctions advanced by Craik or by Glanzer and Cunitz (1966). Because the model can reproduce Craik's data without reference to two memory systems, our simulations support the arguments against the two-store idea.

The holographic account explains the negative-recency effect in terms of overwriting the lexicon across successive free-recall trials. Note that although the final free-recall trial was not overwritten by a succeeding free-recall trial, it was overwritten by *instructions to report from any list.* The instructions had the effect of damping the lexicon by the equivalent of a list of about 1.1 items. From the perspective of our model, the idea of interference (overwriting) is correct, but the idea that it is selective to one of two memory systems is not.

Both Craik (1970) and Glanzer and Cunitz (1966) exploited a distraction technique to clean the contents of primary memory. In Craik's case, distraction was provided by successive free-recall trials. In Glanzer and Cuntiz's case, distraction was provided by requiring subjects to count backward before report. Unfortunately, as Neath (1993) demonstrated, distraction is a more complicated operation than they acknowledged. To test the idea of a labile STM—the same issue addressed by Craik and by Glanzer and Cunitz—Hebb (1961) asked subjects to complete a series of immediate serial-recall memory-span tasks. Every third list in the series was repeated, and Hebb found considerable savings from one repeated list to the next. As Hebb noted, the pattern of savings across a complete report of two series denies the idea of a simple labile STM. Clearly, the interference administered by Craik and by Glanzer and Cunitz is more complicated than the simple idea of knocking items out of a STM store. Curiously, neither Craik nor Glanzer and Cunitz cited Hebb's work.

In earlier work, Franklin and Mewhort (2013) applied the current model to other extensions of the free-recall task. For example, the model successfully captures both part-list cuing (Slamecka, 1968; Sloman, Bower, & Rohrer, 1991) and the list-strength effect

(Tulving & Hastie, 1972). The list-strength effect refers to a report advantage conferred on repeated items in a list coupled with a disadvantage conferred on nonrepeated items. The part-list cuing effect refers to a report disadvantage when, after subjects have studied a list, they are told that they do not have to report a subset of the items. In both cases, the model suggests that the phenomena reflect changes in rehearsal leading to changes in order of report induced by the repetition manipulation (list-strength) and the presentation of for-free items (part-list), respectively.

Single-Trial Serial Recall

In a simple serial-recall task, subjects are asked to study a list of words and to report them in order, from first to last. Hence, unlike the unconstrained case of free recall, serial recall constrains the subject's order of report. To respond to the extra constraint, the model changes several details concerning how report is assembled.

First, because subjects are required to report from first to last, the model uses the begin-study cue (Item 0) as the initial response cue. As in free recall, any response produced is used as a cue for the next response; hence, the early reports are ordered primarily by the chain of associative links. When recall halts, to honour the idea that the subject uses all information available as in free recall, we allow recall to restart. In free recall, report is restarted by using Item 0 as a prompt. Because the two paradigms involve very different constraints, serial recall requires a different restart mechanism.

In serial recall, the words studied most recently are strongly represented in the lexicon. But the task requires report in order. The two facts—that recently studied words are strongly represented in the lexicon and that report must be in first-to-last order—are in conflict. To resolve the conflict, the model uses order information derived from each word's strength of representation in the lexicon. To do so, the model changes the criterion for report. Instead of reporting the word whose momentary strength is closest to 1.0 (and, thereby, risking a report that is out of order), the model picks the word whose *item strength* is within the response boundary and is closest to the boundary. That is, the model abandons associative information in favour of item information. Subsequently, if no response is available, report halts.

Because order is constrained in serial recall, most recent studies of serial recall have focussed on context information that subjects presumably use to order materials and on the pattern of errors made during recall. Note that the model does not use context information to guide serial report; instead, it guides recall using the same information that is used in free recall (but it changes the way the information is used). We want to acknowledge, however, that some tasks (such as a list-discrimination task; Jacoby, 1991) require context information. To respond to the requirements of such tasks, the model would have to encode context information. We will discuss the use of environmental context information later.

We start our demonstration of serial recall by considering the immediate serial-recall experiment by Klein, Addis, and Kahana (2005). Their subjects studied lists of 19 words for immediate serial recall. We fit the model to obtain parameters appropriate to their task; the results shown in Figure 16.
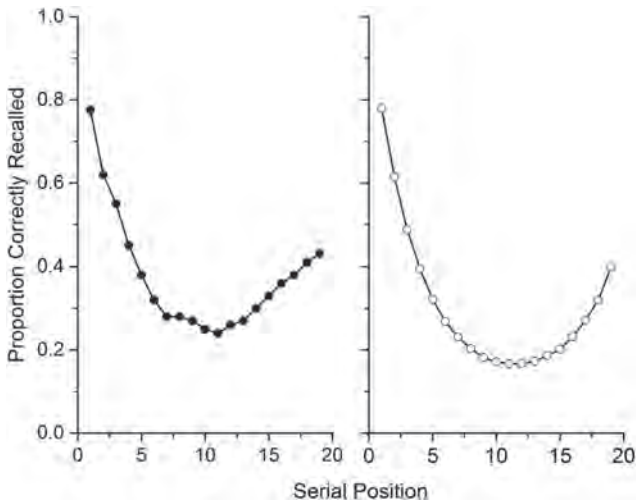
Figure 16. Accuracy as a function of serial position in serial recall. The closed symbols present the first trial averaged across subjects from Klein et al. (2005). The open symbols present the simulation averaged over 20 independent runs of the paradigm.

The left panel of Figure 16 shows Klein et al.'s (2005) data from their Figure 2(A); the right panel summarizes the simulated data. As shown in Figure 16, the simulation matches the data well.

Figure 17 shows the CRP function taken from Klein et al.'s (2005) data along with the corresponding function taken from the simulation shown in Figure 17. As before, the open circles show the CRP function from the simulation shown in Figure 16. The principle result in both the data and the simulation is a strong preference to report in forward direction. It is important to note that the simulation of the serial-position curve was obtained by fitting the data to obtain appropriate parameters. The CRP function from the simulation falls out without refitting, that is, the CRP

function falls out for free as a natural consequence of the memory system.

We used our simulation of Klein et al.'s (2005) experiment to assess the model's ability to capture two characteristics of immediate serial recall documented much later by Farrell, Hurlstone, and Lewandowsky (2013). Figure 18 shows the simulated recall probability as a function of the lag between successive responses, *when conditioned on the first-order error*. There are two points of interest. First, the pattern matched the empirical pattern of the majority of the 13 serial-recall experiments summarised by Farrell et al. (see their Figure 4). Second, the lag function in Figure 18 shows a higher probability for the item preceding the item just recalled. The usual CRP function shows a higher probability for the item following the item just recalled (as shown in Figure 17). Farrell et al. cite the reversal of the CRP following the first-order error as evidence against simple chaining models (see also Henson, Norris, Page, & Baddeley, 1996). Although the holographic model uses item-to-item associations—as do simple chaining models—it is also able to predict the reversal of the CRP function.

Figures 17 and 18 illustrate a unique, almost contradictory, aspect of serial recall. Both fall out of the model without refitting. They reflect the commutativity of circular convolution and the dynamic nature of the holographic lexicon. We know of no other model that can capture both effects.

Farrell et al. (2013) have documented a second characteristic of serial recall. They examined sequential dependencies during serial report across 13 studies. Figure 19 shows the proportion of items recalled as a function of their distance from the correct position (Lag 0). The data for the figure were drawn from the same simulation that reproduced the serial-position curve in Figure 16; the match to empirical data comes for free. Items recalled most often are recalled earliest in the output stream. Items recalled less often tend to be recalled later. As is clear in the figure, the most likely response from the model is the word presented in that position. When simulated subjects err, they are more likely to

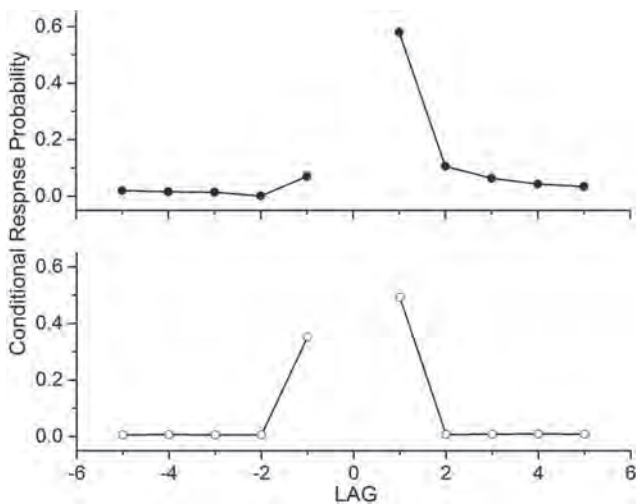*Figure 17.* Conditional response probability as a function of distance from the last item recalled. The filled symbols present data from Klein et al. (2005). The open symbols show the simulation based on 20 independent runs of the paradigm.
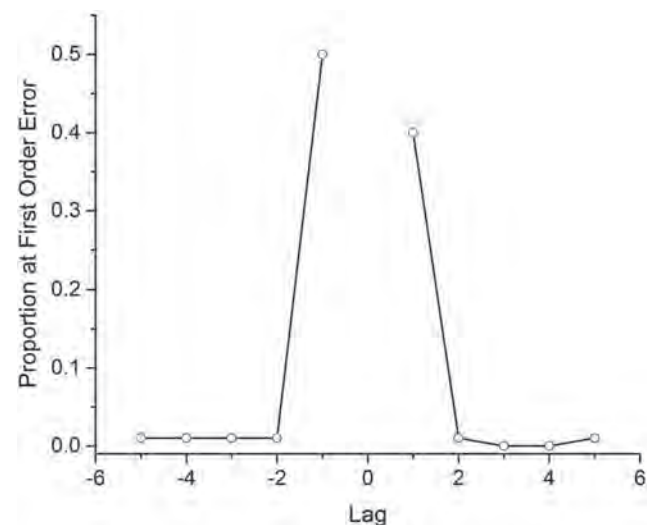

Figure 18. Recall probability as a function of distance from the first-order error. The data are based on the 20 independent runs from the simulation of the Klein et al. (2005) paradigm.
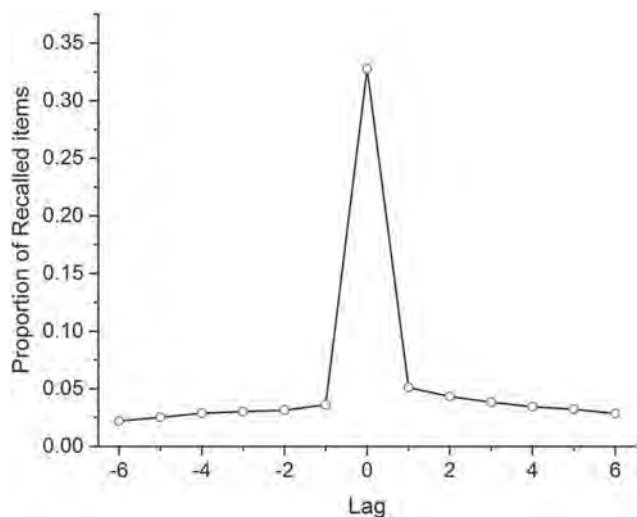
Figure 19. Proportion of items recalled as a function of their distance from the correct position (Lag 0). The data are based on the 20 independent runs from the simulation of the Klein et al. (2005) paradigm.

report from an adjacent position. The pattern illustrated in Figure 19 matched the patterns shown by Farrell et al. for 13 serial-recall studies. Items recalled most often are recalled earliest in the output stream. Items recalled less often tend to be recalled later. The strong negative correlation between the serial-position curve and the output-position curve was also reported by Deese and Kaufman (1957).

Figures 16 through 19 describe the basic characteristics of report in a typical serial-recall task. As before, composite measures and conditional analyses should be interpreted with caution. That said, taken together, the data provide a clear picture of performance in the task. It is important to note that we obtained parameters for the model by fitting the serial-position curve shown in Figure 16. The data shown in the remaining figures are based on data from the simulation shown in Figure 16. We conclude that the model captures the basic facts of report in serial recall.

Following Lashley's (1951) cri de coeur, serial recall is often treated as a laboratory model for the larger issue of serial organisation of sequential behaviour. Recent accounts have stressed the role of environmental context. As the holographic model shows, however, such context is not needed to account for simple verbal experiments. It is definitely needed for exotic tasks such as split-span tasks that put temporal and spatial context in conflict (e.g., Broadbent, 1954, Experiment 2). We will discuss the holographic model's ability to incorporate environmental context later.

Learning Paradigms

Multitrial free recall. The cases illustrated so far involve simple study-recall examples. In a learning experiment, subjects study the materials repeatedly. Klein et al. (2005) compared learning in free recall under two presentation conditions. In the first, a list of words was presented in the same order on each trial. In the second, the words were presented in different orders on successive trials.

Presenting words in the same order on successive trials makes interitem associations available, and as Klein et al. (2005) show, subjects are able to capitalize on them. In particular, the primacy part of the serial-position curve rose more quickly with learning using same-order presentations relative to scrambled presentations. The shift is consistent with the model: Because the same-order condition preserves interitem associations and because associative information is strongest for items in the primacy portion of the serial-position curve, the model predicts faster learning in the primacy part of the serial-position curve.

To implement learning, feedback to the lexicon on successive learning trials was modified to acknowledge both associative and item feedback. In addition, the amount of feedback was based on the strength of the information in memory. The rationale for basing feedback on the current strength in memory is based on work by Rundus (1974). He argued that when attempting to learn a list, subjects pay more attention to (and rehearse more often) those items whose strength is weaker in memory than those that are stronger in memory. Hence, feedback was determined on successive learning trials using the expressions:

$$\text{Item} : \left(0.51 - [\mathbf{a} \bullet \mathbf{L}]\right) \times \delta,$$

$$\text{Associative} : \left(0.51 - ([\mathbf{a} * \mathbf{b}] \bullet \mathbf{L}\right) \times \varphi,$$

where $\mathbf{a}$ is an arbitrary item studied on the previous trial and presented again on the current trial, $\mathbf{a} * \mathbf{b}$ is an arbitrary association studied on the previous trial, $\mathbf{L}$ is the lexicon, δ and φ are item and associative feedback parameters, respectively. The constant 0.51 was set just above the upper bound, at which items become available for recall.

The learning variant of the model has a switch that turns on the use of learning feedback. Feedback is used only if item and associative information are above their respective criteria. The criterion for item information is set to 0.30. For associations the criterion is the value

$$\omega_0 \times (\lambda^{[\mathbf{LL}+1]}) + 0.01$$

Figure 20 presents data from five free-recall learning trials. The left panel presents data drawn from Klein et al. (2005) from a condition in which the study lists were presented in the same order on each trial. The right panel presents the simulated output of the model. As is apparent in Figure 20, left panel the amount of learning between trials is greatest for those items in the asymptotic portion of the serial-position curves. In addition, note that the amount of learning decreased across trials. The simulated data in the right panel anticipates both of these trends. Note that both trends are captured, even though the learning parameters remained constant across list position and across trials in our simulation. As both Klein et al.'s data and our simulation of the data show, when words are presented in the same order on successive trials, interitem associations are available, and both subjects and the model are able to capitalize on them.

Multitrial serial recall. In multitrial serial-recall paradigms, subjects are usually given the same list of items in the same order. In addition to the same-order condition described in Figure 20, Klein et al. (2005) also examined the standard serial-recall paradigm. Figure 21 presents their data from five serial-recall learning trials. When items are presented in the same order on each trial,
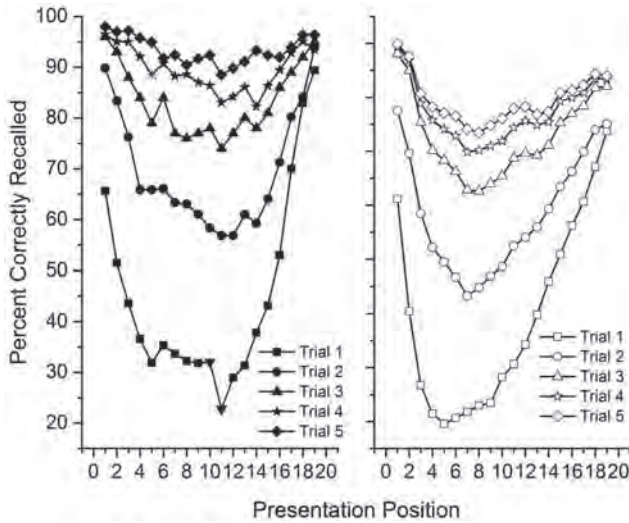
Figure 20. Accuracy as a function of presentation position and practice in free recall. The lists are presented in the same order. The left panel presents data from Klein et al. (2005); the right panel presents the simulation of the data averaged over the 20 independent runs of the paradigm.

serial-recall learning and free-recall learning have much in common: The amount of learning between trials is greatest for those items in the asymptotic portion of the serial-position curves and the amount of learning decreases across trials.

Figure 21 shows the original data of Klein et al. (2005). Figure 22 presents a simulation of learning for a 19-item list. Because we did not attempt to fit the Klein et al. learning data, we present the data on a separate figure. Even without fitting, the model captured all trends seen in the original data. Note that the pattern of learning is similar in both free and serial learning. For example, both show greatest learning at the asymptotic portion of the serial-position
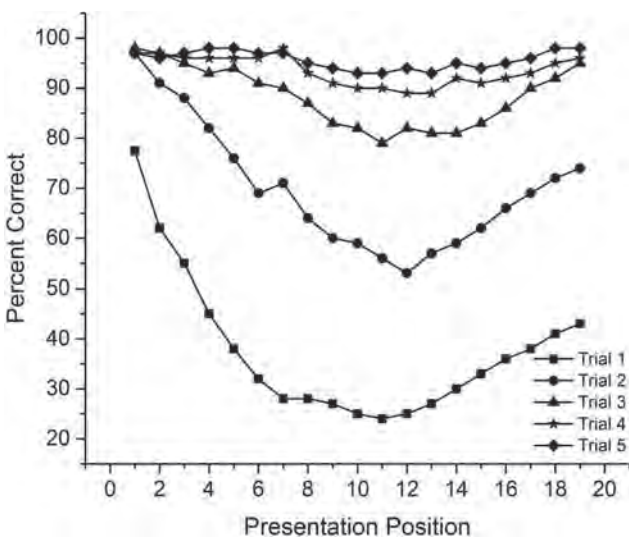
Figure 21. Accuracy as a function of serial position and practice. The data were drawn from Klein et al. (2005, Figure 2). Subjects received five serial-learning trials of a 19-word list.
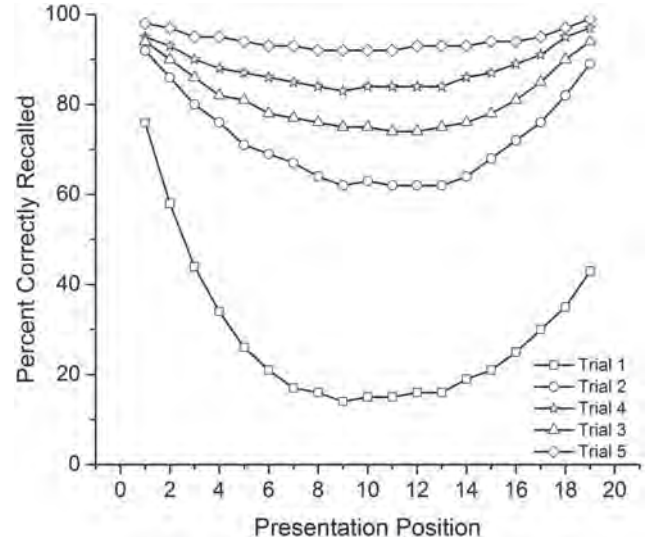
Figure 22. Accuracy as a function of serial position and practice. The data are a simulation averaged over the 20 independent runs of the model. The parameters derived by fitting the first trial of Klein et al.'s (2005) serial-learning experiment and the learning parameters from the free-recall simulation shown in Figure 20.

curve, and in both cases, the amount of learning decreases across trials.

Associative learning is typically described as a gradual increase in excitatory and inhibitory connections among stimulus units—a delta rule. The basic idea is that associative strength accumulates across repetitions (see Rescorla & Wagner, 1972; but see Jamieson, Crump, & Hannah, 2012). Jones and Mewhort (2007) present a learning algorithm that assembles information from different sentences to settle on a meaning rather than on a gradual increment in the strength of an association. Because each word starts as a random vector in their model, the settling operation is, in effect, a noise-reduction algorithm. The present model combines the two ideas. On the one hand, studied items and associations settle at a near optimal value (0.5) across trials, but to the extent that background items in the hologram represent noise, optimizing the strength of studied words and associations reduces noise.

Subjective organisation. Bousfield (1953) asked subjects to study words from various categories and found that the subjects tended to cluster items from like categories when they recalled the items. Such clustering is widely accepted as evidence that subjects organized the material semantically. To generalise the idea, Tulving (1962) used unrelated words and showed that subjects cluster the words into idiosyncratic categories. He referred to the phenomenon as *subjective organisation* to highlight the fact that subjects cluster material in an idiosyncratic fashion.

In their study of learning in free recall using words presented in the same order, Klein et al. (2005) preserved interitem associations. To minimise the temptation to exploit such interitem associations, one should present the words in different orders (as they did in a second condition). The ideal method would be to ensure (a) that each word appeared in each serial position equally often, and (b) that each word followed each other word equally often.

Fortunately, the technique for ordering words in such a way was pioneered by Tulving (1962) in his work on subjective organisa-

tion.[3] He required subjects to learn a list of 16 words using a 16-trial study-test procedure. On each study trial, the list was presented in a unique order, and after each study trial, the subjects were required to report many words as possible. Across a series of 16 study-test trials, each word was presented equally often in each serial position, and each word followed each other word equally often. The subjects learned the list with the mean number of words reported correctly increasing exponentially across the 16 trials from approximately five to approximately 15.

In addition to measuring accuracy of report, Tulving (1962) measured the first-order structure imposed during report and found that as subjects learned the material, they imposed an increasing amount of structure in their report. Indeed, the correlation between accuracy of report and amount of structure was astonishingly high (r = .96). In light of the correlation, it is hard to avoid the conclusion that, for Tulving's subjects, learning was organisation.

To study learning in free recall, we administered 16 study-test trials to the model. Each study trial presented the words to be learned in a different order organized across the simulated subjects using a balanced Latin square constructed, as recommended by Alimena (1962). Each simulated subject was given a unique vocabulary. After each study trial, the simulated subjects were required to recall as many words as possible. In addition to keeping track of accuracy across trials, we also scored Tulving's (1962) subjective organisation (SO) measure. Consistent with his empirical results, the model learned across trials: Accuracy rose from a mean of 3.8 words to a mean of 15.7. SO scores also increased, and as in Tulving's study, the correlation between accuracy and SO was high (ρ = .91).

The accuracy and SO curves for our simulated data are shown in Figure 23. It is clear that the model captures both learning and the development of organisation during free-recall experiments. The model has no organisational rules. The increase in SO across trials is a product of natural processing and storage mechanisms.

We were surprised by the success of the free-recall model in capturing SO. We had expected that subjective organisation would depend on semantic information. With recent developments in

building vectors that represent word meaning (e.g., LSA, Landauer & Dumais, 1997; BEAGLE, Jones & Mewhort, 2007), and subject to the constraints documented by Kelly et al. (2013), it is possible to include semantic information. That said, the model captures organisation without exploiting semantics. Its success reflects associative feedback linking pairs of recalled items combined with the variance of moment-to-moment changes in the strength of items and associations in the lexicon.

If we were to use vectors that include semantic information, the model should exploit the new information and the feedback parameters should change. In this connection, we note that Tulving's (1962) demonstration experiment used words that were unrelated semantically but that were stratified into subsets by a wide range of word-frequency values. For the simulation shown in Figure 23, we did not attempt to reproduce the word-frequency manipulation.

Discussion

In this article, we have advanced a new theory for ordering recall in free- and serial-recall tasks. The heart of the theory is a holographic lexicon that exhibits two important properties: (a) the hologram is robust to loss of medium (illustrated in Figure 4), a property desirable in a biological system; and (b) the hologram is a dynamic store so that altering the strength an item or an association stored in the hologram also alters the strength of all items and associations as a function of their similarity to the item altered (illustrated in Figure 6).

According to the theory, the hologram holds a representation (vector) for each word and for word-to-word association among the words. The hologram can store additional information—such as an association between a word and a context cue—but we have focussed on only two sources of information: words and word-to-word associations. Studying a word, or an association, reinforces the corresponding representation in the hologram. In addition, feedback is created during report, and because the hologram is a dynamic store, both study and report alter the strength of all items in the hologram. The nature and amount of feedback created during study and report depend on the experimenter's instructions (translated into the subject's intentions). In particular, feedback is different when subjects are asked to learn a list across trials. Learning can occur without instructions, but is stronger when subjects are instructed to learn the list.

According to the theory, subjects try to use all of the all information available in the hologram when selecting words to report. In the demonstrations reported here, we have explored the use of item (word) and associative (word-to-word) information; we acknowledge, however, that word-to-context associations may also be used in the special circumstances explored by Brown et al. (2000) or by Bryden (1967). In addition to the information stored in the hologram proper, the theory proposes control mechanisms designed to help extract all useful information from the hologram.
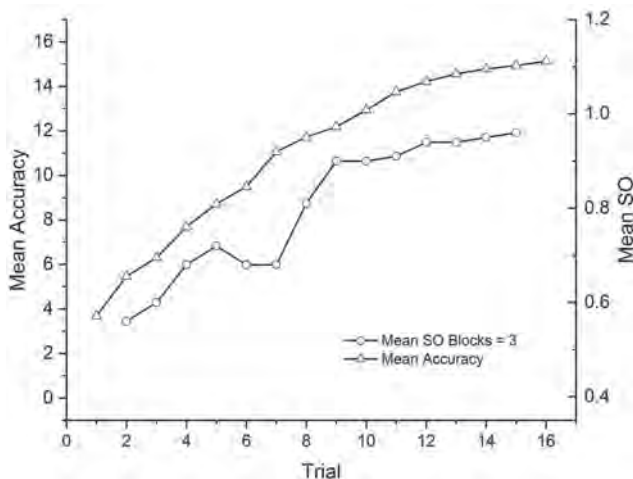
Figure 23. Accuracy (left scale) and subjective organisation (right scale) as a function of practice. The open triangles present accuracy from averaged across 20 runs of Tulving's (1962) paradigm. The open circles present the corresponding subjective organisation scores.

---

[3] Tulving used a 16×16 completely balanced Latin Square to order the materials. He noted that a balanced $N \times N$ square can be constructed easily if ($N + 1$) is a prime number. Clearly, Tulving was aware that remote carry-over effects are not controlled using a simple balanced square; Alimena (1962) has provided an algorithm for eliminating remote carry-over effects in Latin Squares provided that ($N + 1$) is a prime number.

One characteristic of the model deserves emphasis. It will not work well using a small lexicon (e.g., 50 to 250 words). The present simulations are based on a lexicon of 2,000 words and 1,999,000 word-to-word associations. A toy implementation will not work. If the dimensionality of the item vectors is too small, the number of items that can be stored is restricted, and the interactions within the hologram are too small to drive the kind of results shown here. We have replicated the simulations presented here using a both a larger and a smaller vocabulary (with a corresponding modification in the dimensionality of the item vectors), but the important point is that the model's parameters will adapt to the dimensionality of the item vectors and to the number of items.

The single-store account stands in contrast to dual-store models (e.g., the modal model). Several differences deserve emphasis. The dual-store account assumes that STM is labile and subject to interference, whereas long-term memory is stable but subject to error. The holographic account assumes dynamic storage that is both permanent and subject to momentary changes in strength. As noted earlier, the dual-store idea has been subject to severe criticism (e.g., Bjork & Whitten, 1974; Neath, 1993). The fundamental change implied by a holographic model is the focus on process rather than on static storage. The holographic model assumes that subjects know their vocabulary and that activation of particular items depends on the processing currently undertaken. There is no need to transfer material from a short-term store to the knowledge base. The focus on process makes terms such as "top down" versus "bottom up" obsolete. In terms of conventional theory, the holographic account is, perhaps, closest to Cowan's (1988, 1993) idea that STM is an activated subset of long-term memory. He postulated an attentional mechanism activates a subset of memory. The holographic idea does not require a separate attentional mechanism and escapes questions concerning the capacity of various buffer memories.

From the perspective of the holographic model, STM is an unnecessary construct. The holographic model's success in capturing Craik's (1970) data calls the idea of a STM into question once more. The fact that the concept is unnecessary and has been misapplied in the analysis of particular experiments—misapplication highlighted by Bjork and Whitten (1974) and by Neath (1993)—does not, by itself, mean that there is no short-term store; it means the basic evidence supporting the concept has been blunted. In light of the success of the single-store model presented here, the onus is on those who promote the two-store idea to present new empirical evidence to support the idea. Perhaps first, however, proponents of the dual-storage idea need to come up with a working model that can match the range of phenomena illustrated here.

The single-store holographic account offers an alternative that is both more parsimonious (one vs. two storage mechanisms) and in keeping with biological evidence. Kandel (2007, 2009) has argued that the same neural tissue supports both short-term (chemical) and long-term (structural) changes in memory. To the extent that a dual-storage mechanism implies distinct anatomical loci, Kandel's evidence denies its feasibility.

The psychological literature too frequently includes concepts that are *consistent with data* but that are not *forced by the data*. We need experiments that force a conclusion; without such data, theorists risk reifying mechanisms that lack adequate empirical support. The danger, of course, is that people start to believe in the reified concept. Some may even seek its presence in the brain.[4] Too often, search for brain correlates is driven by theory known to be problematic.

The holographic model captures both item and order of report as a by-product of storage and report operations; it does not require separate order information. The question of what information subjects use to order report has been confused in the literature with the mechanism by which the information is used. For example, several accounts associate a series of temporal cues with items to be reported, and by rerunning the temporal series, order recall by recovering the associated items (Brown et al., 2000; Howard & Kahana, 2002). In our view, such a theoretical move focuses too much on environmental information at the expense of internal information associated with the way words and their associates are stored.

The principles of our account allow several sources of environmental information to be used in combination with simple item and associative information: Report is based on momentary activation, a value derived by combining item and associative information. An easy extension to the mechanism would be to add temporal and/or spatial environmental information (in the form of an association between each word and a corresponding context vector). Such an extension would allow the ordering mechanism to combine internal and environmental context to handle the temporal–spatial conflicts addressed by Bryden (1967), the uneven timing cases explored by Brown et al. (2000), and list discrimination studies of the sort pioneered by Jacoby (1991). At issue, is a fundamental computing-science question of data versus process. We have explored the use of two kinds of information using a mechanism flexible enough to combine them. The mechanism is also flexible enough to admit other sources of information. Although we have focussed on the flexibility provided by the hologram, the thrust of our account is that control of report rests on the integration of different sources of information.

Finally, a major limitation of the model deserves mention. The model uses vectors of random numbers to represent words. The vectors have not included structure that we know affects performance. For example, we know from Baddeley's (1986) work that phonological characteristics of words are important in memory, especially in experiments using short lists. Because the model does not represent phonological structure, it is unable to handle such data. Future work should address such holes in the model by extending its representation assumptions to include the missing structure.

-----

[4] The problem is illustrated nicely in the history of physics. At one point, people believed that light required a medium to support its wave properties; the literature invented the aether as the appropriate medium. The concept was so reified that authorities did not question it until the famous Michelson–Morley experiment (see Michelson & Morley, 1886).

Résumé

Nous présentons une théorie holographique de la mémoire humaine. Selon cette théorie, le vocabulaire d'un sujet réside dans une représentation distribuée dynamique, soit un hologramme. Le fait d'étudier ou de se rappeler d'un mot altère autant la représentation actuelle de ce mot dans l'hologramme que tous les mots qui y sont associés. Le rappel est toujours déclenché par une consigne de rappel (soit une directive de départ ou le mot venant d'être évoqué). La séquence d'évocation est une fonction commune entre l'item et l'information associative résidant dans l'hologramme au moment de l'évocation. Nous appliquons le modèle aux données d'archive impliquant une variété de processus tels que le simple

rappel libre, l'apprentissage en mode de rappel libre multi-essais, le simple rappel sériel et l'apprentissage en mode de rappel sériel multi-essais. Le modèle capture l'exactitude et la séquence d'évocation tant dans un contexte de rappel sériel que de celui d'un rappel libre. Il capture aussi les processus d'organisation subjective et d'apprentissage dans un contexte de rappel libre multi-essais. Nous proposons ce modèle comme alternative à la théorie de la mémoire à court et à long terme présentée par le modèle modal.

Mots-clés : hologramme, mémoire, apprentissage, rappel.

References

Alimena, B. S. (1962). A method of determining unbiased distribution in the Latin square. *Psychometrika, 27,* 315–317. http://dx.doi.org/10.1007/BF02289627

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of learning and motivation* (Vol. 2, pp. 89–195). New York, NY: Academic Press. http://dx.doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A. D. (1986). *Working memory*. New York, NY: Oxford University Press.

Beaman, C. P., & Morton, J. (2000). The separate but related origins of the recency effect and the modality effect in free recall. *Cognition, 77,* B59–B65. http://dx.doi.org/10.1016/S0010-0277(00)00107-4

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6,* 173–189. http://dx.doi.org/10.1016/0010-0285(74)90009-7

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113,* 201–233. http://dx.doi.org/10.1037/0033-295X.113.2.201

Bousfield, W. A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology, 49,* 229–240. http://dx.doi.org/10.1080/00221309.1953.9710088

Broadbent, D. E. (1954). The role of auditory localization in attention and memory span. *Journal of Experimental Psychology, 47,* 191–196. http://dx.doi.org/10.1037/h0054182

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539–576. http://dx.doi.org/10.1037/0033-295X.114.3.539

Brown, G. D. A., Preece, T., & Hulme, C. (2000). Oscillator-based memory for serial order. *Psychological Review, 107,* 127–181. http://dx.doi.org/10.1037//0033-295X.107.1.127

Bryden, M. P. (1967). A model for the sequential organization of behaviour. *Canadian Journal of Psychology, 21,* 37–56. http://dx.doi.org/10.1037/h0082960

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review, 106,* 551–581. http://dx.doi.org/10.1037/0033-295X.106.3.551

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82,* 407–428. http://dx.doi.org/10.1037/0033-295X.82.6.407

Corballis, M. C. (1967). Serial order in recognition and recall. *Journal of Experimental Psychology, 74,* 99–105. http://dx.doi.org/10.1037/h0024500

Cowan, N. (1988). Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological Bulletin, 104,* 163–191. http://dx.doi.org/10.1037/0033-2909.104.2.163

Cowan, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition, 21,* 162–167. http://dx.doi.org/10.3758/BF03202728

Craik, F. I. M. (1970). The fate of primary memory items in free recall. *Journal of Verbal Learning & Verbal Behavior, 9,* 143–148. http://dx.doi.org/10.1016/S0022-5371(70)80042-1

Deese, J., & Kaufman, R. A. (1957). Serial effects in recall of unorganized and sequentially organized verbal material. *Journal of Experimental Psychology, 54,* 180–187. http://dx.doi.org/10.1037/h0040536

Dennis, S. (2009). Can a chaining model account for serial recall? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the XXXI annual conference of the Cognitive Science Society* (pp. 2813–2818). Austin, TX: Cognitive Science Society.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Teachers College, Columbia University. (Original work published 1885)

Farrell, S., Hurlstone, M. J., & Lewandowsky, S. (2013). Sequential dependencies in recall of sequences: Filling in the blanks. *Memory & Cognition, 41,* 938–952. http://dx.doi.org/10.3758/s13421-013-0310-0

Farrell, S., & Lewandowsky, S. (2002). An endogenous distributed model of ordering in serial recall. *Psychonomic Bulletin & Review, 9,* 59–79. http://dx.doi.org/10.3758/BF03196257

Farrell, S., & Lewandowsky, S. (2012). Response suppression contributes to recency in serial recall. *Memory & Cognition, 40,* 1070–1080. http://dx.doi.org/10.3758/s13421-012-0212-6

Franklin, D. R. J. (2013). *Control processes in free recall* (Unpublished doctoral dissertation). Queen's University, Kingston, Ontario.

Franklin, D. R. J., & Mewhort, D. J. K. (2002). An analysis of immediate memory: The free-recall task. In N. J. Dimpoloulos & K. F. Li (Eds.), *High performance computing systems and applications 2000* (pp. 465–479). New York, NY: Kluwer. http://dx.doi.org/10.1007/978-1-4615-0849-6_30

Franklin, D. R. J., & Mewhort, D. J. K. (2013). Control processes in free recall. In R. L. West & T. C. Stewart (Eds.), *Proceedings of the 12th international conference on cognitive modelling* (pp. 179–184). Ottawa, Canada: Carleton University.

Gabor, D. (1968). Improved holographic model of temporal recall. *Nature, 217,* 1288–1289. http://dx.doi.org/10.1038/2171288a0

Gabor, D. (1969). Associative holographic memories. *IBM Journal of Research and Development, 13,* 156–159. http://dx.doi.org/10.1147/rd.132.0156

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning & Verbal Behavior, 5,* 351–360. http://dx.doi.org/10.1016/S0022-5371(66)80044-0

Glucksberg, S., & McCloskey, M. (1981). Decisions about ignorance: Knowing that you don't know. *Journal of Experimental Psychology: Human Learning and Memory, 7,* 311–325. http://dx.doi.org/10.1037/0278-7393.7.5.311

Grossberg, S., & Pearson, L. R. (2008). Laminar cortical dynamics of cognitive and motor working memory, sequence learning and performance: Toward a unified theory of how the cerebral cortex works. *Psychological Review, 115,* 677–732. http://dx.doi.org/10.1037/a0012618

Gruneburg, M. M. (1970). A dichotomous theory of memory—Unproved and unprovable. *Acta Psychologica, Amsterdam, 34,* 489–496. http://dx.doi.org/10.1016/0001-6918(70)90042-9

Hebb, D. O. (1961). Distinctive features of learning in the higher animal. In J. F. Delafresnaye (Ed.), *Brain mechanisms and learning* (pp. 37–46). London, UK: Oxford University Press.

Helstrup, T. (1984). Serial position phenomena: Training and retrieval effects. *Scandinavian Journal of Psychology, 25,* 227–250. http://dx.doi.org/10.1111/j.1467-9450.1984.tb01015.x

Henson, R. N. A., Norris, D., Page, M. P. A., & Baddeley, A. D. (1996). Unchained memory: Error patterns rule out chaining models of immediate serial recall. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology, 49,* 80–115. http://dx.doi.org/10.1080/713755612

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269–299. http://dx.doi.org/10.1006/jmps.2001.1388

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language, 30,* 513–541. http://dx.doi.org/10.1016/0749-596X(91)90025-F

Jamieson, R. K., Crump, M. J. C., & Hannah, S. D. (2012). An instance theory of associative learning. *Learning & Behavior, 40,* 61–82. http://dx.doi.org/10.3758/s13420-011-0046-2

Johns, B. T., & Jones, M. N. (2010). Evaluating the random representation assumption of lexical semantics in cognitive models. *Psychonomic Bulletin & Review, 17,* 662–672. http://dx.doi.org/10.3758/PBR.17.5.662

Johnson, G. J. (1975). Positional mediation of intraserial associations. *The American Journal of Psychology, 88,* 49–63. http://dx.doi.org/10.2307/1421664

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language, 55,* 534–552. http://dx.doi.org/10.1016/j.jml.2006.07.003

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114,* 1–37. http://dx.doi.org/10.1037/0033-295X.114.1.1

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24,* 103–109. http://dx.doi.org/10.3758/BF03197276

Kandel, E. R. (2007). *In search of memory: The emergence of a new science of mind*. New York, NY: Norton.

Kandel, E. R. (2009). The biology of memory: A forty-year perspective. *The Journal of Neuroscience, 29,* 12748–12756. http://dx.doi.org/10.1523/JNEUROSCI.3958-09.2009

Kelly, M. A., Blostein, D., & Mewhort, D. J. K. (2013). Encoding structure in holographic reduced representations. *Canadian Journal of Experimental Psychology, 67,* 79–93. http://dx.doi.org/10.1037/a0030301

Klein, K. A., Addis, K. M., & Kahana, M. J. (2005). A comparative analysis of serial and free recall. *Memory & Cognition, 33,* 833–839. http://dx.doi.org/10.3758/BF03193078

Koppenaal, L., & Glanzer, M. (1990). An examination of the continuous distractor task and the "long-term recency effect." *Memory & Cognition, 18,* 183–195. http://dx.doi.org/10.3758/BF03197094

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 211–240. http://dx.doi.org/10.1037/0033-295X.104.2.211

Lashley, K. S. (1951). The problem of serial order in behavior. In R. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon symposium* (pp. 112–136). New York, NY: Wiley.

Lewandowsky, S., & Farrell, S. (2008). Phonological similarity in serial recall: Constraints on theories of memory. *Journal of Memory and Language, 58,* 429–448. http://dx.doi.org/10.1016/j.jml.2007.01.005

Lewandowsky, S., & Murdock, B. B. (1989). Memory for serial order. *Psychological Review, 96,* 25–57. http://dx.doi.org/10.1037/0033-295X.96.1.25

Longuet-Higgins, H. C. (1968). Holographic model of temporal recall. *Nature, 217,* 104. http://dx.doi.org/10.1038/217104a0

Melton, A. W. (1963). Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior, 2,* 1–21. http://dx.doi.org/10.1016/S0022-5371(63)80063-8

Mewhort, D. J. K. (1973). Retrieval tags and order of report in dichotic listening. *Canadian Journal of Psychology, 27,* 119–126. http://dx.doi.org/10.1037/h0082461

Mewhort, D. J. K. (1974). Accuracy and order of report in tachistoscopic identification. *Canadian Journal of Psychology, 28,* 383–398. http://dx.doi.org/10.1037/h0082004

Mewhort, D. J. K., & Campbell, A. J. (1981). Toward a model of skilled reading: An analysis of performance in tachistoscopic tasks. In G. E. MacKinnon & T. G. Waller (Eds.), *Reading research: Advances in theory and practice* (Vol. 3, pp. 39–118). New York, NY: Academic Press.

Mewhort, D. J. K., & Popham, D. (1991). Serial recall of tachistoscopic letter strings. In W. E. Hockley & S. Lewandowsky (Eds.), *Relating theory and data: Essays on human memory in honor of Bennet B. Murdock* (pp. 425–443). Hillsdale, NJ: Erlbaum.

Michelson, A. A., & Morley, E. W. (1886). Influence of motion of the medium on the velocity of light. *American Journal of Science, 31,* 377–385. http://dx.doi.org/10.2475/ajs.s3-31.185.377

Murdock, B. B., Jr. (1962). The serial position effect of free recall. *Journal of Experimental Psychology, 64,* 482–488. http://dx.doi.org/10.1037/h0045106

Murdock, B. B., Jr. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89,* 609–626. http://dx.doi.org/10.1037/0033-295X.89.6.609

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 689–697. http://dx.doi.org/10.1037/0278-7393.19.3.689

Neath, I. (1993). Contextual and distinctive processes and the serial position function. *Journal of Memory and Language, 32,* 820–840. http://dx.doi.org/10.1006/jmla.1993.1041

Nipher, F. E. (1878). On the distribution of errors in numbers written from memory. *Transactions of the Academy of Science of St. Louis, 3,* ccx–ccxi.

Norman, D. A. (1970). Appendix: Serial position curves. In D. A. Norman (Ed.), *Models of human memory* (pp. 511–518). New York, NY: Academic Press.

Plate, T. A. (2003). *Holographic reduced representations*. CSLI Lecture Notes Number 150. Stanford, CA: CSLI Publications.

Poggio, T. (1973). On holographic models of memory. *Kybernetik, 12,* 237–238. http://dx.doi.org/10.1007/BF00270577

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in FORTRAN The art of scientific computing* (2nd ed.). New York, NY: Cambridge University Press.

Raaijmakers, J. G. W., & Shiffrin, R. M. (1981). Search of associative memory. *Psychological Review, 88,* 93–134. http://dx.doi.org/10.1037/0033-295X.88.2.93

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Roberts, W. A. (1972). Free recall of word lists varying in length and rate of presentation: A test of total-time hypotheses. *Journal of Experimental Psychology, 92,* 365–372. http://dx.doi.org/10.1037/h0032278

Rundus, D. (1974). Output order and rehearsal in multi-trial free recall. *Journal of Verbal Learning & Verbal Behavior, 13,* 656–663. http://dx.doi.org/10.1016/S0022-5371(74)80053-8

Rundus, D., & Atkinson, R. C. (1970). Rehearsal processes in free recall: A procedure for direct observation. *Journal of Verbal Learning & Verbal Behavior, 9,* 99–105. http://dx.doi.org/10.1016/S0022-5371(70)80015-9

Shiffrin, R., Ratcliff, R., Murnane, K., & Nobel, P. (1993). TODAM and the list-strength and list-length effects: Comment on Murdock and Kahana (1993a). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19,* 1445–1449. http://dx.doi.org/10.1037/0278-7393.19.6.1445

Shiffrin, R. M. (1999). 30 years of memory. In C. Izawa (Ed.), *On human memory: Evolution, progress, and reflections on the 30th anniversary of the Atkinson-Shiffrin model* (pp. 17–33). Mahwah, NJ: Erlbaum.

Slamecka, N. J. (1968). An examination of trace storage in free recall. *Journal of Experimental Psychology, 76,* 504–513. http://dx.doi.org/10.1037/h0025695

Sloman, S. A., Bower, G. H., & Rohrer, D. (1991). Congruency effects in part-list cuing inhibition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17,* 974–982. http://dx.doi.org/10.1037/0278-7393.17.5.974

Tulving, E. (1962). Subjective organization in free recall of "unrelated" words. *Psychological Review, 69,* 344–354. http://dx.doi.org/10.1037/h0043150

Tulving, E., & Hastie, R. (1972). Inhibition effects of intralist repetition in free recall. *Journal of Experimental Psychology, 92,* 297–304. http://dx .doi.org/10.1037/h0032367

Ward, G., Tan, L., & Grenfell-Essam, R. (2010). Examining the relationship between free recall and immediate serial recall: The effects of list length and output order. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1207–1241. http://dx.doi.org/10.1037/a0020122

Watkins, M. J., & Watkins, O. C. (1974). Processing of recency items for free recall. *Journal of Experimental Psychology, 102,* 488–493. http:// dx.doi.org/10.1037/h0035903

Waugh, N. C., & Norman, D. A. (1965). Primary Memory. *Psychological Review, 72,* 89–104. http://dx.doi.org/10.1037/h0021797

Young, R. K. (1962). Tests of three hypotheses about the effective stimulus in serial learning. *Journal of Experimental Psychology, 63,* 307–313. http://dx.doi.org/10.1037/h0038534

The Psychology of Intelligence Analysis: Drivers of Prediction Accuracy in World Politics

Barbara Mellers, Eric Stone, Pavel Atanasov,
Nick Rohrbaugh, S. Emlen Metz, Lyle Ungar,
Michael M. Bishop, and Michael Horowitz
University of Pennsylvania

Ed Merkle
University of Missouri

Philip Tetlock
University of Pennsylvania

This article extends psychological methods and concepts into a domain that is as profoundly consequential as it is poorly understood: intelligence analysis. We report findings from a geopolitical forecasting tournament that assessed the accuracy of more than 150,000 forecasts of 743 participants on 199 events occurring over 2 years. Participants were above average in intelligence and political knowledge relative to the general population. Individual differences in performance emerged, and forecasting skills were surprisingly consistent over time. Key predictors were (a) dispositional variables of cognitive ability, political knowledge, and open-mindedness; (b) situational variables of training in probabilistic reasoning and participation in collaborative teams that shared information and discussed rationales (Mellers, Ungar, et al., 2014); and (c) behavioral variables of deliberation time and frequency of belief updating. We developed a profile of the best forecasters; they were better at inductive reasoning, pattern detection, cognitive flexibility, and open-mindedness. They had greater understanding of geopolitics, training in probabilistic reasoning, and opportunities to succeed in cognitively enriched team environments. Last but not least, they viewed forecasting as a skill that required deliberate practice, sustained effort, and constant monitoring of current affairs.

Keywords: forecasting, predictions, skill, probability judgment, accuracy

Supplemental materials: http://dx.doi.org/10.1037/xap0000040.supp

Predicting the future is an integral part of human cognition. We reach for an umbrella when we expect rain. We cross the street when the light turns green and expect cars to stop. We help others and expect reciprocity—they will help us in future situations. Without some ability to generate predictions, we could neither plan for the future nor interpret the past.

Psychologists have studied the accuracy of intuitive predictions in many settings, including eyewitness testimony (Loftus, 1996), affective forecasting (Wilson & Gilbert, 2005), and probability judgment (Kahneman, Slovic, & Tversky, 1982). This literature paints a disappointing picture. Eyewitness testimonies are often faulty (Wells, 2014; Wells & Olson, 2003), affective forecasts stray far from affective experiences (Schkade & Kahneman, 1998), and probability estimates are highly susceptible to overconfidence, base rate neglect, and hindsight bias (Fischhoff & Bruine de Bruin, 1999; Fischhoff, Slovic, & Lichtenstein, 1977; Kahneman et al., 1982).

To make matters worse, intuitive predictions are often inferior to simple statistical models in domains ranging from graduate school admissions to parole violations (Dawes, Faust, & Meehl, 1989; Swets, Dawes, & Monahan, 2000). In political forecasting, Tetlock (2005) asked professionals to estimate the probabilities of events

up to 5 years into the future—from the standpoint of 1988. Would there be a nonviolent end to apartheid in South Africa? Would Gorbachev be ousted in a coup? Would the United States go to war in the Persian Gulf? Experts were frequently hard-pressed to beat simple actuarial models or even chance baselines (see also Green and Armstrong, 2007).

A Forecasting Competition

It was against this backdrop that the National Academy of Sciences issued a report on the quality of intelligence analysis (Fischhoff & Chauvin, 2011). A key theme was the need to systematically track the accuracy of probabilistic forecasts that analysts routinely (albeit covertly) make. In response, the Intelligence Advanced Research Projects Activity (IARPA), the research and development branch of the Office of the Director of National Intelligence, launched a large-scale forecasting tournament designed to monitor the accuracy of probabilistic forecasts about events that occurred around the world. Five university-based research groups competed to develop methods to elicit and aggregate forecasts to arrive at the most accurate predictions. Our research group consistently outperformed the other groups 2 years in a row.

Within the tournament, accuracy of probability judgments was assessed by the Brier scoring rule (Brier, 1950), a widely used measure in fields ranging from meteorology (Murphy & Winkler, 1984) to medical imaging (Itoh et al., 2002; Steyerberg, 2009). The Brier scoring rule is "strictly proper" in the sense that it incentivizes forecasters to report their true beliefs—and avoid making false-positive versus false-negative judgments. These scores are sums of squared deviations between probability forecasts and reality (in which reality is coded as "1" for the event and "0" otherwise). They range from 0 (best) to 2 (worst). Suppose a forecaster reported that one outcome of a two-option question was 75% likely, and that outcome occurred. The forecaster's Brier score would be $(1 - 0.75)^2 + (0 - 0.25)^2 = 0.125$. This measure of accuracy is central to the question of whether forecasters can perform well over extended periods and what factors predict their success.

Consistency in Forecasting Skills

In this article, we study variation in the degree to which people possess, and are capable of developing, geopolitical forecasting skill. Skill acquisition and expertise has been examined in numerous domains. We were unsure whether it was even possible to develop skill in this domain. Geopolitical forecasting problems can be complex, requiring a balance of clashing causal forces. It is no wonder that some attribute forecasting success to skill, whereas others attribute it to luck. Skeptics argue that accurate forecasts are fortuitous match-ups between reality and observers' preconceptions in a radically unpredictable world. From this perspective, we would find little or no consistency in individual accuracy across questions (Almond & Genco, 1977; Taleb, 2007).

Our prediction task involves several factors usually associated with poor performance, including a dynamic prediction environment, a long delay before feedback on most questions, the lack of empirically tested decision aids, and a reliance on subjective judgment (Shanteau, 1992). Indeed, Reyna, Chick, Corbin, and Hsia (2014) showed that intelligence analysts were more suscep-

tible to risky-choice framing effects than either college students or postcollegiate adults, perhaps because they had developed bad habits in an "unfriendly" environment. Although experts may, on average, be poor at exercising good judgment in complex domains like geopolitical forecasting, others suspect that there are systematic individual differences—and that some forecasters will consistently outperform others (Bueno de Mesquita, 2009; Tetlock, 2005).

As we shall soon show, striking individual differences in forecasting accuracy emerged, and these differences created the opportunity to test hypotheses about which assortment of dispositional variables (e.g., cognitive abilities and political understanding), situational variables (e.g., cognitive-debiasing exercises), and/or behavioral variables (e.g., willingness to revisit and update one's beliefs) could predict judgmental accuracy. Insofar as all three classes of variables matter, how are they interrelated? And what are the characteristics of the best forecasters?

Dispositional Variables

Accurate predictions of global events require an array of skills. One needs diverse pockets of content knowledge, a judicious capacity to choose among causal models for integrating and applying content knowledge, and a rapid-fire Bayesian capacity to change one's mind quickly in response to news about shifting base rates and case-specific cues. A natural starting hypothesis is intelligence, a well replicated predictor of success, including job performance (Ree & Earles, 1992; Schmidt & Hunter, 2004), socioeconomic status (Strenze, 2007), academic achievement (Furnham & Monsen, 2009), and decision competence (Del Missier, Mäntylä, & Bruine de Bruin, 2012; Parker & Fischhoff, 2005).

Intelligence

Theories of intelligence vary in complexity, starting with the single-factor model widely known as *g* (Spearman, 1927), the two-factor fluid/crystallized intelligence framework (Cattell, 1963; Cattell & Horn, 1978), the seven basic abilities (Thurstone & Thurstone, 1941), and, finally, the 120-factor cube derived from combinations of content, operation, and product (Guilford & Hoepfner, 1971). Carroll (1993) reanalyzed over 400 data sets that measured cognitive abilities and found overwhelming evidence for a general intelligence factor (interpreted as *g*, fluid intelligence, with domain-specific forms of crystallized intelligence defining additional factors).

Three aspects of intelligence seem most relevant to geopolitical forecasting. One is the ability to engage in *inductive reasoning*, or make associations between a current problem—say, the likelihood of an African leader falling from power—and potential historical analogies. Individuals must look for regularities, form hypotheses, and test them. The second is *cognitive control* (also known as cognitive reflection). Someone with greater cognitive control has the ability to override seemingly obvious but incorrect responses and engage in more prolonged and deeper thought. The third skill is *numerical reasoning*. Numeracy would be especially important for economic questions such as, "Will the price per barrel for November, 2011 Brent Crude oil futures exceed $115 by a given date?" A more numerate forecaster would be likelier to recognize that the answer hinged, in part, on how close the current price was

to the target price and how often price fluctuations of the necessary magnitude occurred within the specified time frame. Our first hypothesis is therefore as follows:

Hypothesis 1: Individuals with greater skill at inductive reasoning, cognitive control, and numerical reasoning will be more accurate forecasters.

Researchers disagree on the relationship between intelligence and expertise. Some claim that experts, such as chess grandmasters, possess exceptional cognitive abilities that place them high up in the tail of the distribution (Plomin, Shakeshaft, McMillan, & Trzaskowski, 2014). Others claim that, beyond a certain moderately high threshold, intelligence is not necessary; what really matters is deep deliberative practice that promotes expertise by enabling the neural networking and consolidation of performance-enhancing knowledge structures (Ericsson, 2014).

The forecasting tournament let us explore the relationship between intelligence and skill development. If the correlation between intelligence and accuracy was positive and remained constant throughout the tournament, one could argue that superior intelligence is necessary for expertise. But if the correlation between intelligence and accuracy were stronger at the beginning and weaker toward the end of the tournament (after deliberative practice), one could argue that deliberative practice is a cognitive leveler, at least within the ability range of the above-average IARPA forecasters.

Thinking Style

Cognitive styles capture *how* people typically think—as opposed to what they think about (e.g., causal theories) and how well they can think (ability). There are as many frameworks for cognitive styles as taxonomies of cognitive abilities (Riding & Cheema, 1991; Vannoy, 1965; Witkin, Oltman, Raskin, & Karp, 1971).

A relevant cognitive style is the tendency to evaluate arguments and evidence without undue bias from one's own prior beliefs—and with recognition of the fallibility of one's judgment (Nickerson, 1987). High scorers on this dimension are *actively open-minded thinkers*. They avoid the "myside bias"—the tendency to bolster one's own views and dismiss contradictory evidence (Baron, 2000). Actively open-minded thinkers have also been found to be more accurate at estimating uncertain quantities (Haran, Ritov, & Mellers, 2013), a task that is arguably similar to estimating the likelihood of future events.

Actively open-minded thinkers also have greater tolerance for ambiguity and weaker *need for closure* (the tendency to want to reach conclusions quickly, often before all the evidence has been gathered, coupled with an aversion to ambiguity; Kruglanski & Webster, 1996; Webster & Kruglanski, 1994). Previous research has found that experts with a greater need for closure reject counterfactual scenarios that prove their theories wrong while embracing counterfactual scenarios that prove their theories right (Tetlock, 1998), a form of motivated reasoning that is likely to hinder attempts to accurately model uncertainty in the real world.

In a related vein, the concept of *hedgehogs versus foxes*, developed by Tetlock (2005), draws on need for closure and taps into a preference for parsimony in political explanations (the hedgehog knows one big thing) versus a preference for eclectic blends of causal precepts (the fox knows many, not-so-big things). Tetlock found that the foxes were less prone to overconfidence in their political predictions. Although we measured actively open-minded thinking, need for closure, and hedgehog versus fox separately, these constructs reflect distinct but related features of cognitive flexibility. Given the strong family resemblance among openness to self-correction, cognitive flexibility, foxiness, and tolerance for ambiguity, we bundle them into our next hypothesis. Forecasters with more open-minded and flexible cognitive styles should be more nuanced in applying pet theories to real-world events—or, more simply,

Hypothesis 2: More open-minded forecasters will be more accurate forecasters.

Political Knowledge

Political knowledge refers to content information necessary for answering factual questions about states of the world. Even the most intelligent and open-minded forecasters need political knowledge to execute multidimensional comparisons of current events with pertinent precedents. Consider the question, "Will the United Nations General Assembly recognize a Palestinian state by September, 30, 2011?" Forecasters with no institutional knowledge would be at a disadvantage. They might read headlines that a majority of the General Assembly favored recognition and infer that recognition was imminent. But someone who knew more about the United Nations might know that permanent members of the Security Council have many ways to delay a vote, such as "tabling the resolution" for a later point in time. This brings us to our third hypothesis:

Hypothesis 3: More politically knowledgeable forecasters will be more accurate forecasters.

Situational Variables

Forecasting accuracy also depends on the environment; forecasters need opportunities for deliberative practice to cultivate skills (Arkes, 2001; Ericsson, Krampe, & Tesch-Romer, 1993; Kahneman & Klein, 2009). Some environments lack these opportunities. Cue-impoverished environments stack the odds against forecasters who wish to cultivate their skills. Environments with delayed feedback, misleading feedback, or nonexistent feedback also restrict learning (Einhorn, 1982).

Mellers, Ungar, et al. (2014) reported two experimentally manipulated situational variables that boosted forecasting accuracy. One was training in probabilistic reasoning. Forecasters were taught to consider comparison classes and take the "outside" view. They were told to look for historical trends and update their beliefs by identifying and extrapolating persistent trends and accounting for the passage of time. They were told to average multiple estimates and use previously validated statistical models when available. When not available, forecasters were told to look for predictive variables from formal models that exploit past regularities. Finally, forecasters were warned against judgmental errors, such as wishful thinking, belief persistence, confirmation bias, and hindsight bias. This training module was informed by a large literature that investigates methods of debiasing (see Lichtenstein

and Fischhoff, 1980; Soll, Milkman, and Payne, in press; and Wilson and Brekke, 1994, for reviews).

The second situational factor was random assignment to teams. Drawing on research in group problem-solving (Laughlin, 2011; Laughlin, Hatch, Silver, & Boh, 2006; MacCoun, 2012; Steiner, 1972), Mellers, Ungar, et al. (2014) designed teams with the goal of ensuring the "process gains" of putting individuals into groups (e.g., benefits of diversity of knowledge, information sharing, motivating engagement, and accountability to rigorous norms) exceeded the "process losses" from teaming (e.g., conformity pressures, overweighting common information, poor coordination, factionalism). The manipulation was successful. Teamwork produced enlightened cognitive altruism: Forecasters in teams shared news articles, argued about the evidence, and exchanged rationales using self-critical epistemic norms. Forecasters who worked alone were less accurate. Here, we explore whether the dispositional variables discussed earlier add to the predictive accuracy of forecasting over and beyond the two situational variables already known to promote accuracy. Our fourth hypothesis is

Hypothesis 4: Dispositional variables, such as intelligence, open-mindedness, and political knowledge will add to the prediction of forecasting accuracy, beyond situational variables of training and teamwork.

Behavioral Variables

Dweck (2006) argues that those with growth mind-sets who view learning and achievement as cultivatable skills are likelier to perform well than those who view learning as innately determined. More accurate forecasters are presumably those with growth mind-sets. In the tournament, behavioral indicators of motivation included the numbers of questions tried and the frequency of belief updating. Engaged forecasters should also spend more time researching, discussing, and deliberating before making a forecast. Our fifth hypothesis is

Hypothesis 5: Behavioral variables that reflect engagement, including the number of questions tried, frequency of updating, and time spent viewing a question before forecasting will add to the prediction of forecasting accuracy, beyond dispositional and situational variables.

Overview

After testing these hypotheses, we build a structural equation model to summarize the interrelationships among variables. Then we develop a profile of the best forecasters. Finally, we take a practical stance and ask, when information is limited, which variables are best? Imagine a forecaster who "applies for the job" and takes a set of relevant tests (i.e., dispositional variables). We might also know the forecaster's working conditions (i.e., situational variables). We could then "hire" the forecaster and monitor work habits while "on the job" (i.e., behavioral variables). Which type of variables best identifies those who make the most accurate forecasts?

Method

The forecasting tournament was conducted over 2 years, with each year lasting about 9 months. The first period ran from September 2011 to April 2012, and the second one ran from June 2012 to April 2013. We recruited forecasters from professional societies, research centers, alumni associations, and science blogs, as well as word of mouth. Entry into the tournament required a bachelor's degree or higher and completion of a battery of psychological and political knowledge tests that took approximately 2 hr. Participants were largely U.S. citizens (76%) and males (83%), with an average age of 36. Almost two thirds (64%) had some postgraduate training.

Design

In Year 1, participants were randomly assigned to a 3 × 3 factorial design of Training (probabilistic-reasoning training, scenario training, and no training) × Group Influence (independent forecasters, crowd-belief forecasters, and team forecasters).[1] Training consisted of instructional modules. Probabilistic-reasoning training, consisted of tips about what information to look for and how to avoid judgmental biases. Scenario training taught forecasters to generate new futures, actively entertain more possibilities, use decision trees, and avoid overconfidence.

Group influence had three levels. We staked out a continuum with independent forecasters who worked alone at one end, and interdependent forecasters who worked in teams of approximately 15 people and interacted on a website at the other end. We also included a compromise level (crowd belief forecasters) in which forecasters worked alone, but had knowledge of others' beliefs. The benefit of this approach is that forecasters had access to a potentially potent signal—the numerical distribution of the crowd's opinions, but they could avoid the potential costs of social interaction, such as mindless "herding" or free-riding.

Those in team conditions also received training in how to create a well-functioning group. Members were encouraged to maintain high standards of proof and seek out high-quality information. They were given strategies for explaining their forecasts to others, offering constructive critiques, and building an effective team. Members could offer rationales for their thinking and critiques of others' thinking. They could share information, including their forecasts. But there was no systematic display of team members' predictions. Instructions, training, tests, and forecasting questions are available in the online supplemental materials.

At the end of Year 1, we learned that probabilistic-reasoning training was the most effective instructional module, and teamwork was the most effective form of group interaction. We decided to replicate only the most effective experimental methods from Year 1 using a reduced 2 × 2 factorial design of Training (probabilistic-reasoning training and no training) by Group Influence (team vs. independent forecasters).

We added another intervention—the tracking of top performers. We skimmed off the top 2% of forecasters and put them in five elite teams. This small group of forecasters had a distinctly different experience from others (see Mellers, Stone, et al., 2014) and therefore was not included in the analyses. However, results did not change when these forecasters were included.

---

[1] Results from the prediction market are discussed elsewhere because individual accuracy measures (in the form of Brier scores) cannot be computed (Atanasov et al., 2014).

The smaller experimental design of Year 2 required reassignment of forecasters from Year 1 conditions that were not continued in Year 2 (i.e., crowd belief forecasters and scenario training). Assignment of forecasters proceeded as follows: (a) if a Year 1 condition remained in Year 2, forecasters stayed in that condition; (b) crowd-belief forecasters were randomly assigned to independent or team conditions; (c) scenario trainees were randomly assigned to no training or probabilistic-reasoning training. Our analyses in this article focus only on the effectiveness of two situational variables: probabilistic-reasoning training and teamwork.

Questions

Questions were released throughout the tournament in small batches, and forecasters received 199 questions over 2 years. Questions covered political-economic issues around the world and were selected by the IARPA, not by the research team. Questions covered topics ranging from whether North Korea would test a nuclear device between January 9, 2012, and April 1, 2012, to whether Moody's would downgrade the sovereign debt rating of Greece between October 3, 2011, and November 30, 2011. Questions were open for an average of 109 days (range = 2 to 418 days).

Participants were free to answer any questions they wished within a season. There were no constraints on how many, except that payment for the season required that participants provide forecasts for at least 25 questions. One question asked, "Will there be a significant outbreak of H5N1 in China in 2012?" The word "significant" was defined as at least 20 infected individuals and five casualties. The question was launched on February 21, 2012, and was scheduled to close on December 30, 2012. If the outcome occurred prior to December 30, the question closed when the outcome occurred. Forecasters could enter their initial forecast or update their prior forecast until the resolution of the outcome.

One hundred fifty questions were binary. One binary question, released on November 8, 2011, asked, "Will Bashar al-Assad remain President of Syria through January, 31 2012?" Answers were "yes" or "no." Some questions had three to five outcomes. A three-option question, released on October 4, 2011, asked, "Who will win the January 2012 Taiwan Presidential election?" Answers were "Ma Ying-jeou," "Tsai Ing-wen," or "neither." Some questions had ordered outcomes. One with four ordered outcomes asked, "When will Nouri al-Maliki resign, lose confidence vote, or vacate the office of Prime Minister of Iraq?" Answers were "between July 16, 2012 and Sept 30, 2012," "between Oct 1, 2012 and Dec, 31 2012," between "Jan 1, 2013 and Mar 31, 2013," or "the event will not occur before April 1, 2013." Finally, another set was conditional questions, typically having two antecedents and two outcomes. One question asked,

Before March 1, 2014, will North Korea conduct another successful nuclear detonation (a) if the United Nations committee established pursuant to Security Council resolution 1718 adds any further names to its list of designated persons or entities beforehand or (b) if the United Nations committee established pursuant to Security Council resolution 1718 *does not* add any further names to its list of designated persons or entities beforehand?

Answers to both possibilities were "yes" or "no."

All forecasters were given a brief Brier score tutorial and learned that their overarching goal was to minimize Brier scores.

Feedback given to forecasters during the tournament included Brier scores, averaged over days within a question and across questions. Forecasters were incentivized to answer questions if they believed they knew more than the average forecaster in their condition. If they did not answer a question, they received the average Brier score that others in their condition received on that question. Whenever a question closed, we recalculated individual Brier scores, thereby providing forecasters with constant feedback.

Brier scores used in our analyses *did not* include the average scores of others if a forecaster did not answer a question. Instead, we simply computed the Brier score for each forecast made by a participant and averaged over Brier scores if that participant made multiple forecasts on a given question. Inclusion of averages from others would simply have reduced differences among individuals.

Measures

Prior to each forecasting season, we administered a battery of psychological tests. Intelligence was measured by four scales. Inductive pattern recognition was assessed by a short form of the Ravens Advanced Progressive Matrices (Ravens APM; Bors & Stokes, 1998), which circumvents cultural or linguistic knowledge by testing spatial problem-solving skills. Cognitive control was measured by the three-item Cognitive Reflection Test (CRT; Frederick, 2005) and the four-item extension of the CRT (Baron, Scott, Fincher, & Metz, 2014), with questions such as, "All flowers have petals. Roses have petals. If these two statements are true can we conclude that roses are flowers?" Mathematical reasoning was measured by a three-item Numeracy scale. The first item came from Lipkus, Samsa, and Rimer (2001), and the second two were from Peters et al. (2006).

We had three measures of open-mindedness. The first was a seven-item actively open-minded thinking test (Haran et al., 2013) that used a 7-point rating scale (1 = *completely disagree* and 7 = *completely agree*). Actively open-minded thinking predicts both persistence in information searches and accuracy in estimating uncertain quantities (Haran et al., 2013). The second was an 11-item Need-For-Closure scale (Kruglanski & Webster, 1996; Webster & Kruglanski, 1994). Responses were made on the same 7-point rating scale. The third was a single question: "In a famous essay, Isaiah Berlin classified thinkers as hedgehogs and foxes: The hedgehog knows one big thing and tries to explain as much as possible using that theory or framework. The fox knows many small things and is content to improvise explanations on a case-by-case basis. When it comes to making predictions, would you describe yourself as more of a hedgehog or more of a fox?" Responses were made on a 5-point rating scale (1 = *very much more fox-like*; 5 = *very much more hedgehog-like*).

Political knowledge was assessed by two true–false tests of current affairs, one given each year. The first was a 35-item test with items such as "Azerbaijan and Armenia have formally settled their border dispute." The second was a 50-item test with items such as "India's liberalization reforms now allow for 100% Foreign Direct Investment (FDI) stake in ventures" or "The GINI coefficient measures the rate of economic expansion."

Participants

Year 1 began with 1,593 survey participants who were randomly assigned to nine conditions, with an average of 177 per condition. Attrition was 7%. Year 2 started with 943 respondents. Attrition in Year 2 fell to 3%, perhaps because most participants were returnees and familiar with the task. We wanted forecasters who made many predictions and for whom we could get stable estimates of forecasting ability. To that end, we used only 743 forecasters who participated in both years of the tournament and had made at least 30 predictions.

Payments

Forecasters who met the minimum participation requirements received $150 at the end of Year 1 and $250 at the end of Year 2, regardless of their accuracy. Those who returned from Year 1 received a $100 retention bonus. Forecasters also received status rewards for their performance via leader boards that displayed Brier scores for the top 20 forecasters (10%) in each condition and full Brier score rankings of teams. Team Brier scores were the median of scores for individuals within a team.

Results

Individual Differences and Consistency Over Time

Our primary goal was to investigate the consistency and predictability of individual forecasting accuracy, defined as a Brier score averaged over days and questions. Participants attempted an average of 121 forecasting questions. Figure 1 shows the distribution of overall Brier scores, revealing a wide range of forecasting abilities.

A common approach in studies of accuracy is to compare intuitive predictions with simple benchmarks, such as the score one would receive by assigning a uniform distribution over outcomes for all questions. The raw Brier score would be 0.53, on a scale ranging from 0 (best) to 2 (worst). The average raw Brier score of our participants was 0.30, much better than random guessing, $t(741) = -61.79$, $p < .001$. Overall, forecasters were significantly better than chance.
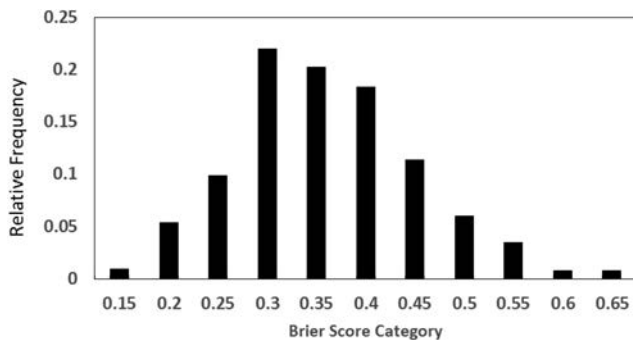
Figure 1. Distribution of Brier scores over forecasters plotted against category bins of size .049. The category labeled .15 refers to Brier scores between .10 and .149.

An alternative measure of forecast accuracy is the proportion of days on which forecasters' estimates were on the correct side of 50%. This measure is calculated by counting the days on which forecasters were active and correct (i.e., they placed estimates of 51% or above for events that occurred and 49% or below for events that did not occur). For multinomial questions, forecasts were considered correct if the realized option was associated with the highest probability. We counted all days after the first forecast was placed, and we carried forward estimates until a participant updated his or her forecast or the question closed. A perfect score would be 100%, and a chance score for binary questions would be 50%. For all questions in the sample, a chance score was 47%. The mean proportion of days with correct estimates was 75%, significantly better than random guessing for binary questions, $t(740) = 79.70$, $p < .001$.

Figure 2 shows the distribution. The correlation between mean Brier score and proportion of correct days was very high, $r = .89$, $t(741) > 54.25$, $p < .0001$. The proportion of correct days is just another way to illustrate accuracy. All subsequent analyses focus on Brier scores, unless otherwise specified.

Next we turn to the question of consistency, but first we make a simple adjustment to the accuracy metric. Forecasters selected their own questions, and this feature of the experimental design allowed people to get a better Brier score if they could select events that were easier to predict. To handle this problem, we standardized Brier scores within questions. Standardization minimizes differences in difficulty across questions and allowed us to focus on relative, rather than absolute, performance. If accuracy were largely attributable to luck, there would be little internal consistency in Brier scores over questions. However, Cronbach's alpha (a gauge of the internal consistency of Brier scores on questions) was 0.88, suggesting high internal consistency. Figure 3 illustrates how the best and worst forecasters differed in skill across time. We constructed two groups based on average standardized Brier scores after the first 25 questions had closed and forecasters had attempted an average of 15 questions. The black and gray lines represent the 100 best and worst forecasters, respectively. Figure 3 tracks their progress over time; average Brier scores are presented for each group on 26th to the 199th question, plotted against the order that questions closed.[2] Using relatively little initial knowledge about forecaster skill, we could identify differences in performance that continued for a period of 2 years. These groups differed by an average of 0.54—more than half a standard deviation—across the tournament. If we could identify good forecasters early, there was a reasonable chance they would be good later.

There are several ways to look for individual consistency across questions. We sorted questions on the basis of response format (binary, multinomial, conditional, ordered), region (Eurzone, Latin America, China, etc.), and duration of question (short, medium, and long). We computed accuracy scores for each individual on each variable within each set (e.g., binary, multinomial, conditional, and ordered) and then constructed correlation matrices. For all three question types, correlations were positive; an individual who scored high on binary questions tended to score higher on

[2] For each forecaster, we averaged predictions over days, regardless of the day on which the prediction was made.
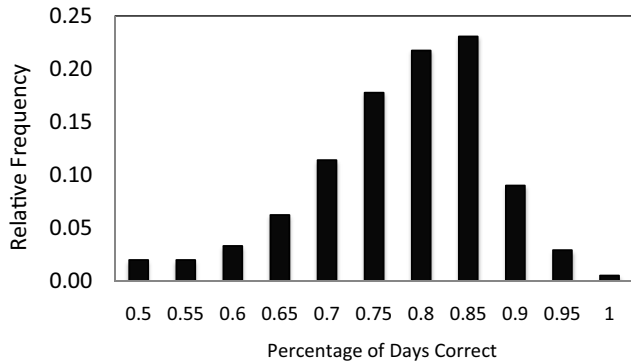
95

Figure 2. Distribution of days on which estimates were on the correct side of 50% plotted against bins of size .049. The category labeled 0.55 refers to forecasters who were correct for 55% to 59.9% of the days on which they had active forecasts.

Table 1
Descriptive Statistics

	Mean	SD	Min	Max	Alpha
Standardized Brier score	0	0.29	−0.57	1.32	0.88
Ravens	8.56	2.43	0	12	0.71
Cog Reflection Test	2.1	0.97	0	3	0.55
Extended Cog Reflection Test	3.37	1.03	0	4	0.70
Numeracy	2.71	0.53	0	3	0.11
Actively open-minded think	5.91	0.6	4	7	0.65
Need for closure	3.34	0.58	1.45	5.09	0.55
Fox vs. Hedgehog	3.82	0.54	1.9	6	
Political knowledge Year 1	28.79	3.07	18	35	0.53
Political knowledge Year 2	36.5	4.64	19	48	0.64
Number predictions per Q	1.58	0.77	1	6.33	
Number of questions	21	51	13	199	
Deliberation time (s)	3.6	0.71	2	4	

Note. Min = minimum; Max = maximum; SD = signaled donation; Cog = cognition.
[a] Cronbach's alpha for Brier scores is calculated at the question level. All other alphas are calculated at the participant level. Alphas are not reported for scales with three or fewer items and for behavioral variables.

multinomial questions. Then we conducted factor analyses. For each question type, a large proportion of the variance was captured by a single factor, consistent with the hypothesis that one underlying dimension was necessary to capture correlations among response formats, regions, and question duration.

Dispositional Variables

Are individual dispositional variables of intelligence, openmindedness, and political knowledge associated with forecasting accuracy? Table 1 shows means and variances of predictor variables. The mean score on the short version of the Ravens APM was 8.56 out of 12, which was considerably higher than 7.07, the mean score of a random sample of first-year students at the University of Toronto (Bors & Stokes, 1998). Our forecasters scored 2.10 on the CRT, virtually equivalent to 2.18, the average score of MIT students (Frederick, 2005). The extended CRT and the numeracy items had no comparable norms.

Table 1 also presents reliability estimates, when applicable. Values were .71 for the Ravens APM, .55 for the three-item CRT,
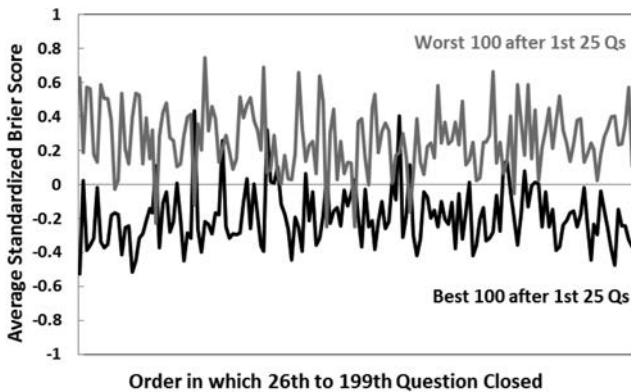
Figure 3. Average scores for 100 best forecasters (black) and 100 worst forecasters (gray) defined after the close of the first 25 questions. The *x*-axis represents the order in which questions closed throughout the rest of the tournament. Differences defined early in the tournament remained for 2 years, as seen by the space between the two lines.

.70 for the extended CRT, .11 for Numeracy, .65 for actively open-minded thinking, .55 for Need for Closure, .53 for the Year 1 political knowledge test, and .64 for the Year 2 test. The most troubling reliability estimate was that of Numeracy. Most people found it very easy; the percentages correct on the three items were 93%, 92%, and 86%.

Table 2 shows all possible pairwise correlations. Three of the four measures of intelligence—the Ravens APM, the CRT, and the extended CRT—were significantly correlated with standardized Brier score accuracy: Correlations were −.23, −.15, and −.14, respectively, $t(741) = −6.38, p < .001, t(741) = −4.17, p < .001$, and $t(599) = −3.56, p = .001$.[3] Lower Brier scores indicate higher accuracy, so negative correlations mean that greater accuracy is associated with higher intelligence scores. We combined these variables into a single factor using the first dimension of a principal axis factor analysis. The correlation between standardized Brier scores and factor scores was −.22, $t(741) = −5.51, p < .001$. Greater intelligence predicted better performance, consistent with our first hypothesis.

Next we turn to open-mindedness and examined whether three measures—actively open-minded thinking, need for closure, and hedgehog–fox orientation—predicted forecasting accuracy. The average score on actively open-mindedness was 5.91, relatively high on a 1 to 7 response scale. The average need for closure score was 3.34, close to the middle of the scale, and the average fox–hedgehog response was 3.82, which indicated that, on average, forecasters viewed themselves as slightly more hedgehoglike. More actively open-minded participants had less need for closure, $r = −.20, t(742) = −5.56, p < .001$, and more hedgehoglike participants had more need for closure, $r = .24, t(742) = −6.73, p < .001$. Only one of the measures, actively open-minded thinking, was significantly related to standardized

---

[3] The correlation between reaction time on the Ravens APM test and forecasting accuracy was also significant; those who spent more time on the Ravens APM test also tended to be better forecasters, $r = −0.12, t(741) = 3.29, p < .001$.

Table 2
Correlations Among Dispositional, Situational, and Behavioral Variables

	Std BS	Ravens	CRT	ExCRT	Numeracy	AOMT	Nfclo	Foxhed	PKY1	PKY2	Train	Teams	Npredq	Nquest
Std BS	1.00													
Ravens	**−0.23**	1.00												
CRT	**−0.15**	**0.38**	1.00											
ExCRT	**−0.14**	**0.34**	**0.39**	1.00										
Numeracy	−0.09	**0.16**	**0.12**	**0.14**	1.00									
AOMT	**−0.10**	**0.10**	0.08	**0.22**	0.09	1.00								
Nfclo	0.03	0.03	−0.02	−0.05	0.10	**−0.20**	1.00							
Foxhed	0.09	0.05	0.01	0.02	0.02	−0.09	**0.24**	1.00						
PKY1	**−0.18**	0.05	0.06	0.08	0.03	**0.13**	−0.03	−0.03	1.00					
PKY2	**−0.20**	0.08	0.08	0.12	0.01	**0.12**	−0.07	−0.09	**0.59**	1.00				
Train	**−0.17**	0.02	−0.01	0.06	0.06	0.05	−0.03	0.02	0.04	0.02	1.00			
Teams	**−0.30**	−0.04	0.01	0.01	0.04	0.04	−0.05	−0.06	0.02	0.02	0.00	1.00		
Npredq	**−0.49**	**0.17**	**0.12**	**0.12**	0.09	0.05	0.01	−0.02	**0.14**	**0.19**	0.08	**0.11**	1.00	
Nquest	0.07	−0.02	0.04	0.04	−0.05	−0.02	0.06	0.07	0.07	0.07	−0.02	**−0.17**	**0.23**	1.00
Del time	**−0.30**	0.08	−0.09	−0.05	0.03	0.05	−0.09	−0.08	−0.01	0.05	0.06	**0.28**	0.15	**−0.25**

Note. Bold values are significant at the .001 level. Std BS = Standardized Brier score; CRT = cognitive reflection test; ExCRT = extended cognitive reflection test; AOMT = actively open-minded thinking; Nfclo = need for closure; Foxhed = fox versus hedgehog; PKY1 = political knowledge year 1; PKY2 = political knowledge year 2; train = training; Npredq = number of predictions per question; Nonquest = number of questions answered.

Brier score accuracy, $r = −.10$, $t(742) = −2.51$, $p < .01$. Thus, we had partial support for the second hypothesis that flexible and open-minded cognitive styles predicted forecasting accuracy.

The third hypothesis stated that political knowledge would predict Brier score accuracy. Percent correct scores on these true–false questions were 82% and 76%, respectively. Test scores were highly correlated with forecasting accuracy, $r = .59$, $t(742) = −19.91$, $p < .001$. We have no comparable norms, but the obvious difficulty of the tests makes these scores seem high. The correlation between political knowledge scores in Years 1 and 2 and relative forecasting accuracy was −.18 and −.20, respectively, $t(741) = −4.85$, $p < .001$, and $t(648) = −5.06$, $p < .001$. Again, we constructed a single measure of content knowledge using the first factor of a principal axis factor analysis. The correlation between relative accuracy and these factor scores was −.22, $t(599) = −5.52$, $p < .001$. Political knowledge was predictive of forecasting accuracy, consistent with our third hypothesis.

Earlier, we mentioned the debate about the role of intelligence versus deliberative practice in the development of expertise. One hypothesis was that the correlation between intelligence and performance would be strongest early on and gradually disappear as forecasters engage in more deliberate practice. In past studies, the Ravens APM is a common measure of cognitive ability (e.g., Ruthsatz, Detterman, Griscom, & Cirullo, 2008). We correlated Ravens APM scores with accuracy early and late in the tournament. "Early" and "late" are vague terms, so we used multiple definitions, including the first 50 and last 50 questions, the first 40 and last 40 questions, and the first 30 and last 30 questions (out of 199).

Correlations between the Ravens APM scores and accuracy based on the first and last 50 questions representing early and late stages were −.22 and −.10, respectively. The relationship between intelligence and performance was stronger earlier than it was later, $t(624) = 2.57$, $p < .01$. Similar results occurred with cutoffs of 30 and 40 questions. This analysis is based on only 2 years of deliberative practice, not 10,000 hr (i.e., the length of

deliberative practice necessary to achieve expertise, according to Ericsson et al., 1993). Nonetheless, the difficulty of the questions and the breadth of topics suggest that one would do poorly in our tournament without some degree of sustained effort and engagement.

Situational Variables

Mellers, Ungar, et al. (2014) showed that forecasters who were trained in probabilistic reasoning and worked together in teams were more accurate than others. However, effect sizes in the form of correlations were not presented. Table 2 shows that relative accuracy and training in probabilistic reasoning had a correlation of −.17, $t(741) = −4.56$, $p < .001$. In addition, relative accuracy of team participation had a correlation of −.30, $t(741) = −8.55$, $p < .001$. These findings illustrate how the prediction environment influences forecaster accuracy, independent of all else.

To test the fourth hypothesis—dispositional variables predict forecasting skill beyond situational variables—we conducted a multiple regression predicting standardized Brier scores from intelligence factor scores, actively open-minded thinking, political knowledge factor scores, probability training, and teamwork. The latter two variables were dummy coded. The multiple correlation was .43, $F(5, 587) = 26.13$, $p < .001$. Standardized regression coefficients for two of the three dispositional variables—intelligence and political knowledge—were statistically significant, −0.18 and −0.15, $t(587) = −4.65$, $p < .001$, and $t(587) = −3.89$, $p < .001$. Intelligence and political knowledge added to the prediction of accuracy beyond the situational variables.

Behavioral Variables

Effort and engagement manifest themselves in several ways, including the number of predictions made per question (belief updating), the time spent before making a forecast (deliberation time), and the number of forecasting questions attempted. The

average number of predictions made per forecasting question was 1.58, or slightly more than 1.5 forecasts per person per question. Deliberation time, which was only measured in Year 2, was transformed by a logarithmic function (to reduce tail effects) and averaged over questions. The average length of deliberation time was 3.60 min, and the average number of questions tried throughout the 2-year period was 121 out of 199 (61% of all questions). Correlations between standardized Brier score accuracy and effort were statistically significant for belief updating, $-.49$, $t(740) = -15.29$, $p < .001$, and deliberation time, $-.30$, $t(694) = -8,28$, $p < .001$, but not for number of forecasting questions attempted. Thus, two of three behavioral variables predicted accuracy, in partial support of the fourth hypothesis.

The fifth hypothesis stated that behavioral variables would contribute to the predictability of skill over and beyond dispositional and situational variables. To test this hypothesis, we conducted a multiple regression predicting standardized Brier scores from belief updating, deliberation time, and number of questions attempted, as well as intelligence factor scores, actively open-minded thinking, political knowledge factor scores, training, and teamwork. The multiple correlation was .64, $F(8, 581) = 52.24$, $p < .001$. Behavioral variables with significant standardized regression coefficients were belief updating, -0.45, $t(581) = -12.89$, $p < .001$, and deliberation time, -0.13, $t(581) = -3.69$, $p < .001$. Results were thus consistent with the fifth hypothesis that behavioral variables provide valuable independent information, in addition to dispositional and situational variables.

Structural Equation Model

To further explore interconnections among these variables, we used a structural equation model that enabled us to synthesize our results in a single analysis, incorporate latent variables, and perform simultaneous regressions to test our hypotheses. The initial model only included variables that were significant predictors of standardized Brier score accuracy on their own. These variables included two latent constructs (political knowledge and intelligence), open-mindedness, probabilistic training, teamwork, belief updating, and deliberation time.

We then conducted mediation analyses to see whether behavioral variables mediated the relationship between dispositional variables and accuracy, and situational variables and accuracy.[4] Results are shown in Table 3. For simplicity, we removed pathways whose inclusion neither improved nor changed the model fit significantly, and ultimately we arrived at the model in Figure 4. Yellow ovals are latent dispositional variables, yellow rectangles are observed dispositional variables, pink rectangles are experimentally manipulated situational variables, and green rectangles are observed behavioral variables.

Dispositional variables of political knowledge and intelligence had direct and indirect effects on Brier score accuracy. Better forecasters had greater knowledge and ability, and part of that relationship was accounted for by greater belief updating. Actively open-minded thinking, our best cognitive-style predictor, had only direct effects on accuracy. Situational variables of teamwork and training had direct effects on accuracy, but teamwork also had indirect effects. Those in teams did better than those who worked

alone, especially when they updated their beliefs often and deliberated longer.

The structural equation modeling required the fit of five simultaneous regressions shown in Table 4. In one regression, the latent variable of fluid intelligence was predicted from the Ravens APM, the CRT, the extended CRT, and Numeracy. The coefficient for the Ravens APM was set to 1.0, and others were estimated relative to it. In the next regression, the latent variable of political knowledge was predicted from tests in Years 1 and 2. In the third regression, belief updating was predicted by the two latent variables and teamwork, and in the fourth, deliberation time was predicted from teamwork. The last regression was the prediction of forecaster accuracy from fluid intelligence, political knowledge, intelligence, actively open-minded thinking, teams, probability training, belief updating, and deliberation time. Coefficients for these regressions with observed variables, along with standard errors, Z statistics, p values, and bootstrapped confidence intervals (when appropriate) are provided in Table 4.

This model provided a reasonable account of relative forecaster accuracy. The Tucker-Lewis Reliability Index was 0.92, and the comparative fit index was 0.95. The root mean squared error of approximation was 0.04. In addition to fitting a model to individuals' average relative accuracy scores, we fit models to their first and last forecasts. Using only first forecasts, the effect of belief updating disappeared, as expected, but the remaining associations remained strong. Using only last forecasts, the effect of belief updating increased, as we would expect, and all other associations did not change.

One way to test the model's stability is to calculate confidence intervals on coefficients using bootstrapping methods. These intervals are presented in Table 4, and all of these exclude zero, supporting the validity of the relationships. Another way to test the model's stability is to conduct cross validations which examine the extent to which the model would have fit with different subsets of questions or of participants. For the cross-validation of questions, we randomly assigned each question to one of 10 folds (or subsets of questions) with equal numbers of questions in each fold. In each validation, we used 90% of the questions, computed average standardized Brier scores for each participant, and refit the structural model. We repeated this process for each fold and examined the distributions of resulting coefficients over all 10 validations. Coefficients for all of the parameters were consistent with the full model.[5] We conducted a similar cross-validation analysis using subsets of participants, and again, coefficients for all of the parameters were consistent with the full model in each of the 10 validation sets.

Which Types of Variables Best Predict Relative Accuracy?

Table 5 shows multiple correlations and tests of nested comparisons. Using only dispositional information (intelligence, political

---

[4] We conducted mediation analyses to determine which of our predictors of accuracy might be mediated by effort. We used those results to determine which pathways to include in the structural equation model. Results of the mediation analyses are summarized in Table 3 and Figure 3.

[5] Consistency refers to all coefficients in the validation models maintaining significance ($p \leq .05$), and similar magnitude to the full model, across all 10 cross-validation folds.

Table 3
Indirect and Total Contributions in Mediation Analyses

Independent	Mediator	Dependent	Indirect	*p* value	Total	*p* value
IQ	Bel updating	Std Br score	−0.18	<0.001	−0.54	<0.001
Pol know	Bel updating	Std Br score	−0.16	<0.001	−0.36	<0.001
AOMT	Bel updating	Std Br score	−0.03	0.10	−0.12	<0.001
Train	Bel updating	Std Br score	−0.07	0.03	−0.32	<0.001
Team	Bel updating	Std Br score	−0.08	0.01	−0.53	<0.001
Team	Deliberation time	Std Br score	−0.07	<0.001	−0.29	<0.001

Note. Independent refers to the independent variable, and Dependent is the dependent variable. Indirect is the product of the correlation between the independent variable and the mediator and that between the mediator and the dependent variable. Total is the indirect plus the direct effects, where the direct effect is the correlation between the independent and dependent variable. IQ = intelligence; Pol know = political knowledge; AOMT = actively open-minded thinking; Train = probablistic reasoning training; Teams = on a collaborative team.

knowledge, and actively open-minded thinking), the multiple correlation for accuracy was .31, close to the famous .3 barrier in personality research, which is sometimes supposed to be the upper bound of the validities on many personality scales. Using only situational variables that describe the conditions under which forecasters worked, the multiple correlation was .34, similar in magnitude to that obtained from the dispositional variables, replicating the more general conclusion of Funder and Ozer (1983) that many individual difference effects and situational manipulations appear to have similar effect-size upper bounds.

Adding behavioral information on forecaster belief updating and deliberation time, predictability improved and the multiple correlation rose to .54. Not surprisingly, it is harder to identify the best forecasters from abstract dispositional constructs than it is from specific features of their behavior in the prediction environment. Nonetheless, as we saw in the structural model, and confirm here, the best model uses dispositional, situational, and behavioral variables. The combination produced a multiple correlation of .64. Each model provided a better fit as more variables were included. *F* tests for the nested model deviance showed that larger models provided a significantly better fit than their simpler counterparts. The person, the situation, and related behavior all contribute to identifying superior geopolitical forecasting performance.[6]

Figure 4. Structural equation model with standardized coefficients. See the online article for the color version of this figure.

Discussion

We examine the geopolitical forecasting skills of participants in a 2-year tournament. Drawing on diverse literatures, we tested three categories of hypotheses about the psychological drivers of judgmental accuracy: (a) dispositional variables measured prior to the tournament, such as cognitive abilities and styles; (b) situational variables that were experimentally manipulated prior to the competition, such as training in probabilistic reasoning and collaborative engagement in teams organized around self-critical, epistemic norms; and (c) behavioral variables measured during the competition, such as the willingness to revisit and update older beliefs.

The best forecasters scored higher on both intelligence and political knowledge than the already well-above-average group of forecasters. The best forecasters had more open-minded cognitive styles. They benefited from better working environments with probability training and collaborative teams. And while making predictions, they spent more time deliberating and updating their forecasts.

These predictors of accuracy proved robust over several subsets of questions. With few exceptions, variables that captured the best forecasters overall predicted accuracy across different temporal periods within a question (early, middle, and late), across questions that differed in length (short, medium, and long durations), and across questions that differed in mutability (close calls vs. clear-cut outcomes).

We offer a structural equation model to capture the interrelationships among variables. Measures that reflected behavior within tournaments served as mediators. Belief updating partly mediated the relationship between intelligence and accuracy, between political knowledge and accuracy, and teamwork and accuracy. Deliberation time mediated the relationship between teamwork and accuracy. This association has different causal interpretations. Those with more political knowledge and greater intelligence might have enjoyed the task more—and that enjoyment may have motivated engagement. Alternatively, once forecasters became more engaged, they may have become more politically knowledgeable. Furthermore, those who worked in teams may also have been
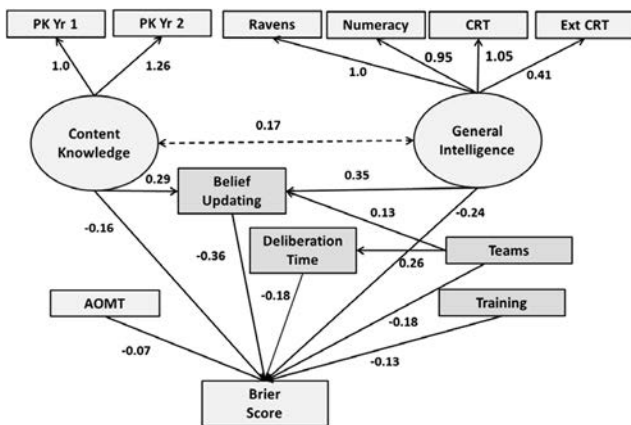
---

[6] These correlations were fit directly to the data. Cross-validated correlations would obviously be smaller.

Table 4
Regressions in Structural Equation Model

Regressions	Estimate	Std error	z	p	Bootstrap CI	
1. Intelligence (FS)						
CRT	1.01	0.12	8.73	0.00		
Ex-CRT	0.89	0.10	8.63	0.00		
Numeracy	0.35	0.08	4.17	0.00		
2. Pol knowledge (FS)						
Pol know Year 2	1.22	0.22	5.49	0.00		
3. Belief updating						
Teams	0.12	0.04	2.91	0.00	0.04	0.21
Pol know	0.30	0.08	3.97	0.00	0.15	0.44
Intelligence	0.34	0.09	3.86	0.00	0.17	0.51
4. Deliberation time						
Teams	0.27	0.04	7.18	0.00	0.20	0.34
5. Std Brier score						
Belief updating	−0.35	0.03	−10.72	0.00	−0.42	−0.29
Deliberation time	−0.17	0.04	−4.82	0.00	−0.24	−0.10
Pol know	−0.16	0.06	−2.80	0.01	−0.28	−0.05
Intelligence	−0.24	0.07	−3.54	0.00	−0.38	−0.11
Teams	−0.19	0.03	−5.65	0.00	−0.26	−0.12
P train	−0.13	0.03	−4.15	0.00	−0.20	−0.07
AOMT	−0.07	0.03	−2.13	0.03	−0.13	−0.01

Note. FS refers to factor score. The regression coefficients for Ravens and the political knowledge test Year 1 were set to 1.0 in the first and second regression, respectively. CRT = cognitive reflection test; ExCRT = extended cognitive reflection test; Pol know = political knowledge; P train = probablistic reasoning training; AOMT = actively open-minded thinking.

motivated to do well for the sake of the group, which could also produce greater updating and ultimately greater accuracy.

Actively open-minded thinking predicted accuracy but less consistently than other variables. This cognitive style is associated

fewer cognitive errors, including the myside bias, biased argument evaluation, superstitious thinking, and outcome bias (Stanovich & West, 1997, 1998, 2007). Laboratory evidence shows that actively open-minded thinking predicts accuracy of estimates of uncertain

Table 5
Predicting Overall Forecasting Accuracy From Different Types of Variables

	Multiple R	F	Sig
Variable type			
Dispositional (2 latent, 1 observed variable)	0.31	21.12	p < .001
Situational (2 variables)	0.34	49.44	p < .001
Behavioral (2 variables)	0.54	142.52	p < .001
Dis + Sit (2 latent, 3 observed variables)	0.45	30.17	p < .001
Dis + Beh (2 latent, 3 variables)	0.58	60.06	p < .001
Sit + Beh (4 variables)	0.58	89.43	p < .001
Dis + Sit + Beh (7 variables)	0.64	52.99	p < .001

	Delta SS	Delta DF	F	Sig
Nested comparisons				
Dis vs. Dis + Sit	64.76	2	40.38	p < .001
Sit vs. Dis + Sit	56.16	7	10.01	p < .001
Dis vs. Dis + Beh	154.51	3	79.41	p < .001
Beh vs. Dis + Beh	22.75	7	5.01	p < .001
Sit vs. Beh + Sit	148.88	3	77.79	p < .001
Beh vs. Sit + Beh	25.72	2	25.72	p < .001
Dis + Sit vs. Dis + Sit + Beh	115.51	3	63.51	p < .001
Dis + Beh vs. Dis + Sit + Beh	25.77	2	21.25	p < .001
Sit + Beh vs. Dis + Sit + Beh	22.80	7	5.37	p < .001

Note. Dispositional variables refer to principle factor scores for general intelligence (Ravens, CRT, exCRT, numeracy) and political knowledge (Year 1 and Year 2 tests), as well as actively open-minded thinking. Situational variables refer to teams and training. Behavioral variables are the average number of forecasts made per question and average time spent deliberating about a question. The F test for nested comparisons tests the probability that the difference in sum of squares between the smaller and larger models is >0 by comparing the mean square error of the smaller model to the residual sum of squares for the larger one.

quantities (Haran et al., 2013), but no prior studies have demonstrated an association between actively open-minded thinking and forecasting performance on real-world problems. We believe this cognitive style translates into more accurate political forecasts through its association with better norms of thinking.

Kahneman and Klein (2009) argued that for any type of skill to develop, two conditions must be present: (a) an environment with sufficient deterministic stability to permit learning, and (b) opportunities for practice. Skill development occurs to the extent that people care enough to engage in deliberative rehearsal (Ericsson, 2006). Our forecasters received constant feedback in the form of Brier scores and leaderboard rankings. They had many chances to learn; there were 199 questions over a period of 2 years, and, on average, forecasters made predictions for 121 of them. These conditions enabled a process of learning-by-doing and help to explain why some forecasters achieved far-better-than-chance accuracy.

In the real world, intelligence analysts use data from diverse sources. They frequently make nonnumerical forecasts that are vague and hard to score for accuracy, so feedback is often absent. Intelligence analysts shift their response thresholds depending on the cost of the errors. That is, they are likelier to say "signal" when the costs of a miss are high, and they are likelier to say "noise" when the costs of a false alarm are high. Although our forecasters knew that the Brier-score costs of errors were symmetric, the real world is much more complicated.

Analysts also operate under bureaucratic-political pressure—and are tempted to respond to previous mistakes by shifting their response thresholds. They are likelier to say "signal" when recently accused of underconnecting the dots (i.e., 9/11) and to say "noise" when recently accused of overconnecting the dots (i.e., weapons of mass destruction in Iraq). Tetlock and Mellers (2011) describe this process as accountability ping-pong. One escape from this cycle is to make a public organizational commitment to using tournaments to monitor long-term accuracy, not just avoidance of the most recent mistake (McGraw, Todorov, & Kunreuther, 2011).

Our study is the first to keep score and track categories of variables that predict performance in the politically sensitive domain of intelligence analysis. The study demonstrates the value of tournaments in identifying top forecasters. If the National Academy of Science Report on improving intelligence analysis is correct about the power of measuring accuracy and providing feedback to boost performance, tournaments should become a regular feature of research on improving accuracy in organizational systems and evaluating the track records of intelligence analysts.

References

Almond, G. A., & Genco, S. J. (1977). Clouds, clocks, and the study of politics. *World Politics, 29,* 489–522. http://dx.doi.org/10.2307/2010037

Arkes, H. R. (2001). Overconfidence in judgmental forecasting. In S. Armstrong (Ed.), *Principles of forecasting* (pp. 495–515). New York, NY: Springer.

Atanasov, P., Rescober, P., Stone, E., Servan-Schreiber, E., Tetlock, P., Ungar, L., & Mellers, B. (2014). *Crowd forecasting with prediction polls and prediction markets.* Manuscript submitted for publication.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). New York, NY: Cambridge University Press.

Baron, J., Scott, S. E., Fincher, K., & Metz, S. E. (2014). Why does the Cognitive Reflection Test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory & Cognition.* Advance online publication. http://dx.doi.org/10.1016/j.jarmar.2014.09.003

Bors, D. A., & Stokes, T. L. (1998). Raven's Advanced Progressive Matrices: Norms for first-year university students and the development of a short form. *Educational and Psychological Measurement, 58,* 382–398. http://dx.doi.org/10.1177/0013164498058003002

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review, 78,* 1–3. http://dx.doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2

Bueno de Mesquita, B. (2009). *The predictioneer's game: Using the logic of brazen self-interest to see and shape the future.* New York, NY: Random House.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511571312

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology, 54,* 1–22. http://dx.doi.org/10.1037/h0046743

Cattell, R. B., & Horn, J. L. (1978). A check on the theory of fluid and crystallized intelligence with description of new subtest designs. *Journal of Educational Measurement, 15,* 139–164. http://dx.doi.org/10.1111/j.1745-3984.1978.tb00065.x

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668–1674. http://dx.doi.org/10.1126/science.2648573

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making, 25,* 331–351. http://dx.doi.org/10.1002/bdm.731

Dweck, C. (2006). *Mindset: The new psychology of success.* New York, NY: Random House.

Einhorn, H. J. (1982). Learning from experience and suboptimal rules in decision making. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 268–286). Cambridge, UK: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477.020

Ericsson, K. A. (2006). The influence of experience and deliberative practice on the development of superior expert performance. In K. A. Ericsson, N. Charness, R. Hoffman, & P. Feltovich (Eds.), *Cambridge handbook of expertise and expert performance* (pp. 683–704). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511816796.038

Ericsson, K. A. (2014). Why expert performance is special and cannot be extrapolated from studies of performance in the general population: A response to criticisms. *Intelligence, 45,* 81–103. http://dx.doi.org/10.1016/j.intell.2013.12.001

Ericsson, K. A., Krampe, R. T., & Tesch-Romer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100,* 363–406. http://dx.doi.org/10.1037/0033-295X.100.3.363

Fischhoff, B., & Bruine De Bruin, W. (1999). Fifty–Fifty=50%? *Journal of Behavioral Decision Making, 12,* 149–163. http://dx.doi.org/10.1002/(SICI)1099-0771(199906)12:2<149::AID-BDM314>3.0.CO;2-J

Fischhoff, B., & Chauvin, C. (Eds.). (2011). *Intelligence analysis: Behavioral and social scientific foundations.* Washington, DC: National Academies Press.

Fischhoff, B., Slovic, P., & Lichtenstein, S. (1977). Knowing with certainty: The appropriateness of extreme confidence. *Journal of Experimental Psychology: Human Perception and Performance, 3,* 552–564. http://dx.doi.org/10.1037/0096-1523.3.4.552

Frederick, S. (2005). Cognitive reflection and decision making. *The Journal of Economic Perspectives, 19,* 25–42. http://dx.doi.org/10.1257/089533005775196732

Funder, D., & Ozer, D. (1983). Behavior as a function of the situation. *Journal of Personality and Social Psychology, 44,* 107–112. http://dx.doi.org/10.1037/0022-3514.44.1.107

Furnham, A., & Monsen, J. (2009). Personality traits and intelligence predict academic school grades. *Learning and Individual Differences, 19,* 28–33. http://dx.doi.org/10.1016/j.lindif.2008.02.001

Green, K. C., & Armstrong, J. S. (2007). Global warming: Forecasts by scientists versus scientific forecasts. *Energy & Environment, 18,* 997–1021. http://dx.doi.org/10.1260/095830507782616887

Guilford, J. P., & Hoepfner, R. (1971). *The analysis of intelligence*. New York, NY: McGraw-Hill.

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8,* 188–201.

Itoh, S., Ikeda, M., Mori, Y., Suzuki, K., Sawaki, A., Iwano, S., . . . Ishigaki, T. (2002). Lung: Feasibility of a method for changing tube current during low-dose helical CT. *Radiology, 224,* 905–912. http://dx.doi.org/10.1148/radiol.2243010874

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64,* 515–526. http://dx.doi.org/10.1037/a0016755

Kahneman, D., Slovic, P., & Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases*. New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511809477

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "seizing" and "freezing." *Psychological Review, 103,* 263–283. http://dx.doi.org/10.1037/0033-295X.103.2.263

Laughlin, P. R. (2011). *Group problem solving*. Princeton, NJ: Princeton University Press.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individuals on letters-to-numbers problems: Effects of group size. *Journal of Personality and Social Psychology, 90,* 644–651. http://dx.doi.org/10.1037/0022-3514.90.4.644

Lichtenstein, S., & Fischhoff, B. (1980). Training for calibration. *Organizational Behavior & Human Performance, 26,* 149–171. http://dx.doi.org/10.1016/0030-5073(80)90052-5

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21,* 37–44. http://dx.doi.org/10.1177/0272989X0102100105

Loftus, E. (1996). *Eyewitness testimony*. Cambridge, MA: Harvard University Press.

MacCoun, R. J. (2012). The burden of social proof: Shared thresholds and social influence. *Psychological Review, 119,* 345–372. http://dx.doi.org/10.1037/a0027121

McGraw, P., Todorov, A., & Kunreuther, H. (2011). A policy maker's dilemma: Preventing terrorism or preventing blame? *Organizational Behavior and Human Decision Processes, 115,* 25–34. http://dx.doi.org/10.1016/j.obhdp.2011.01.004

Mellers, B. A., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., . . . Tetlock, P. (2014). *Improving probabilistic predictions by identifying and cultivating "superforecasters"*. Manuscript submitted for publication.

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., . . . Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science, 25,* 1106–1115. http://dx.doi.org/10.1177/0956797614524255

Murphy, A. H., & Winkler, R. L. (1984). Probability forecasting in meteorology. *Journal of the American Statistical Association, 79,* 489–500.

Nickerson, R. S. (1987). *Understanding understanding*. New York, NY: Bolt Beranek and Newman.

Parker, A., & Fischhoff, B. (2005). Decision making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making, 18,* 1–27. http://dx.doi.org/10.1002/bdm.481

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science, 17,* 407–413. http://dx.doi.org/10.1111/j.1467-9280.2006.01720.x

Plomin, R., Shakeshaft, N. G., McMillan, A., & Trzaskowski, M. (2014). Nature, nurture, and expertise. *Intelligence, 45,* 46–59. http://dx.doi.org/10.1016/j.intell.2013.06.008

Ree, M. J., & Earles, J. A. (1992). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89. http://dx.doi.org/10.1111/1467-8721.ep10768746

Reyna, V. F., Chick, C. F., Corbin, J. C., & Hsia, A. N. (2014). Developmental reversals in risky decision making: Intelligence agents show larger decision biases than college students. *Psychological Science, 25,* 76–84. http://dx.doi.org/10.1177/0956797613497022

Riding, R., & Cheema, I. (1991). Cognitive styles—An overview and integration. *Educational Psychology, 11,* 193–215. http://dx.doi.org/10.1080/0144341910110301

Ruthsatz, J., Detterman, D. K., Griscom, W. S., & Cirullo, B. A. (2008). Becoming an expert in the musical domain: It takes more than just practice. *Intelligence, 36,* 330–338. http://dx.doi.org/10.1016/j.intell.2007.08.003

Schkade, D. A., & Kahneman, D. (1998). Does living in California make people happy? A focusing illusion in judgments of life satisfaction. *Psychological Science, 9,* 340–346. http://dx.doi.org/10.1111/1467-9280.00066

Schmidt, F. L., & Hunter, J. (2004). General mental ability in the world of work: Occupational attainment and job performance. *Journal of Personality and Social Psychology, 86,* 162–173. http://dx.doi.org/10.1037/0022-3514.86.1.162

Shanteau, J. (1992). Competence in experts: The role of task characteristics. *Organizational Behavior and Human Decision Processes, 53,* 252–266. http://dx.doi.org/10.1016/0749-5978(92)90064-E

Soll, J. B., Milkman, K. L., & Payne, J. W. (in press). A user's guide to debiasing. In G. Wu & G. Keren (Eds.), *Handbook of judgment and decision making*. New York: Wiley.

Spearman, C. (1927). *The abilities of man: Their nature and measurement*. New York, NY: Macmillan.

Stanovich, K. E., & West, R. F. (1997). Reasoning independently of prior belief and individual differences in actively open-minded thinking. *Journal of Educational Psychology, 89,* 342–357. http://dx.doi.org/10.1037/0022-0663.89.2.342

Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of Experimental Psychology: General, 127,* 161–188. http://dx.doi.org/10.1037/0096-3445.127.2.161

Stanovich, K. E., & West, R. F. (2007). Natural myside bias is independent of cognitive ability. *Thinking & Reasoning, 13,* 225–247. http://dx.doi.org/10.1080/13546780600780796

Steiner, I. D. (1972). *Group processes and group productivity*. New York, NY: Academic Press.

Steyerberg, E. W. (2009). *Clinical prediction models*. New York, NY: Springer. http://dx.doi.org/10.1007/978-0-387-77244-8

Strenze, T. (2007). Intelligence and socioeconomic success: A meta-analytic review of longitudinal research. *Intelligence, 35,* 401–426. http://dx.doi.org/10.1016/j.intell.2006.09.004

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26. http://dx.doi.org/10.1111/1529-1006.001

Taleb, N. N. (2007). Black swans and the domains of statistics. *The American Statistician, 61,* 198–200. http://dx.doi.org/10.1198/000313007X219996

Tetlock, P. E. (1998). Close-call counterfactuals and belief-system defenses: I was not almost wrong but I was almost right. *Journal of*

Personality and Social Psychology, 75, 639–652. http://dx.doi.org/10.1037/0022-3514.75.3.639

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tetlock, P. E., & Mellers, B. A. (2011). Intelligent management of intelligence agencies: Beyond accountability ping-pong. *American Psychologist, 66,* 542–554. http://dx.doi.org/10.1037/a0024285

Thurstone, L. L., & Thurstone, T. G. (1941). *Factorial studies of intelligence.* Chicago, IL: Psychometric Monographs.

Vannoy, J. S. (1965). Generality of cognitive complexity-simplicity as a personality construct. *Journal of Personality and Social Psychology, 2,* 385–396. http://dx.doi.org/10.1037/h0022270

Webster, D. M., & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology, 67,* 1049–1062. http://dx.doi.org/10.1037/0022-3514.67.6.1049

Wells, G. L. (2014). Eyewitness identification: Probative value, criterion shifts, and policy regarding the sequential lineup. *Current Directions in Psychological Science, 23,* 11–16. http://dx.doi.org/10.1177/0963721413504781

Wells, G. L., & Olson, E. A. (2003). Eyewitness testimony. *Annual Review of Psychology, 54,* 277–295. http://dx.doi.org/10.1146/annurev.psych.54.101601.145028

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin, 116,* 117–142. http://dx.doi.org/10.1037/0033-2909.116.1.117

Wilson, T. D., & Gilbert, D. (2005). Affective forecasting: Knowing what to want. *Current Directions in Psychological Science, 14,* 131–134. http://dx.doi.org/10.1111/j.0963-7214.2005.00355.x

Witkin, H. A., Oltman, P. K., Raskin, E., & Karp, S. A. (1971). *A manual for the group embedded figures test.* Palo Alto, CA: Consulting Psychologists Press.

OBSERVATION

What Can 1 Billion Trials Tell Us About Visual Search?

Stephen R. Mitroff, Adam T. Biggs, Stephen H. Adamo,
Emma Wu Dowd, Jonathan Winkle, and Kait Clark
Duke University

Mobile technology (e.g., smartphones and tablets) has provided psychologists with a wonderful opportunity: through careful design and implementation, mobile applications can be used to crowd source data collection. By garnering massive amounts of data from a wide variety of individuals, it is possible to explore psychological questions that have, to date, been out of reach. Here we discuss 2 examples of how data from the mobile game *Airport Scanner* (Kedlin Co., http://www.airportscannergame.com) can be used to address questions about the nature of visual search that pose intractable problems for laboratory-based research. *Airport Scanner* is a successful mobile game with millions of unique users and billions of individual trials, which allows for examining nuanced visual search questions. The goals of the current Observation Report were to highlight the growing opportunity that mobile technology affords psychological research and to provide an example roadmap of how to successfully collect usable data.

Keywords: visual search, big data, mobile applications, airport scanner, multiple-target search

Psychological researchers have never lacked for good questions. However, sometimes method and technology lag behind—thereby withholding the necessary tools for addressing certain questions. As such, each new technological and/or methodological advance provides the field with a possible means to move forward. Cognitive psychology, for example, has grown hand-in-hand with each new available technology (e.g., Wilhelm Wundt's laboratory devices, the tachistoscope, personal computers, eye-tracking devices, brain-imaging techniques).

At present, we are in the early stages of another breakthrough capable of pushing psychological forward. Namely, researchers have begun to "crowd source" experiments to obtain large amounts of data from many people in short order. Building off ingenious ideas, such as Luis von Ahn's "ESP Game" (a "game" that was the basis for how Google matches words to images; von Ahn & Dabbish, 2004), researchers have turned to outlets such as Amazon's Mechanical Turk (e.g., Buhrmester, Kwang, & Gosling, 2011) to rapidly distribute an experiment to many different participants.

Another outlet for rapid distribution of experiments is through the use of mobile applications—"apps" created for mobile devices, such as Apple and Android products. Psychologists have a history of using games and game-like interfaces to make experiments more palatable to participants (e.g., Anguera et al., 2013; Boot et al., 2010; Mané & Donchin, 1989; Miranda & Palmer, 2014), and mobile devices offer an exciting new means to crowd source an experiment in a game-like form.

Some mobile apps have been specifically designed to assess and/or train cognitive abilities, and they can address open questions with data voluntarily contributed by users. Other mobile apps just happen to tap into cognitive abilities in a manner that can be analyzed by researchers, even though that might not have been the apps' intended purpose. For example, some games challenge players to look for differences between images presented side-by-side—a game version of change detection tasks (e.g., Simons & Rensink, 2005). Similarly, other games tap into abilities related to the cognitive processes of working memory (memory match games), go/no-go ("whack-a-mole" games), and visual search (search-and-find games).

Using Mobile Technology for Research

There are clear advantages to crowd sourcing data collection through mobile technology. The most obvious benefit is the potential for gathering "big data"—massive datasets that provide the ability to examine nuanced questions with sufficient statistical power. Likewise, this can provide a means to collect relatively cheap data in an automated and continuous manner. Lastly, this process can mimic real-world aspects that are difficult to address in a laboratory environment (e.g., realistic distributions of variables).

There are also clear disadvantages to consider. First, researchers either need to have the necessary skills to create a fun game or need to partner with a developer. Gathering data through a mobile

game is only worthwhile if people will play the game, and people are more likely to play if the game is fun (e.g., Miranda & Palmer, 2014). Second, large amounts of data may not necessarily result in high-quality data; it is critical to carefully select what research questions are to be addressed and how they are addressed through the game interface. Finally, by collecting data through crowd sourcing, there is an inherent lack of control over who is playing and under what conditions (e.g., there is no way to know what percentage of the data is collected while participants are on the toilet).

Assuming the advantages outweigh the disadvantages and that the disadvantages can be addressed, the largest benefit of data collection through mobile technology is the potential for analyzing big data. With millions (or billions) of trials, it is possible to examine experimental variables that are too difficult to assess in a laboratory environment. While using mobile technology for research purposes may seem like a simple methodological advance, it has the potential to greatly inform psychology theory. Here we discuss our recent efforts focusing on the specific cognitive task of *visual search*.

Examples of Using Mobile Technology for Research

Visual search is the act of looking for target items among distractor items. Decades of research have sought to understand this ubiquitous cognitive process and to determine how humans, nonhuman animals, and computers successfully identify targets (see Eckstein, 2011; Horowitz, 2014; Nakayama & Martini, 2011, for recent reviews). Visual search has a history of using big data analyses—in 1998, Jeremy Wolfe collated data from 2,500 experimental sessions to ask "What can 1 million trials tell us about visual search" (Wolfe, 1998). This endeavor confirmed some open hypotheses and challenged others, while also demonstrating the value of big data for visual search analyses.

The downside of Wolfe's approach was that it took 10 years to collect—as is to be expected in typical laboratory experiments. Mobile apps offer the potential to collect data far more expeditiously. In the current report, we discuss results from our recent partnership with Kedlin Co., the creators of *Airport Scanner* (https://www.airportscannergame.com). In *Airport Scanner*, players are tasked with searching for illegal items in simulated x-ray bag images in an airport security environment. Players view one bag at a time and use finger-taps to identify illegal items on a touchscreen (see Figure 1 for gameplay examples). Players are provided with a logbook of illegal and legal items, and the logbook expands (going from a handful of possible targets to hundreds) as players progress through the game.

As of November 2014, there were over two billion trials from over seven million mobile devices available for research purposes. Data are collected in accordance with the terms and conditions of the standard Apple User Agreement and those provided by Kedlin Co. Each player consents to the terms and conditions when installing the application, and the Duke University Institutional Review Board provided approval for secondary data analyses (see Biggs, Adamo, & Mitroff, 2014; Mitroff & Biggs, 2014, for more details). Here we provide a brief overview of two examples of how we have used the *Airport Scanner* data for research purposes.

Use of Airport Scanner Data

Example 1: Ultra-rare Targets

In a recently published article (Mitroff & Biggs, 2014), we explored how visual search performance is affected when specific targets rarely appeared. While maintaining an overall target prevalence rate of 50% (half of the *bags* in *Airport Scanner* had at least one target present), the frequency with which any given target
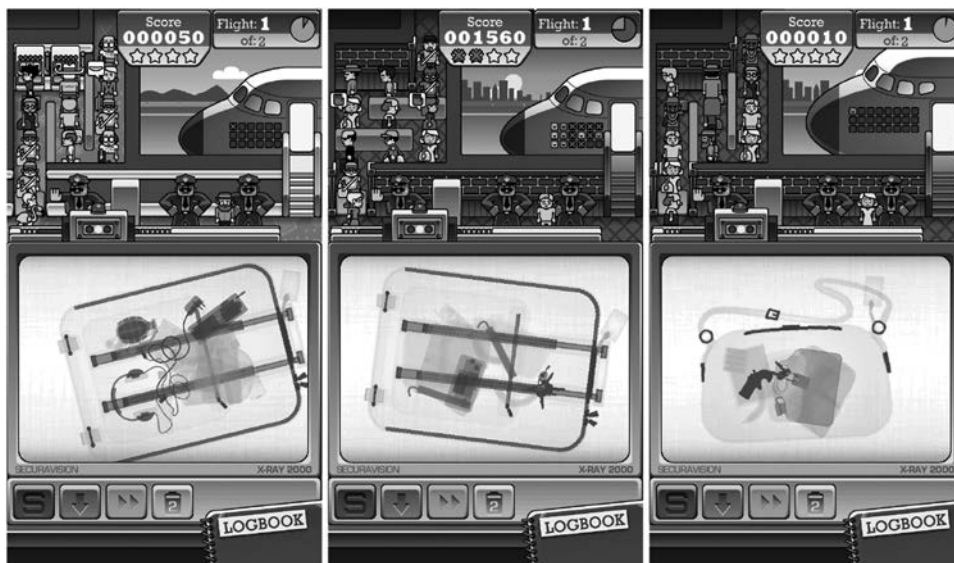
Figure 1. Sample images from *Airport Scanner*: the left image contains one target (*hand grenade*), the middle contains two identical targets (two exemplars of the *dynamite stick* target type), and the right image contains two different target types (*derringer*, *gasoline can*). *Airport Scanner* images appear with permission from Kedlin Co. Copyright 2014 by the Kedlin Company. See the online article for the color version of this figure.

could appear varied greatly (e.g., a *hammer* appeared as a target in 3% of the trials while a *switchblade* appeared as a target in only 0.08% of the trials). Critically, nearly 30 of the targets were "ultra-rare"—they appeared in less than 0.15% of all trials. To examine the effects of such extreme target rarity on visual search performance in a laboratory would be difficult for even one target item. For example, to assess accuracy for targets that only appeared in 0.1% of trials, 1,000 trials would be needed for a single occurrence. To obtain sufficient statistical power (e.g., at least 20 occurrences), too many total trials would be needed to realistically test such a question in a laboratory. However, with the large *Airport Scanner* dataset, we were able to look at hundreds of cases for each of the nearly 30 "ultra-rare" targets. Comparing the relationship between search accuracy and target frequency across 78 unique target types of various frequency rates revealed an extremely strong logarithmic relationship (adjusted $R^2 = .92$) such that the "ultra-rare" items were much more likely to be missed than the more frequently occurring targets (Mitroff & Biggs, 2014). This example highlights the more obvious benefits and drawbacks of using big data to address research questions. The primary benefit is clear—a question that could have taken decades to answer in a laboratory setting can be answered using big data in a fraction of the time. The *Airport Scanner* app also bypasses most of the potential downsides mentioned above given that it is a popular game with an interface that is conducive to research. However, there is an inherent lack of control over the nature of data collected via mobile technology, and there is no obvious means to counter this lack of control. Analogous to a "speed/accuracy trade-off" (a well-studied juxtaposition between performing quickly vs. performing accurately; e.g., Pachella, 1974), big data might engender a "volume/control trade-off"—a juxtaposition between the amount of available data and the methodological control over the data.

Example 2: Multiple-Target Search Theories

Many real-world visual searches can have more than one target present within the same search array (e.g., more than one abnormality in a radiological x-ray; more than one prohibited item in a bag going through airport security). Unfortunately, multiple-target searches are highly susceptible to errors such that additional targets are less likely to be detected if one target has already been found (see Berbaum, 2012, for a review). This effect was originally termed the "satisfaction of search" phenomenon, but we have recently renamed it the "subsequent search misses" (SSM) phenomenon (Adamo, Cain, & Mitroff, 2013). SSM errors are a stubborn source of errors, and several efforts (e.g., Berbaum, 2012; Cain, Adamo, & Mitroff, 2013) have attempted to identify their underlying cause(s).

Three primary theories of SSM have been proposed. First, the original explanation—and the source of the "satisfaction of search" name—suggests that searchers become "satisfied" with the meaning of the search on locating a first target and terminate their search prematurely (Smith, 1967; Tuddenham, 1962). Second, a resource depletion account (e.g., Berbaum et al., 1991; Cain & Mitroff, 2013) suggests that cognitive resources (e.g., attention, working memory) are consumed by a found target and leave fewer resources available to detect additional targets during subsequent search. Finally, a perceptual set account suggests that searchers

become biased to look for additional targets similar to the first found target (Berbaum et al., 1990; Berbaum et al., 1991; Berbaum et al., 2010); for example, if you just found a tumor, you might enter "tumor mode" and be less likely to subsequently detect a fracture that appeared in the same x-ray image.

There have been empirical tests of the satisfaction and resource depletion theories (see Berbaum, 2012), but no substantial tests have been offered for the perceptual set account. Previous investigations have employed a small number of possible target types; for example, Fleck, Samei, and Mitroff (2010) asked observers to search for targets that were T-shaped items among L-shaped distractor items. There were two different forms of the target Ts—those that were relatively light and those that were relatively dark. If a perceptual set operates via a priming-like influence, such a design might be suboptimal, because repeated exposure to each target could result in elevated priming across all trials for all targets (i.e., you only need to see so many lightly colored T-shaped targets before you become generally biased for lightly colored T-shaped targets across the entire experiment).

A large and unpredictable set of targets could generate more short-lived priming during a visual search task, which is more in line with real-world scenarios in which a perceptual set could meaningfully impact performance. However, such an experimental design—many and varied targets spread over an immense number of trials—is not practical to administer in a laboratory environment, and it is not easily assessed in real-world scenarios such as an airport security checkpoint. Here we used the *Airport Scanner* gameplay data to address this idea.

More details about the nature of the data and the gameplay are available elsewhere (e.g., Mitroff & Biggs, 2014); however, we highlight in Table 1 how we filtered the gameplay variables to address our specific research question at hand. Each trial (*bag*) could contain 0–3 illegal target items with approximately 43% of all trials containing one target, 6% with two targets, and less than 1% with three targets. We analyzed SSM errors by comparing search accuracy for a specific target item on single-target trials to search accuracy for the same target on dual-target trials when another target had been detected first (e.g., Biggs & Mitroff, 2014). Players identified a target by tapping directly onto the target location, and we excluded all cases in which one tap captured two targets. Analyses were limited to target types with at least 20 instances within our dataset (i.e., the *shotgun* target type was filtered from our SSM analyses for only contributing seven instances).

We first determined whether SSM errors occurred in the *Airport Scanner* gameplay as we have observed in the laboratory (e.g., Cain et al., 2013; Fleck et al., 2010). This analysis was performed across 78 target items without considering the identity of the first found target; for example, when calculating the SSM error rate for the *pistol* as a second target, all data were included whether or not the first found target was another *pistol*, a *grenade*, a *knife*, and so forth. A significant overall SSM error rate was found ($M = 14.00\%$, $SE = 1.11\%$), $t(77) = 12.62$, $p < .001$.

Next, we assessed SSM errors when the two targets in the same bag were identical (e.g., two exemplars of the *dynamite stick* target type; e.g., the second panel of Figure 1) as opposed to when the two targets in the same bag were not identical (e.g., one *pistol* and one *hand grenade*; e.g., the third panel of Figure 1). Thirty-three target types met the 20-occurrence minimum for inclusion into

Table 1
Game Elements and Nature of Trial Filtering for the Multiple-Target Visual Search Example

Game element/variable			
Game variable	Description	Cases	Filtering for SSM errors example
Airport	6 levels; increase in difficulty	*Trainee*, *Honolulu*, *Las Vegas*, *Chicago*, *Aspen*, *London*	Exclude *Trainee*
Rank	5 levels; player's experience level	*Trainee*, *Operator*, *Pro*, *Expert*, *Elite*	Only *Elite* players (here = 62,606 devices)
Day	Sessions within Airport level	5 *Days* per *Airport*; additional *Challenge* levels for some *Airports*	Exclude *Challenge*
Mission type	Game play mode	*Career*, *Challenge*	Only *Career* mode
Replay	Repeat a Day after completing it	Replays allowed or disallowed	Replays allowed
Day status	How the Day session ended	Completed, timed out, security breach	No exclusions
Bag type	Shape and size of search array	> 15 unique *Bag* types	*Briefcase*, *carry-on*, *duffle*, and *purse* included
Passenger type	Difficulty of Bag	*Easy*: ≈ 0–8 legal items present *Medium*: ≈ 9–13 legal items present *Hard*: ≈ 14–20 legal items present *Impossible*: Requires upgrades	Exclude *Impossible*
In-game upgrades	Add-ons to help with gameplay	> 10 unique upgrades	Exclude all that affect search performance
Special passengers/items	Nontypical gameplay events	*Air Marshals*, *Flight Crew*, *First Class Passengers*, *Delay Passengers*, *Rare Targets (special items)*	*Air Marshal* and *rare-target Bags* excluded
Illegal item count	No. of target items present	0 targets: ≈ 50% of all trials 1 target: ≈ 43% of all trials 2 targets: ≈ 6% of all trials 3 targets: ≈ 1% of all trials	Excluded 0-illegal and 3-illlegal item *Bags*
Legal item count	No. of distractor items present	0–20	No exclusions
Specific illegal items	Various target items (see Figure 1)	> 200	1-target accuracy: $N = 78$ 2-target accuracy: $N = 33$
Specific legal items	Various distractor items (see Figure 1)	> 200	No exclusions
How data filtering affected trial counts			
Total trials available as of 11/18/14			2,236,844,667
Total trials for example analysis date range of 04/15/13 to 08/26/13			1,098,098,764
Total trials for 1-target trial accuracy analyses after all filters applied			1,795,907
Total trials for 2-target trial accuracy analyses after all filters applied			126,579

Note. SSM = subsequent search misses.

analyses. We observed significant SSM errors when the first and second targets were identical ($M = 6.53\%$, $SE = 1.62\%$), $t(32) = 4.04$, $p < .001$, and when the first and second targets were not identical ($M = 19.21\%$, $SE = 1.36\%$), $t(32) = 14.13$, $p < .001$. Importantly, there was a significant difference in SSM error rates for identical targets versus nonidentical targets ($M_{\text{difference}} = 12.69\%$, $SE = 1.69\%$), $t(32) = 7.52$, $p < .001$, such that SSM errors were substantially reduced when the targets were identical than when the two targets were not identical.

In this particular example, the power of big data allowed us to answer a nuanced question that required a substantial number of trials to appropriately assess. Specifically, this analysis focused on trials that contained two targets and that met several exclusionary criteria (see Table 1), which resulted in analyzing only about 125,000 trials out of a dataset of more than 1,000,000,000 trials (0.01%). It was necessary to examine accuracy for specific target types within a framework containing a wide variety of target types to prevent participants from becoming overexposed to any one target. This was best accomplished through the use of mobile technology, which allowed for the accumulation of the necessary data while providing

players an enjoyable experience. Importantly, we expanded current understanding about the mechanisms underlying SSM errors by revealing that the errors are, in fact, partially due to a perceptual set mechanism.

Conclusion

Using a game interface to assess cognitive abilities is not new to psychological research (e.g., Boot et al., 2010; Mané, & Donchin, 1989), but mobile technology offers a phenomenal opportunity to examine cognitive processes on a large scale. Here we discussed two specific examples of how we analyzed data from the *Airport Scanner* game to address psychological questions, and much more is possible.

However, using mobile apps for research purposes is easier said than done. Researchers can build games for data collection purposes and have complete control over the design and implementation. However, there is no guarantee that any game will be successful enough to garner data—simply producing a game does not ensure anyone will play it. Alternatively, researchers can partner with developers to create apps or can partner with devel-

opers of already existing apps, which can be a great opportunity for both groups; developers can benefit from researchers' insight and added press, and researchers can benefit from developers' skill and access to established games. Our partnership with Kedlin Co. exemplifies this beneficial relationship and has been successful enough to lead to federal funding opportunities for further research implementations of *Airport Scanner*.

With the proliferation of mobile technology, it is time to aim high. In 1998, 1 million trials on a specific cognitive task was a mind-blowing amount of data (Wolfe, 1998). Today, we collect over a million trials a day through *Airport Scanner*. Inflation has hit visual search, and as researchers, we should (responsibly and carefully) look for ways to take advantage of this opportunity.

References

Adamo, S. H., Cain, M. S., & Mitroff, S. R. (2013). Self-induced attentional blink: A cause of errors in multiple-target search. *Psychological Science, 24,* 2569–2574. http://dx.doi.org/10.1177/0956797613497970

Anguera, J. A., Boccanfuso, J., Rintoul, J. L., Al-Hashimi, O., Faraji, F., Janowich, J., . . . Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. *Nature, 501,* 97–101. http://dx.doi.org/10.1038/nature12486

Berbaum, K. S. (2012). Satisfaction of search experiments in advanced imaging. *Proceedings of the Society for Photo-Instrumentation Engineers, 8291,* 82910V. http://dx.doi.org/10.1117/12.916461

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Coffman, C. E., Cornell, S. H., . . . Smith, T. P. (1991). Time course of satisfaction of search. *Investigative Radiology, 26,* 640–648. http://dx.doi.org/10.1097/00004424-199107000-00003

Berbaum, K. S., Franken, E. A., Jr., Dorfman, D. D., Rooholamini, S. A., Kathol, M. H., Barloon, T. J., . . . Montgomery, W. J. (1990). Satisfaction of search in diagnostic radiology. *Investigative Radiology, 25,* 133–140. http://dx.doi.org/10.1097/00004424-199002000-00006

Berbaum, K. S., Franklin, E. A., Jr., Caldwell, R. T., & Schartz, K. M. (2010). Satisfaction of search in traditional radiographic imaging. In E. Samei & E. Krupinski (Eds.), *The handbook of medical image perception and techniques* (pp. 107–138). New York, NY: Cambridge University Press.

Biggs, A. T., Adamo, S. H., & Mitroff, S. R. (2014). Rare, but obviously there: Effects of target frequency and salience on visual search accuracy. *Acta Psychologica, 152,* 158–165. http://dx.doi.org/10.1016/j.actpsy.2014.08.005

Biggs, A. T., & Mitroff, S. R. (2014). Different predictors of multiple-target search accuracy between nonprofessional and professional visual searchers. *The Quarterly Journal of Experimental Psychology, 67,* 1335–1348. http://dx.doi.org/10.1080/17470218.2013.859715

Boot, W. R., Basak, C., Erickson, K. I., Neider, M., Simons, D. J., Fabiani, M., . . . Kramer, A. F. (2010). Transfer of skill engendered by complex task training under conditions of variable priority. *Acta Psychologica, 135,* 349–357. http://dx.doi.org/10.1016/j.actpsy.2010.09.005

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Per-*spectives on Psychological Science, 6,* 3–5. http://dx.doi.org/10.1177/1745691610393980

Cain, M. S., Adamo, S. H., & Mitroff, S. R. (2013). A taxonomy of errors in multiple-target visual search. *Visual Cognition, 21,* 899–921. http://dx.doi.org/10.1080/13506285.2013.843627

Cain, M. S., & Mitroff, S. R. (2013). Memory for found targets interferes with subsequent performance in multiple-target visual search. *Journal of Experimental Psychology: Human Perception and Performance, 39,* 1398–1408. http://dx.doi.org/10.1037/a0030726

Eckstein, M. P. (2011). Visual search: A retrospective. *Journal of Vision, 11*(5), 14. http://dx.doi.org/10.1167/11.5.14

Fleck, M. S., Samei, E., & Mitroff, S. R. (2010). Generalized "satisfaction of search": Adverse influences on dual-target search accuracy. *Journal of Experimental Psychology: Applied, 16,* 60–71. http://dx.doi.org/10.1037/a0018629

Horowitz, T. S. (2014). Exit strategies: Visual search and the quitting time problem. In R. Metzler, G. Oshanin, & S. Redner (Eds.), *First-passage phenomena and their applications* (pp. 390–415). Hackensack, NJ: World Scientific Press.

Mané, A., & Donchin, E. (1989). The Space Fortress game. *Acta Psychologica, 71,* 17–22. http://dx.doi.org/10.1016/0001-6918(89)90003-6

Miranda, A. T., & Palmer, E. M. (2014). Intrinsic motivation and attentional capture from gamelike features in a visual search task. *Behavior Research Methods, 46,* 159–172. http://dx.doi.org/10.3758/s13428-013-0357-7

Mitroff, S. R., & Biggs, A. T. (2014). The ultra-rare-item effect: Visual search for exceedingly rare items is highly susceptible to error. *Psychological Science, 25,* 284–289. http://dx.doi.org/10.1177/0956797613504221

Nakayama, K., & Martini, P. (2011). Situating visual search. *Vision Research, 51,* 1526–1537. http://dx.doi.org/10.1016/j.visres.2010.09.003

Pachella, R. (1974). The interpretation of reaction time in information processing research. In B. H. Kantowitz (Ed.), *Human information processing: Tutorials in performance and cognition* (pp. 41–82). Hillsdale, NJ: Erlbaum.

Simons, D. J., & Rensink, R. A. (2005). Change blindness: Past, present, and future. *Trends in Cognitive Sciences, 9,* 16–20. http://dx.doi.org/10.1016/j.tics.2004.11.006

Smith, M. J. (1967). *Error and variation in diagnostic radiology*. Springfield, IL: C. C. Thomas.

Tuddenham, W. J. (1962). Visual search, image organization, and reader error in roentgen diagnosis. Studies of the psychophysiology of roentgen image perception. *Radiology, 78,* 694–704. http://dx.doi.org/10.1148/78.5.694

von Ahn, L., & Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems,* 319–326. http://dx.doi.org/10.1145/985692.985733

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? *Psychological Science, 9,* 33–39. http://dx.doi.org/10.1111/1467-9280.00006

Sending Your Grandparents to University Increases Cognitive Reserve: The Tasmanian Healthy Brain Project

Megan E. Lenehan
University of Tasmania

Mathew J. Summers
University of the Sunshine Coast and University of Tasmania

Nichole L. Saunders
University of Tasmania, Hobart

Jeffery J. Summers
University of Tasmania, and Liverpool John Moores University

David D. Ward
University of Tasmania, Hobart

Karen Ritchie
INSERM, Montpellier, France

James C. Vickers
University of Tasmania

Objective: Increasing an individual's level of cognitive reserve (CR) has been suggested as a nonpharmacological approach to reducing the risk for Alzheimer's disease. We examined changes in CR in older adults participating over 4 years in the Tasmanian Healthy Brain Project. *Method:* A sample of 459 healthy older adults between 50 and 79 years of age underwent a comprehensive annual assessment of current CR, neuropsychological function, and psychosocial factors over a 4-year period. The intervention group of 359 older adults (M = 59.61 years, SD = 6.67) having completed a minimum of 12 months part-time university study were compared against a control reference group of 100 adults (M = 62.49 years, SD = 6.24) who did not engage in further education. *Results:* Growth mixture modeling demonstrated that 44.3% of the control sample showed no change in CR, whereas 92.5% of the further education participants displayed a significant linear increase in CR over the 4 years of the study. These results indicate that older adults engaging in high-level mental stimulation display an increase in CR over a 4-year period. *Conclusion:* Increasing mental activity in older adulthood may be a viable strategy to improve cognitive function and offset cognitive decline associated with normal aging.

Keywords: aging, education, cognitive reserve, age-related cognitive decline

One nonpharmacological approach to reducing the risk of rapid age-related cognitive decline and Alzheimer's disease is to increase cognitive reserve (CR). CR is a theoretical construct describing the capacity of an individual to utilize preexisting brain networks efficiently (neural reserve) as well as to enlist alternate brain networks (neural compensation) when under the duress of brain pathology (Stern, 2002; Tucker & Stern, 2011). Life experiences and innate intelligence are proposed to impart CR on individuals (Stern, 2002). Research evidence has supported the role of occupational attainment (Valenzuela & Sachdev, 2006), intelligence (Whalley et al., 2000), education (e.g., Anstey & Christensen, 2000), and involvement in cognitively stimulating activities (Scarmeas & Stern, 2003) in modifying an individual's risk for dementia. It is inferred that the modification of an individual's risk for dementia is a result of modifications to the level of CR that the person displays.

CR is a theoretical construct, and therefore it is imperative to recognize that what is measured (latent variable, observed score on a task or test, or physical property) is not the same thing as the construct (Zumbo, 2007). At best, attempts to operationalize and measure CR (Harrison et al., 2015) represent proxy measures with differing levels of construct validity. Various studies have used single-proxy measures to infer the impact of CR on cognitive performance and rate of age-related cognitive decline. For example, individuals with lower occupational status have shown lower performance on measures of global cognitive function in later life (Dartigues et al., 1992; Frisoni, Rozzini, Bianchetti, & Trabucchi, 1993; Jorm et al., 1998). Similarly, a socially engaged lifestyle in later life is associated with superior cognitive performance and a reduced rate of age-related cognitive decline (Barnes, Mendes de Leon, Wilson, Bienias, & Evans, 2004; Ertel, Glymour, & Berkman, 2008; Lövdén, Ghisletta, & Lindenberger, 2005).

A key contributor to CR is thought to be education. Education is seen as increasing CR through fostering the development of new cognitive strategies (Manly, Byrd, Touradji, Sanchez, & Stern, 2004). Educational attainment not only is associated with a decreased risk of dementia (Valenzuela & Sachdev, 2006) but also modifies the association between a direct measure of brain pathology and performance on measures of cognitive function (Bennett et al., 2003; Dufouil, Alpérovitch, & Tzourio, 2003). Despite mixed results, higher levels of education in early adulthood have been associated with superior performance on measures of cognitive function (Anstey & Christensen, 2000; Lenehan, Summers, Saunders, Summers, & Vickers, 2015). Therefore, regardless of whether education influences the rate of normal age-related cognitive decline, enhancing an individual's level of cognitive function has the potential of preserving normal cognitive function for a longer period of time in the presence of neuropathological changes in the brain.

A recent advancement in the area of CR research has been the development of a multidimensional proxy measure of CR (Ward, Summers, Saunders, & Vickers, 2015). Previous research has typically utilized a single proxy measure, such as years of education or occupational attainment, to infer an individual's level of CR. However, this approach may not be accurate given that education, occupational attainment, and leisure activities differentially contribute to CR (Foubert-Samier et al., 2012). Acknowledging the multivariate nature of CR, we developed two factor–analysis-defined latent proxy measures of CR (Ward et al., 2015). Prior CR combines proxy measures traditionally associated with CR, including education, preexisting intellectual capacity, and five subscores from the Life Experience Questionnaire (Valenzuela & Sachdev, 2007). However, because CR theoretically develops in response to new life experiences throughout the life span, we developed a second proxy measure of CR designed to assess dynamic change in CR (Ward et al., 2015). This measure of current CR incorporates cognitive tests suitable for repeated assessment including current intellectual capacity and academic ability (Ward et al., 2015). Whereas prior CR enables CR set earlier in life to be determined, current CR enables possible increases in CR following an intervention to be quantified. University study typically involves complex mental and social stimulation that is increasingly being accessed by older populations.

The Tasmanian Healthy Brain Project (THBP) is the first prospective study in the world to examine the potential of university level of education in later life to reduce age-related cognitive decline (Summers et al., 2013). The THBP has recruited a sample of older adults, age 50–79 years at commencement in the study, from the island state of Tasmania, Australia. The THBP adopts a mixed-group longitudinal design, comparing older adults who engaged in later-life tertiary study with a control group who do not undertake further education. The THBP undertakes annual assessment of each participant, examining cognitive reserve, neuropsychological/cognitive function, psychosocial function, and genetic factors. This article examines whether engaging healthy older adults in university-level education results in a measureable change in CR when accounting for preexisting CR levels for each individual.

Method

Participants

Data from participants in the THBP as of December 31, 2014, were utilized for this study. The initial sample comprised 566 adults age 50–79 years enrolled in the THBP (Summers et al., 2013). Of these, 19 were excluded from the analysis due to English being a second, rather than primary, language. A further 41 were excluded due to having withdrawn from the project prior to any follow-up testing. Of the remaining 498 participants, a further 39 were missing data necessary to calculate prior CR score. Because prior CR was used as a covariate in the analysis, participants with missing data on this variable were excluded. The final sample used in the analysis consisted of 459 healthy older adults.

Participants were not randomly allocated to conditions but volunteered to participate in either the intervention or control condition. Those in the intervention group ($N = 359$) undertook a minimum of 12 months part-time or full-time university study, with a minimum study load of two units at the undergraduate or postgraduate level. The remaining 100 participants in the control group did not engage in any tertiary-level study. Participants who presented with a medical, neurological, or psychiatric disorder that could potentially influence neuropsychological test performance were precluded from entry into the THBP. The project was approved by the Human Research Ethics Committee (Tasmania) Network, and further details of the study protocol have been published elsewhere (Summers et al., 2013).

Measures

Participants in the THBP completed a comprehensive testing battery. For the full project protocol, refer to Summers et al. (2013). The Dementia Rating Scale–2 (DRS-2; Jurica, Leitten, & Mattis, 2001), the Hospital Anxiety and Depression Scale (HADS; Snaith, 2003), and the Medical Health Status questionnaire (Summers et al., 2013) were completed to ensure participants were free from dementia and of sound psychological and physical health. The Personal Wellbeing Index (PWI; International Wellbeing Group, 2006) and the 18-item version of the Lubben Social Network Scale (LSNS-18; Lubben & Gironda, 2003) are self-report questionnaires and were completed to assess quality of life and perceived social support within the sample.

Prior CR. The tests included in the calculation of prior CR were as specified in Ward et al. (2015): the Wechsler Test of Adult Reading (WTAR; Wechsler, 2001) to estimate baseline intellectual capacity; five subscores (Young Adulthood Specific and Nonspecific; and the Midlife Specific, Nonspecific, and Continuing Education Bonus) from the Life Experience Questionnaire (LEQ; Valenzuela & Sachdev, 2007) to quantify previous lifetime experience in education, occupation, and leisure activities; and the Medical Health Status Questionnaire (Summers et al., 2013) to obtain each individual's total years of prior education.

Current CR. The tests used for the calculation of current CR as specified in Ward et al. (2015) were: the Wechsler Adult Intelligence Scale, Third Edition, Short Form 1 (WAIS-III-SF1; Donnell, Pliskin, Holdnack, Axelrod, & Randolph, 2007) to estimate current intellectual capacity and the spelling and math computation subtests of the Wide Range Achievement Test, Fourth Edition, Progress Monitoring Version (WRAT-4-PMV; Roid & Ledbetter, 2006) to assess current academic ability. The WRAT-4-PMV has four alternate versions of each test, which were utilized to avoid learning effects (e.g., Form 1 at baseline, Form 2 at Year 1 follow-up).

Procedure

The elements of the test battery used in the current analysis were as follows: WTAR, DRS-2, Medical Health Status, LEQ, WAIS-III-SF1, WRAT-4-PMV, HADS, PWI, and LSNS-18. The LEQ and WTAR IQ estimate were collected only once, at baseline. Retesting occurred at 1-year intervals (±1 month). When available, alternate versions of tests were used to minimize familiarity effects (e.g., Forms 1–4 of the WRAT). The full THBP took approximately 4 hr to complete, and participants were encouraged to take short breaks as needed to avoid fatigue (Summers et al., 2013).

Analysis

Calculating prior CR and current CR. Current CR and prior CR were calculated for each participant using factor analysis defined regression coefficients as developed and described by Ward and colleagues (2015). This equation was used to calculate prior CR: .370 (WTAR full-scale IQ) + .408 (prior education in years) + .567 (LEQ Young Adulthood Specific) + .565 (Young Adulthood Nonspecific) + .630 (LEQ Midlife Nonspecific) + .875 (LEQ Midlife Continuing Education Bonus) + 1.004 (LEQ Midlife Specific). This equation was used to calculate current CR: .454 (WAIS-III-SF1) + .369 (WRAT-4-PMV Spelling Level Equivalent Score [LES]) + .463 (WRAT-4-PMV Math Computation LES). Because the regression-based formula for prior CR and current CR are based on z-transformed raw scores, current CR scores for Years 1, 2, and 3 (retesting) were z-transformed against the mean and standard deviation of the entire sample at baseline (Year 0). Therefore, positive CR scores represent an increase in CR relative to baseline CR scores.

Modeling approach. Growth mixture modeling (GMM) was conducted using Mplus 7.0 (Muthén & Muthén, 1998–2012) maximum likelihood with robust standard errors estima-

tion. GMM identifies unobserved, homogenous subgroups of individuals from larger heterogeneous populations on the basis of similar response patterns (Muthén & Muthén, 1998–2012). This is important, given research has shown that various subpopulations exist within a broader population and are differentially impacted by an intervention (Jackson & Sher, 2005). This is particularly relevant in the field of CR research, given that a potential increase in CR could depend on each individual's untapped CR capacity. Taking this into account, the conventional latent curve growth approach to analysis could oversimplify and potentially underestimate change (Jung & Wickrama, 2008). As such, GMM was conducted on the control and intervention groups separately to examine whether each group was characterized by classes of individuals with distinct patterns of change in current CR. We chose to analyze each group separately (relative comparison) to enable an examination of differences between individuals high or low in CR in the rate of CR change over the 4-year period.

The procedure outlined by Jung and Wickrama (2008) for conducting GMM was followed. Because the number of unobserved groups is unknown to the investigator, the suggested procedure is to identify the best fitting single-class latent growth curve model (e.g., linear or quadratic) and then progressively test models with more classes until the model fit is no longer improved by the addition of extra classes (Jung & Wickrama, 2008). In all models time was paramatized with scores that represented years since study entry (0, 1, 2, 3 for the linear term and 0, 1, 4, 9 for the quadratic term). Initially, Mplus default parameters were used. The intercepts of the outcome variable at the four time points were fixed at zero. The intercepts, residual variances, and covariances of the growth factors were estimated and not held equal across classes. The model allowed for the effect of the covariates on the growth parameters for each class to be estimated. Incremental model changes such as fixing growth factor variance to zero were also investigated to find the best fitting model. In each group, initial status of the model represented mean current CR at baseline, the linear growth rate represented the annual rate of change in current CR, and the quadratic growth rate indicated the change in the rate of change (accelerating or decelerating change). Because the models included a covariate (conditional models), the *intercepts* describe the growth factors (i.e., initial starting point, linear term, and quadratic term) after taking into account the effect of covariates, so these are reported throughout.

Model evaluation. In the initial latent growth curve analysis (single class), model fit was assessed by considering a range of fit indices: the likelihood-ratio chi-square, the root-mean-square error of approximation (RMSEA), the standardized root-mean-square residual (SRMR), and the comparative fit index (CFI). As a general rule, a smaller chi-square indicates a better fit. An RMSEA value less than .05 and an SRMR less than .05 indicate a good-fitting model (Geiser, 2013) The CFI should be larger than .95. For GMM, the optimal number of classes was determined by considering both the Bayes information criteria (BIC) and the sample-adjusted BIC. As a general rule, the model with the smallest information criterion is preferred (Geiser, 2013). The interpretability of classes was also considered with reference to theory and prior research (Schaie, 1989).

Results

Descriptive Data

Data from a sample of 459 participants was included in this study. Participants at commencement in the study were 50–79 years of age, of average intelligence, free from dementia, and not clinically depressed or anxious (see Table 1). The control group was significantly older, $t(496) = 4.32, p < .001$, and had lower current CR at baseline, $t(494) = -3.05, p < .01$, compared to the intervention group. However, because there were no significant correlations between age and current CR at any time point in either the control group or the intervention group, the decision was made not to include age as a covariate in further analysis. There were no significant differences between the control and intervention groups across baseline measures of prior CR, global cognition, estimated premorbid IQ, level of anxiety, or level of depression. The mean scores of current CR of the control group were lower at baseline compared to the experimental group, but both groups appeared to increase current CR score over time.

In the control group, the best fitting single class model was a linear model with prior CR included as a time-invariant covariate, $\chi^2(7, N = 100) = 23.00, p = < .01$, RMSEA = .15, confidence interval (CI) [.09, .22], SRMR = .04, CFI = .95. In the intervention group, the best fitting model was a quadratic model with prior CR included as a covariate, $\chi^2(7, N = 359) = 26.45, p = < .001$, RMSEA = .09, CI [.05, .13], SRMR .04, CFI = .98. Zero variance in the linear and quadratic growth factors was specified to avoid an inadmissible model due to negative residual variances. These models were used to progressively test models with more classes in each of the control and intervention groups.

GMM control group. The lowest Adjusted Bayes Information Criterion (ABIC) corresponded to a two-class model. The entropy was calculated at .60, which indicated that the model had a reasonable classification of individuals into classes. Class 1 (maintainers) comprised 44.3% of the control group. In Class 1, the linear slope was not significant, indicating that linear change in current CR did not significantly differ from zero (see Figure 1 and Table 2). The remainder of the control group were in Class 2 (improvers; 55.7%). This class had a significant linear slope, suggesting progressive increase in CR over the 4-year period (see Figure 1 and Table 2). The effect of prior CR was consistent in both classes (see Table 2). Higher prior CR was associated with a higher current CR score at baseline. Prior CR did not have a significant association with the rate of linear change in current CR over time. The classes were examined to determine whether other demographic variables could account for class membership. However, there were no differences between decliners and improvers in sex, age, level of depression, level of anxiety, personal wellbeing, or social connectedness.

Intervention group. The lowest ABIC corresponded to a two-class model in the intervention group also, and the entropy value of .78 indicated good separation of individuals into classes. Class 1 (maintainers) constituted a minority of the intervention group (7.5%). In this class the significant negative linear growth term indicates that current CR score decreased over the 4-year period, and the significant quadratic term suggests that CR change accelerated over time (see Figure 2 and Table 3). The majority of the intervention group were in Class 2 (improvers; 92.5%). The significant linear growth term indicates that the current CR for this class increased over the 4-year period (see Figure 2 and Table 3). The negative quadratic term indicated the rate of increase decelerated over time, though this parameter was not significant (see Table 3). Within Class 1 (maintainers), higher prior CR was associated with lower current CR at baseline. However, within Class 2 (improvers), higher prior CR was associated with higher current CR at baseline. In both classes, prior CR had no association with the rate of linear or quadratic change in current CR over time (see Table 3).

The classes were examined to see whether other demographic variables could describe class membership. However, there were no differences between maintainers and improvers in sex, age,

Table 1

Sample Demographic and Cognitive Reserve (CR) as a Function of Group

Variable	Control (baseline N = 100)		Intervention (baseline N = 359)		p			
	N (%)	M (SD)	N (%)	M (SD)	Independent samples t test	χ^2	Obtained effect size (d)	Power
Female	64 (61)		273 (69.5)			.10		
Baseline age (years)		62.62 (6.34)		59.48 (6.69)	<.001		.482	.828
DRS-2 AEMSS		11.81 (2.27)		11.96 (2.07)	.52		.069	.004
WTAR (est. FSIQ)		112.23 (5.10)		112.65 (5.47)			.079	.005
HADS–Anxiety		5.51 (2.91)		5.24 (3.15)	.35		.090	.006
HADS–Depression		2.86 (2.28)		2.38 (2.26)	.05		.212	.076
Prior CR		−0.36 (2.27)		0.13 (2.28)	.06		.215	.081
Current CR								
Year 0 (baseline)		−0.26 (1.01)		0.07 (0.98)	.002		.332	.354
Year 1		−0.05 (1.12)		0.32 (1.05)	.04		.341	.384
Year 2		0.11 (0.97)		0.34 (1.00)	.11		.234	.108
Year 3		0.22 (1.11)		0.68 (0.98)	.01		.439	.716

Note. DRS-2 AEMSS = Dementia Rating Scale age and education corrected Mayo scaled score; WTAR (est. FSIQ) = Wechsler Test of Adult Reading Scale estimated full-scale IQ; HADS = Hospital Anxiety and Depression Scale.
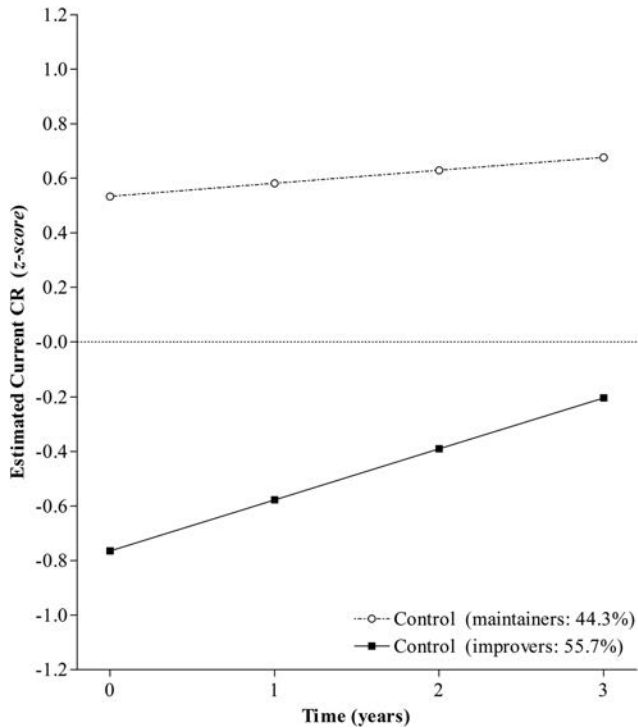
Figure 1. Control Group 2 class model estimated means adjusted for the effect of prior cognitive reserve (CR). The dotted horizontal line indicates the 50th percentile of current CR of the entire cohort at baseline.
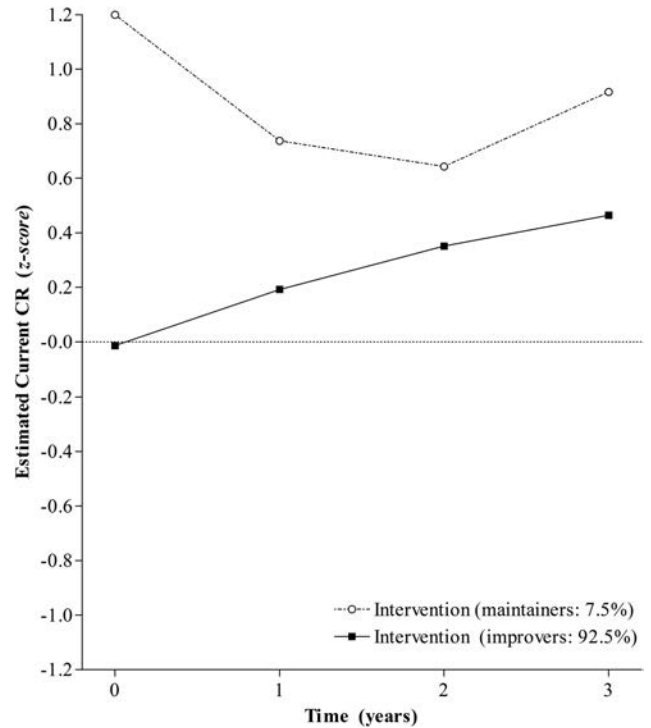
Figure 2. Intervention Group 2 class model estimated means adjusted for the effect of prior cognitive reserve (CR). The dotted horizontal line indicates the 50th percentile of current CR of the entire cohort at baseline.

level of depression, level of anxiety, personal wellbeing, or social connectedness.

Discussion

The hypothesis that individuals who receive an education intervention will display an increase in CR compared to a control group was supported by the results of the study. In both the control group and the intervention group, there appear to be two distinct subgroups of individuals. In the intervention group, approximately 92.5% of the sample displayed a significant increase in CR over time, whereas the remaining 7.5% generally maintained CR across

the 4-year period. Among those in the intervention group, the maintainers displayed higher levels of CR at baseline relative to the improvers. In contrast, among the control group participants, 44.3% displayed no change in CR over time, with the remaining 55.7% displaying a significant increase in CR over the 4 years. The increase in CR seen in this subgroup of control participants was evident in those individuals who displayed below-average CR at baseline. Despite increasing over time, the level of CR of the control improvers remained below the 50th percentile of the baseline CR of the entire cohort.

Table 2

Estimates (and Standard Errors) of Class-Specific Intercept Parameters and the Effect of Prior CR on Class-Specific Growth Terms for the Control Group

| Variable | Model estimate (SE) | | Effect size d |
	Class 1: Maintainers (n = 43)	Class 2: Improvers (n = 57)	
Initial status	.598 (.242)*	−.674 (.114)**	4.34
Linear growth rate	.040 (.044)	.185 (.052)**	1.67
Prior CR covariate			
Initial status	.180 (.078)*	.253 (.058)**	0.62
Linear term	−.022 (.019)	−.004 (.018)	0.55

Note. CR = cognitive reserve; $d = \beta_{11}$(time)/SD_{pooled} (Feingold, 2009).
* $p < .05.$ ** $p < .01.$

Table 3

Estimates (and Standard Errors) of Class-Specific Intercept Parameters and the Effect of Prior CR on Class-Specific Growth Terms for the Intervention Group

| Variable | Model estimate (SE) | | Effect size d |
	Class 1: Maintainers (n = 15)	Class 2: Improvers (n = 344)	
Initial status	1.227 (.229)**	−.038 (.053)	5.17
Linear growth rate	−.664 (.203)	.226 (.051)**	3.79
Quadratic growth rate	.189 (.072)**	−.022 (.018)	2.55
Prior CR covariate			
Initial status	−.208 (.062)**	.182 (.022)**	3.89
Linear term	.133 (.082)**	.024 (.021)	1.13
Quadratic term	−.037 (.028)	−.005 (.008)	0.87

Note. CR = cognitive reserve; $d = \beta_{11}$(time)/SD_{pooled} (Feingold, 2009).
* $p < .05.$ ** $p < .01.$

These results indicate that the overwhelming majority of healthy older adults who engaged in some degree of university-level education for at least 12 months displayed a measureable increase in CR over a 4-year period. The small number of participants who displayed no change in CR over time while attending university already had higher than average CR at baseline ($\sim$1.2 SDs above the cohort at baseline). This tentatively suggests that individuals with already high levels of current CR may lack the capacity for further increases in current CR. This finding should be interpreted with caution, however, due to the small sample size for this group ($n = 15$).

The findings of the present research are consistent with other investigations reporting benefits from cognitive training programs (Ball et al., 2002) and physical activity (Kramer et al., 1999) on cognitive function, presumably through the positive effect these activities have on building CR. The proportion of the control group who showed improvement in current CR despite not receiving the intervention is comparable to that shown in other studies. For example, up to 37% of the no-contact control group in the study by Ball and colleagues (2002) showed increases on a range of cognitive measures despite not receiving a cognitive training program. That 55.7% of the control group in the present study displayed an increase in CR may reflect unreported involvement in mentally complex and stimulating activities outside of the THBP. It would have been informative to have an ongoing measure of noneducational life experiences and activities, beyond baseline, in order to explain control group growth.

For three of the groups, prior CR tended to be associated with higher current CR at baseline. This finding suggests that prior life experience, such as education, promotes higher levels of CR in later life. However, in the intervention-maintainers group, prior CR was associated with lower current CR at baseline. Due to the small sample size of this group ($n = 15$), such associations must be treated with caution. There was no association between prior CR and the rate of linear or quadratic change over time. Thus, prior CR predicts initial levels of current CR for the majority of participants but is not predictive of the rate or degree of change in CR that occurs following exposure to university-level education.

Though the benefit of early-life education on late-life cognitive function is well reported (Anstey & Christensen, 2000; Lenehan et al., 2015), this research is the first to investigate the potential benefit of a period of formal education in later life to enhance CR. It also utilizes a multivariate estimation of both preexisting and current CR in order to provide an accurate evaluation of the potential benefit associated with the education intervention (Ward et al., 2015). However, it is important to note that the modeling approaches utilized rely on extrapolation from an incomplete dataset. The THBP is an ongoing study, and it will be interesting to see whether these findings are robust once the full sample proceeds through all of the time points in future years. There are a number of limitations that should be noted in interpreting the results of the present study. Noticeably, the control group reaches just the minimum sample size of 100, which is typically preferred for latent growth modeling (Curran, Obeidat, & Losardo, 2010). The total number of person-by-time observations influences statistical power (Curran et al., 2010). Additionally, due to the progressive recruitment of participants into the THBP over a 4-year period, the models estimated are based on extrapolation from an incomplete data set, in which some individuals have only one or two obser-vations over time. This may result in increased within-group variability, as indicated by a larger standard error of the mean, which is more evident in the control group and therefore less power to detect significant intercept and slopes. Future research will need to reexamine the findings of the present analysis once the complete THBP participant pool finishes assessment over all time points.

It is also important to note that although unavoidable due to the design of the present study, the recruitment of voluntary participants into the THBP may result in a self-selection bias of older adults with an interest in pursuing further education and a history of higher level secondary school education required for entry into University-level study. Therefore the participants in the THBP are likely to have a higher level of prior education and a greater interest in education than WOULD the wider community. However, it is important to note that the THBP is designed to determine whether increased mental activity in later life is beneficial to cognitive function in an aging population. As such, the THBP has utilized higher education as the tool for stimulating mental activity. A finding of increased cognitive capacity would be evidence of an effect of increased mental activity that could be achieved through the pursuit of mentally stimulating activities distinct from university-level education.

To summarize, the findings of the present study indicate that engaging healthy older adults in university-level education FOR a minimum of 12 months results in a measureable and significant increase in cognitive reserve. Future research is planned to determine whether this increase in cognitive reserve is sufficient to offset age-related cognitive decline and, further, whether this increase in CR mitigates the risk for degenerative conditions such as dementia or delays the onset of clinical symptoms of dementia in those at risk of dementia.

References

Anstey, K., & Christensen, H. (2000). Education, activity, health, blood pressure and apolipoprotein E as predictors of cognitive change in old age: A review. *Gerontology, 46,* 163–177. http://dx.doi.org/10.1159/000022153

Ball, K., Berch, D. B., Helmers, K. F., Jobe, J. B., Leveck, M. D., Marsiske, M., . . . Willis, S. L. (2002). Effects of cognitive training interventions with older adults: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 288,* 2271–2281. http://dx.doi.org/10.1001/jama.288.18.2271

Barnes, L. L., Mendes de Leon, C. F., Wilson, R. S., Bienias, J. L., & Evans, D. A. (2004). Social resources and cognitive decline in a population of older African Americans and whites. *Neurology, 63,* 2322–2326. http://dx.doi.org/10.1212/01.WNL.0000147473.04043.B3

Bennett, D. A., Wilson, R. S., Schneider, J. A., Evans, D. A., Mendes de Leon, C. F., Arnold, S. E., . . . Bienias, J. L. (2003). Education modifies the relation of AD pathology to level of cognitive function in older persons. *Neurology, 60,* 1909–1915. http://dx.doi.org/10.1212/01.WNL.0000069923.64550.9F

Curran, P. J., Obeidat, K., & Losardo, D. (2010). Twelve frequently asked questions about growth curve modeling. *Journal of Cognition and Development, 11,* 121–136. http://dx.doi.org/10.1080/15248371003699969

Dartigues, J.-F., Gagnon, M., Letenneur, L., Barberger-Gateau, P., Commenges, D., Evaldre, M., & Salamon, R. (1992). Principal lifetime occupation and cognitive impairment in a French elderly cohort (Paquid). *American Journal of Epidemiology, 135,* 981–988.

Donnell, A. J., Pliskin, N., Holdnack, J., Axelrod, B., & Randolph, C. (2007). Rapidly-administered short forms of the Wechsler Adult Intel-

ligence Scale-3rd ed. *Archives of Clinical Neuropsychology, 22,* 917–924. http://dx.doi.org/10.1016/j.acn.2007.06.007

Dufouil, C., Alpérovitch, A., & Tzourio, C. (2003). Influence of education on the relationship between white matter lesions and cognition. *Neurology, 60,* 831–836. http://dx.doi.org/10.1212/01.WNL.0000049456.33231.96

Ertel, K. A., Glymour, M. M., & Berkman, L. F. (2008). Effects of social integration on preserving memory function in a nationally representative US elderly population. *American Journal of Public Health, 98,* 1215–1220. http://dx.doi.org/10.2105/AJPH.2007.113654

Feingold, A. (2009). Effect sizes for growth-modeling analysis for controlled clinical trials in the same metric as for classical analysis. *Psychological Methods, 14,* 43–53. http://dx.doi.org/10.1037/a0014699

Foubert-Samier, A., Catheline, G., Amieva, H., Dilharreguy, B., Helmer, C., Allard, M., & Dartigues, J.-F. (2012). Education, occupation, leisure activities, and brain reserve: A population-based study. *Neurobiology of Aging, 33,* 423.e15–423.e25. http://dx.doi.org/10.1016/j.neurobiolaging.2010.09.023

Frisoni, G. B., Rozzini, R., Bianchetti, A., & Trabucchi, M. (1993). Principal lifetime occupation and MMSE score in elderly persons. *Journal of Gerontology, 48*(6), S310–S314. http://dx.doi.org/10.1093/geronj/48.6.S310

Geiser, C. (2013). *Data analysis with MPlus.* New York, NY: Guilford Press.

Harrison, S. L., Sajjad, A., Bramer, W. M., Ikram, M. A., Tiemeier, H., & Stephan, B. C. M. (2015). Exploring strategies to operationalize cognitive reserve: A systematic review of reviews. *Journal of Clinical and Experimental Neuropsychology, 37,* 253–264. http://dx.doi.org/10.1080/13803395.2014.1002759

International Wellbeing Group. (2006). *Personal wellbeing index* (4th ed.). Melbourne: Australian Centre on Quality of Life, Deakin University.

Jackson, K. M., & Sher, K. J. (2005). Similarities and differences of longitudinal phenotypes across alternate indices of alcohol involvement: A methodologic comparison of trajectory approaches. *Psychology of Addictive Behaviors, 19,* 339–351. http://dx.doi.org/10.1037/0893-164X.19.4.339

Jorm, A. F., Rodgers, B., Henderson, A. S., Korten, A. E., Jacomb, P. A., Christensen, H., & Mackinnon, A. (1998). Occupation type as a predictor of cognitive decline and dementia in old age. *Age and Ageing, 27,* 477–483. http://dx.doi.org/10.1093/ageing/27.4.477

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent class growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2,* 302–317. http://dx.doi.org/10.1111/j.1751-9004.2007.00054.x

Jurica, P. J., Leitten, C. L., & Mattis, S. (2001). *Dementia Rating Scale-2 (DRS-2): Professional manual.* Odessa, FL: Psychological Assessment Resources.

Kramer, A. F., Hahn, S., Cohen, N. J., Banich, M. T., McAuley, E., Harrison, C. R., . . . Colcombe, A. (1999, July 29). Ageing, fitness and neurocognitive function. *Nature, 400,* 418–419. http://dx.doi.org/10.1038/22682

Lenehan, M. E., Summers, M. J., Saunders, N. L., Summers, J. J., & Vickers, J. C. (2015). Relationship between education and age-related cognitive decline: A review of recent research. *Psychogeriatrics, 15,* 154–162. http://dx.doi.org/10.1111/psyg.12083

Lövdén, M., Ghisletta, P., & Lindenberger, U. (2005). Social participation attenuates decline in perceptual speed in old and very old age. *Psychology and Aging, 20,* 423–434. http://dx.doi.org/10.1037/0882-7974.20.3.423

Lubben, J., & Gironda, M. (2003). Centrality of social ties to the health and well being of older adults. In B. Berkman & L. Harooytan (Eds.), *Social work and health care in an aging world* (pp. 319–350). New York, NY: Springer.

Manly, J. J., Byrd, D., Touradji, P., Sanchez, D., & Stern, Y. (2004). Literacy and cognitive change among ethnically diverse elders. *International Journal of Psychology, 39,* 47–60. http://dx.doi.org/10.1080/00207590344000286

Muthén, B. O., & Muthén, L. K. (1998–2012). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Roid, G. H., & Ledbetter, M. F. (2006). *WRAT4 progress monitoring version: Professional manual.* Lutz, FL: Psychological Assessment Resources.

Scarmeas, N., & Stern, Y. (2003). Cognitive reserve and lifestyle. *Journal of Clinical and Experimental Neuropsychology, 25,* 625–633. http://dx.doi.org/10.1076/jcen.25.5.625.14576

Schaie, K. W. (1989). The hazards of cognitive aging. *Gerontologist, 29,* 484–493. http://dx.doi.org/10.1093/geront/29.4.484

Snaith, R. P. (2003). The Hospital Anxiety and Depression Scale. *Health and Quality of Life Outcomes, 1*: 29. http://dx.doi.org/10.1186/1477-7525-1-29

Stern, Y. (2002). What is cognitive reserve? Theory and research application of the reserve concept. *Journal of the International Neuropsychological Society, 8,* 448–460. http://dx.doi.org/10.1017/S1355617702813248

Summers, M. J., Saunders, N. L., Valenzuela, M. J., Summers, J. J., Ritchie, K., Robinson, A., & Vickers, J. C. (2013). The Tasmanian Healthy Brain Project (THBP): A prospective longitudinal examination of the effect of university-level education in older adults in preventing age-related cognitive decline and reducing the risk of dementia. *International Psychogeriatrics, 25,* 1145–1155. http://dx.doi.org/10.1017/S1041610213000380

Tucker, A. M., & Stern, Y. (2011). Cognitive reserve in aging. *Current Alzheimer Research, 8,* 354–360. http://dx.doi.org/10.2174/156720511795745320

Valenzuela, M. J., & Sachdev, P. (2006). Brain reserve and dementia: A systematic review. *Psychological Medicine, 36,* 441–454. http://dx.doi.org/10.1017/S0033291705006264

Valenzuela, M. J., & Sachdev, P. (2007). Assessment of complex mental activity across the lifespan: Development of the Lifetime of Experiences Questionnaire (LEQ). *Psychological Medicine, 37,* 1015–1025. http://dx.doi.org/10.1017/S003329170600938X

Ward, D. D., Summers, M. J., Saunders, N. L., & Vickers, J. C. (2015). Modeling cognitive reserve in healthy middle-aged and older adults: The Tasmanian Healthy Brain Project. *International Psychogeriatrics, 27,* 579–589. http://dx.doi.org/10.1017/S1041610214002075

Wechsler, D. (2001). *Wechsler Test of Adult Reading (WTAR).* San Antonio, TX: Harcourt Assessment.

Whalley, L. J., Starr, J. M., Athawes, R., Hunter, D., Pattie, A., & Deary, I. J. (2000). Childhood mental ability and dementia. *Neurology, 55,* 1455–1459. http://dx.doi.org/10.1212/WNL.55.10.1455

Zumbo, B. D. (2007). Validity: Foundational issues and statistical methodology. In C. R. Rao & S. Sinharay (Eds.), *Handbook of statistics: Vol. 26. Psychometrics* (pp. 45–79). Amsterdam, The Netherlands: Elsevier Science B. V.

A Complementary Processes Account of the Development of Childhood Amnesia and a Personal Past

Patricia J. Bauer
Emory University

Personal-episodic or autobiographical memories are an important source of evidence for continuity of self over time. Numerous studies conducted with adults have revealed a relative paucity of personal-episodic or autobiographical memories of events from the first 3 to 4 years of life, with a seemingly gradual increase in the number of memories until approximately age 7 years, after which an adult distribution has been assumed. Historically, this so-called *infantile amnesia* or *childhood amnesia* has been attributed either to late development of personal-episodic or autobiographical memory (implying its absence in the early years of life) or to an emotional, cognitive, or linguistic event that renders early autobiographical memories inaccessible to later recollection. However, neither type of explanation alone can fully account for the shape of the distribution of autobiographical memories early in life. In contrast, the complementary processes account developed in this article acknowledges early, gradual development of the ability to form, retain, and later retrieve memories of personally relevant past events, as well as an accelerated rate of forgetting in childhood relative to adulthood. The adult distribution of memories is achieved as (a) the quality of memory traces increases, through addition of more, better elaborated, and more tightly integrated personal-episodic or autobiographical features; and (b) the vulnerability of mnemonic traces decreases, as a result of more efficient and effective neural, cognitive, and specifically mnemonic processes, thus slowing the rate of forgetting. The perspective brings order to an array of findings from the adult and developmental literatures.

Keywords: autobiographical memory, childhood amnesia, development, episodic memory, forgetting

For better or for worse, we all have a personal past. We have an historical self who has experienced a lifetime of events. Some of the events are mundane and have a short tenure in memory. For example, though we encode where we parked the car when we arrived at work in the morning, after retrieving the information at the end of the day, we commit no further effort to retaining it. Others of our experiences are defining moments in our lives, such as graduation from college, the birth of a child, or the death of a parent. Memories of these types of events not only are long lasting, they also are critically important to our sense of self—our sense of continuity over time rests on memories of events and experiences that took place in the past. In essence, we believe that we are the same person yesterday and today because we have memories of ourselves from the past. However, there is a striking discontinuity in our personal past. That is, most adults have few if any memories from the first 3 to 4 years of life. There is what appears to be a gradually increasing number of memories from the years of age 3

to 7, after which an adult-like distribution of personal memories is assumed. In this review, I advance a novel account of this discontinuity in terms of complementary processes that contribute to increases in the quality of personal memories and to decreases in their vulnerability to forgetting; thus, producing the characteristic distribution that is the hallmark of childhood amnesia.

Since the "amnesia" for events from the first years of life was first identified in the literature at the end of the 19th century (Henri & Henri, 1896, 1898; Miles, 1895), and named at the beginning of the 20th century (Freud, 1905/1953), there have been a number of empirical studies that establish its robust nature. As well, a number of theories as to the source of the amnesia have been proposed. As elaborated below, though the theories differ in specifics, they fall into one or the other of two categories. Some accounts emphasize late development of the ability to remember the past. They suggest that adults suffer from amnesia for early life events because in the period that eventually becomes obscured by the amnesia, children lack the capacity to create personal memories. It is only after the development of some criterial attribute that children begin to form and retain personal memories. By other accounts, early memories are formed but as a result of an emotional, cognitive, or linguistic change, they later become inaccessible to recall and functionally disappear. Thus, these accounts recognize a developmentally early capacity to form and retain personal memories and hypothesize causes for their later loss to recollection.

As will become apparent, a corollary implication of existing theoretical accounts of childhood amnesia is of discontinuous processes in personal memory. In the case of theories that emphasize late emergence of a new ability to remember, there is thought

to be a period of time before onset of the new ability during which individuals are unable to form, retrain, and later retrieve personal memories. In the case of theories that emphasize the later functional disappearance of early memories, there is thought to be a change in emotional, cognitive, or linguistic processes that renders early events inaccessible to later recollection. The discontinuity inherent in these accounts stems from the fact that they emphasize one or the other side—but not both sides—of the mnemonic coin. That is, theories that posit late development of the capacity to create personal memories do not take into account the possibility that memories of early life events were formed but then were lost to recollection. Absent a focus on processes that contribute to the later inaccessibility of memory traces, in these accounts, the cause of later amnesia is failure to form memories in the first place. Conversely, theories that posit events that cause the later functional disappearance of early memories do not take into account the possibility that the memories that subsequently disappeared may have been especially vulnerable to forgetting because of poor quality of the traces themselves. Absent recognition of differences in the qualities of memories that are formed early versus later in life, the cause of later inaccessibility is an event that renders early memories lost to later recollection.

The purpose of the present theoretical review is to advance a novel account of the phenomenon of the amnesia of childhood that draws from each of these traditional categories of explanation, yet is not a straightforward integration of them. Unlike the traditional categories of explanation—that emphasize one or the other, but not both sides of the mnemonic coin—the account I develop is explicitly complementary. On one side of the mnemonic coin are developmental changes in a number of factors and processes that facilitate encoding, consolidation, and later retrieval of memory traces, eventually resulting in a corpus of personal memories. On the other side of the mnemonic coin are a number of factors and processes that undermine the integrity of memory traces, eventually rendering some inaccessible to later recollection. More important, it is only by considering both the processes that function to increase the quality of memory traces and the processes that function to undermine them that we can explain the phenomenon of childhood amnesia—neither set of processes alone is sufficient to account for the full pattern of data summarized in this review.

The complementary processes account also departs from traditional explanations of childhood amnesia in its emphasis on essential continuities in memory over the course of development. That is, the complementary processes account recognizes the roots of personal memory even in infancy, and a continuous course of development throughout childhood. Over time, the developmental changes permit formation of representations of personally relevant past events that feature more, better elaborated, and more tightly integrated, personal-episodic or autobiographical elements, relative to those formed earlier in life. In other words, the capacity to form personal memories does not emerge later in development, but gradually improves over the course of childhood. The account also recognizes normative forgetting processes that operate throughout the period eventually obscured by childhood amnesia (and beyond). Because early in development, memory processes are carried out by a relatively immature neural substrate, memory representations formed early in life are especially vulnerable to forgetting. In other words, early memory representations do not become inaccessible as a result of an event or qualitative change,

but are lost as a result of normative forgetting processes that weaken memory traces over time. The interaction effect is childhood amnesia—the result of lower quality raw materials operated upon by relatively inefficient and ineffective mnemonic processes. The complementary processes account brings order to an array of findings from the adult and developmental literatures. It is grounded in contemporary understanding of neural, cognitive, and specifically mnemonic developmental processes that help to explain the later inaccessibility of memories of early life events and experiences.

Personal-Episodic or Autobiographical Memories: Definition and Importance

Personal memories are not alone in the vast corpus of representations of past events available to healthy adults and children alike. Rather, they are but one of many types or forms of memories. In the parlance of an influential taxonomy of types of memory (Tulving, 1972, 1983), they are episodic memories: memories of specific past events that happened at a particular place and time. Retrieval of memories of these events often is accompanied by a sense of mentally placing oneself in the past as if reliving the experience. This *autonoetic awareness* (e.g., Baddeley, Eysenck, & Anderson, 2009; Tulving, 2002, 2005) of the experience as having taken place in the past is accompanied by vivid recollection of what happened, when, and where. Episodic memory frequently is operationalized in terms of an individual's recall or recognition of items from a studied list, recall of a passage of text, memory for faces, and so forth.

Episodic memory typically is contrasted with semantic memory (Tulving, 1972, 1983), the latter of which comprises our store of world knowledge. Unlike episodic memories, semantic memories are not located in a particular place or time. For example, we may know that the capital of Georgia is Atlanta, yet unless there was something especially noteworthy or significant about the episode during which we learned this fact, we do not have memory for when or where it was acquired. Even more phenomenologically distant from episodic memory is so-called nondeclarative or implicit memory (e.g., Squire, 1987; Squire, Knowlton, & Musen, 1993). These representations are based on past experiences that, like semantic memories, are not located in specific place and time. However, unlike semantic (as well as episodic) memories, they can influence behavior without conscious recollection. Instead, nondeclarative memories are of motor patterns for how to ride a bicycle, for example, or conditioned or reflexive responses to stimuli. They influence our behavior, to be sure, but they do so without being brought to consciousness (see for, e.g., Bauer, 2013 and Squire et al., 1993, for further development of distinctions among types or forms of memory in the developmental and adult literatures, respectively).

The distinctions among different types or forms of memory are critically important to the effort to explain childhood amnesia because the amnesia does not obscure all types of memories. Even infants and very young children learn and remember a great many things. Infants recognize their caregivers over time and across contexts, they learn to walk and talk, and by the preschool years, children have accrued a great deal of semantic or factual knowledge about the world. These mnemonic accomplishments persist beyond infancy and childhood—they are not obscured by child-

hood amnesia. Rather, childhood amnesia is the relative paucity of *episodic* memories for events and experiences from the first years of life. However, it is more than amnesia for items on a list or passages of text, for example. It is the relative paucity of episodic memories of events and experiences that are about one's self, and about which one has emotions, thoughts, reactions, and reflections. Because of their reference and relevance to the self, these memories are considered not only episodic, but self defining and *autobiographical* (see Bauer, 2013, 2014; Fivush & Zaman, 2014). Thus, the period that is obscured by the "peculiar amnesia of childhood" (Freud, 1920/1935) is one from which we are lacking autobiographical memories.

The absence—or at best, sparse representation—of autobiographical memories from the first years of life leaves a salient void in one respect in particular, namely, continuity of self. As noted briefly above, it is on the basis of memories for past events that we recognize ourselves as continuous in time (e.g., Habermas & Köber, 2014). In general, the ability to remember one's self in the past is a precondition for a sense of personal continuity (e.g., Prebble, Addis, & Tippett, 2013), and personal memories ground a stable and enduring representation of the self over time (e.g., Bauer, Tasdemir-Ozdes, & Larkina, 2014; Bluck & Alea, 2008; Conway, 2005; McAdams, 1995; Wilson & Ross, 2003). The fact that most adults suffer a paucity of autobiographical memories from the first years of life means that although they had a physical existence before their earliest memory, they experience a discontinuity of psychological self. The importance of autobiographical memories helps to explain why they have been the focus of significant research attention, as well as why their absence from the first years of life has been of sustained interest for more than a century.

At this point it is important to recognize that there is not universal agreement on the definition of autobiographical memory. There is consensus that to be "admitted" into the category of autobiographical memories, episodic memories must be of specific past events and experiences that are about one's self, and about which one has emotions, thoughts, reactions, and reflections. However, some perspectives would consider these criteria to be necessary, though not sufficient. Additional criteria suggested in the literature are that autobiographical memories are (a) of discrete, one-time-only or unique events, as opposed to recurring events; (b) expressed verbally; and (c) long lasting (e.g., Nelson, 1993; see Bauer, 2007, 2014, for reviews). By other definitions, even this larger set of criteria would be considered insufficient to differentiate autobiographical from episodic memory. In the literature on self identity, for example, autobiographical memory is defined as a capacity that permits construction of a sequence of memories of temporally linked events, such as expressed in a life story or autobiography (e.g., Fivush, Habermas, Waters, & Zaman, 2011; Fivush & Zaman, 2014; Habermas & Bluck, 2000; McAdams, 2001; see also Conway & Pleydell-Pearce, 2000; Thomsen, 2009). From this perspective, retrieval of an autobiographical memory involves *autobiographical consciousness*, defined as a form of consciousness of a present self who is different from—yet temporally linked to—the past self who experienced the event (Fivush, 2012; see also Fivush & Zaman, 2014). It is only in adolescence that individuals begin to create narratives of past events that bear these features, leading to the suggestion that autobiographical memory is a capacity that emerges only in adolescence.

The perspective adopted in this review is that autobiographical memory is adequately defined as a system that supports formation, retention, and later retrieval of episodic memories of specific events that are spatially and temporally localized, as well as self-referential, as evidenced by personal perspective on or evaluation of the experience. It need not be restricted to discrete, one-time only or unique events: recurring events, such as "Sunday dinners at grandma's when I was a kid," can be localized (albeit on a larger temporal scale) and also are self-referential and defining (e.g., Waters, Bauer, & Fivush, 2014). Moreover, Rubin and Umanath (2015) argue that merged representations of recurring events are the psychological equivalents of single events (see also arguments by Brewer, 1986). Verbal expression is important to investigation of autobiographical memory because it makes it easier to determine whether representations of past events feature evidence of self-reference, and whether retrieval is accompanied by autonoetic awareness of the experience as having taken place in the past. When these features are expressed verbally, we can be confident that the underlying memory representation bears these characteristics. However, verbal descriptions are not isomorphic with memory representations. As such, the absence of verbal expression of these features should not be taken as evidence that memory representations are lacking of them. The criterion that episodic memories must be long-lasting to be considered autobiographical lacks specificity (how long is "long"?) and is potentially circular, owing to the tendency to equate "long-lasting" with "personally relevant"—if the memory is not long-lasting, then it must not have been important to the self.

The final suggested definition of autobiographical memory in terms of developments in narrative self expression in adolescence is useful for understanding changes in narrative production that occur at that time. However, I argue that it is not useful to the goal of understanding childhood amnesia. As elaborated below, among adults, childhood amnesia is dense for the first 3 to 4 years of life and then begins to "lift," as evidenced by a steadily increasing number of memories that are available for recollection. In light of this distribution, we must admit either that autobiographical memories are apparent well before adolescence, or that none of the preadolescent memories retained by adults is autobiographical. If one accepts the latter, then a necessary corollary is that autobiographical memory has virtually nothing to do with childhood amnesia. Given the substantial theoretical departure this would represent, it seems better advised to view the developments taking place in adolescence as reflective not of the emergence of autobiographical memory, per se, but of construction of an autobiography, life story, or life narrative, which takes autobiographical memories as its raw materials. This perspective is consistent with Rubin and Umanath's (2015) view that narrative organization is a characteristic of autobiographical memory, but it is not necessary for it.

Discontinuity in the Personal Past: Childhood Amnesia

Since the beginning of the 20th century, the amnesia that adults experience for early life events has been known as *infantile* or *childhood amnesia* (Freud, 1905/1953). It is recognized as having two phases (Pillemer & White, 1989; see Bauer, 2007, for updated discussion), which are schematically depicted in Figure 1. From the first phase—before age 3 to 4 years—most adults have few if

any autobiographical memories (solid gray bars). From the second phase—between the ages of 5 and 7 years—adults have a smaller number of autobiographical memories than would be expected based on forgetting alone (striped bars). It is only from later in the first decade of life that most adults are able to recall a significant number of past events that are spatially and temporally localized, and which have some degree of personal relevance or significance (solid black bars). Although this "peculiar amnesia of childhood" (Freud, 1920/1935) is considered an adult phenomenon, as discussed in a later section, there is a small but increasing body of evidence that by the end of the first decade of life, children also begin to experience it. Before reviewing that literature, I elaborate on the two major phases of childhood amnesia among adults.

Average Age of Earliest Memory

The earliest research on the phenomenon of childhood amnesia among adults was published at the close of the 19th century. Miles (1895) conducted a survey of adults' childhood experiences and among other things, asked them to think about the earliest event they could remember, and how old they were at the time. This and subsequent such surveys (e.g., Dudycha & Dudycha, 1933a, 1933b; Henri & Henri, 1895, 1896, 1898; Kihlstrom & Harackie-wicz, 1982) have produced one of the most consistent and robust findings in the psychological literature, namely, that the average age of earliest memory among adults in Western cultures is age 3 to 4 years (see, e.g., Wang, 2006, 2014, for discussions of cross-cultural differences in average age of earliest memory). Moreover, the same average age of earliest memory is found whether the source of data is a survey, free recall (e.g., Bauer et al., in press; Waldfogel, 1948; Weigle & Bauer, 2000; West & Bauer, 1999), or response to a cue word prompt (e.g., Bauer & Larkina, 2014; Rubin & Schulkind, 1997; though see Wang, Conway, & Hou, 2004, for evidence that repeated probes can produce earlier estimates). The effect also is impervious to age-cohort effects: the same general pattern is obtained from individuals 20 years of age at the time the memories are prompted and individuals 60 to 70 years of age at the time the memories are elicited (see, Rubin, 2000, for review), even though for older adults, many more years have passed since childhood. The same average age of earliest memory is found even when respondents are asked to remember a specific event the date of which is clearly known, such as the birth
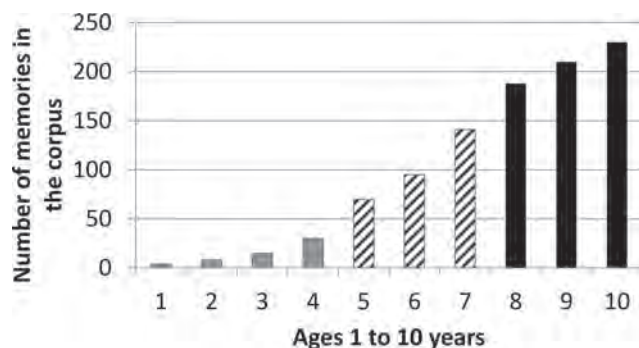

Figure 1. Schematic depiction of the distribution of memories across the first decade of life from a traditional perspective, suggesting a gradually increasing number of memories with age.

of a younger sibling (e.g., Sheingold & Tenney, 1982; Usher & Neisser, 1993).

The robust nature of the average age of earliest memory among adults obscures the substantial individual differences in the phenomenon. From the beginning of research on the topic, individual differences in the age of earliest memory have been apparent. In reports that provide information on variability, most feature at least a small number of instances of memories from before the age of 2 years (e.g., Bauer et al., in press; Dudycha & Dudycha, 1933a, 1933b; Henri & Henri, 1896, 1898; West & Bauer, 1999). Memories from at least some respondents from age 2 years are more the rule than the exception (e.g., Eacott & Crawley, 1998; Usher & Neisser, 1993). Conversely, some adults have earliest memories from later in childhood: their "earliest" memories are from as old as 6 to 9 years of age (e.g., Bauer & Larkina, 2014; West & Bauer, 1999). There also are individual differences among adults in the density of early memories. Some adults recall many memories from their childhood years, whereas others remember only a few, with many months between them (Bauer, Stennes, & Haight, 2003; Jack & Hayne, 2010; Weigle & Bauer, 2000; West & Bauer, 1999).

Characteristic Distribution of Early Memories

The second component of the definition of childhood amnesia is that from the ages of roughly 3 or 4 to 7 years, the number of memories that adults are able to retrieve increases gradually yet is smaller than the number expected based on forgetting alone (Pillemer & White, 1989). After age 7 years, a steeper, more adult-like distribution becomes apparent. The underrepresentation of memories from before age 7 years was empirically demonstrated in a seminal article by Wetzler and Sweeney (1986), using data from Rubin (1982). Rubin asked young adults to think of past events related to each of over 100 cue words (e.g., *cup*, *chair*, and *tree*), and to estimate their age at the time of the event. To the data, Wetzler and Sweeney fitted a power function that in many investigations (e.g., Crovitz & Schiffman, 1974; Rubin & Wenzel, 1996; Rubin, Wetzler, & Nebes, 1986) has been shown to capture the distribution of memories across the life span. As discussed by Rubin and Wenzel (1996), the power function (e.g., Wickelgren, 1974, 1975) implies that equal ratios of time ($t_1/t_2 = t_3/t_4$) will result in equal ratios of recall ($recall_1/recall_2 = recall_3/recall_4$). Thus, for example, if Time 2 recall was 90% of Time 1 recall, then Time 4 recall would be 90% of Time 3 recall (i.e., assuming equal ratios of time). As a result of the constant ratio, over time, forgetting actually slows (i.e., smaller absolute numbers of memories are lost over each unit of time), presumably as a result of memory trace consolidation (see, e.g., Wixted, 2004, for discussion). Wetzler and Sweeney found that the power function was a poor fit to data from birth to age 6 years, implying accelerated forgetting of memories from ages 6 and below. Memories from age 7 years were excluded from the analysis because age 7 years was considered the "inflection point" for childhood amnesia: after age 7 years, the rate of forgetting is assumed to be adult-like. Consistent with this suggestion, Wetzler and Sweeney found that the power function was a good fit to data from age 8 to adulthood (see Bauer, 2007, for additional discussion).

Traditional Explanations of Childhood Amnesia

Given the relevance and importance of autobiographical memories to the self—and the robustness of the finding of a paucity of such memories from the first years of life—it is fitting that childhood amnesia has received a great deal of theoretical attention. Though there are a number of theories that differ in their specifics (see Bauer, 2007, for a review), the explanations can be summarized as belonging to two general categories (Bauer, 2014). As introduced earlier, one category of accounts emphasizes the late emergence of autobiographical memory. By these explanations, adults experience amnesia for early childhood events because as children, they lacked the fundamental capacity to form such memories. By these accounts, autobiographical memory is a later developing achievement, one dependent on general or specific cognitive changes that permit personal memories to be formed, retained, and later retrieved. It is only once the capacity emerges that autobiographical remembering begins.

The second category of accounts of childhood amnesia emphasizes the functional disappearance of early memories—that is, changes that render once accessible memories inaccessible to recollection. By these explanations, autobiographical memories of events from the first years of life are formed and presumably can be recollected by children while they are in the period that eventually becomes obscured by childhood amnesia, yet the memories later become inaccessible and thus functionally disappear.

Traditional explanations of childhood amnesia differ in their emphasis, yet what they have in common is that they implicate one or the other side of the mnemonic coin, but not both. That is, they explain childhood amnesia *either* in terms of something that must develop to permit the capacity for self-referential memories of past events, *or* in terms of something that happens to make early memories inaccessible to later recollection—they fail to take into account the complementary mnemonic processes. That is, the former category puts all of its explanatory eggs into a basket that emphasizes late emergence of autobiographical remembering; the possibility that autobiographical memories are formed early in life but subsequently forgotten is not part of the explanation. The latter category puts all of its explanatory eggs into a basket that emphasizes the later functional disappearance of autobiographical memories; there is no recognition of relative vulnerabilities in the memory traces formed during the period eventually obscured by childhood amnesia. After reviewing these traditional categories of explanation of childhood amnesia, I advance a novel account that integrates the complementary processes implicated in the explanations, by emphasizing increases in the quality of autobiographical memories over the course of development, as well as decreases in the vulnerability of autobiographical memory traces. The account emphasizes the essential continuity of the processes over developmental time.

Emphasis on Late Emergence: Memories Are Not Accessible Because They Were Not Formed

The first major category of perspectives on the source of childhood amnesia is that adults have few autobiographical memories from before the ages of 5 to 7 years because during this period, they lacked the capacity to form and retain them, because of general or more specific cognitive deficits.

The suggestion that general cognitive deficits explain the relative paucity of memories from early in life is perhaps most notably associated with Piaget (1962). Though Piaget did not advance a theory of childhood amnesia, per se, his theoretical perspective provided a compelling explanation for it nonetheless. He maintained that for the first 18 to 24 months of life, infants and children did not have the capacity for symbolic representation. As a result, they could not mentally re-present objects and entities in their absence. Thus, they had no mechanism for recall of past events. Beyond 24 months and through ~5 to 7 years, children were thought to lack the cognitive structures that would permit them to organize events along coherent dimensions that would support later retrieval. One of the most significant dimensions that Piaget suggested preschool-age children lacked was an understanding of temporal order. Specifically, he suggested that it was not until children were ~5 to 7 years of age that they developed the ability to sequence events temporally (Inhelder & Piaget, 1958). Lacking this organizational device, children were not able to form coherent memories of the events of their lives.

A contemporary version of a general cognitive deficit account has been advanced by Olson and Newcombe (2014). They suggest that before the age of 2 years, children lack the ability to bind together or relate elements of representations of events, a prerequisite for formation of episodic memories (discussed in more detail in a later section). Between the ages of 2 and 6 years, children may bind elements of events together, yet do so in a manner that fails to ensure persistence of memories over time. Consistent with this suggestion, Olson and Newcombe highlight young children's difficulties remembering the correct sources of their experiences (so-called source memory, e.g., Drummey & Newcombe, 2002; Riggins, 2014), and their difficulties creating conjunctions between items and their locations (e.g., Bauer, Doydum, Pathman, Larkina, Güler, & Burch, 2012; Sluzenski, Newcombe, & Kovacs, 2006; though see Bauer, Stewart, White, & Larkina, in press).

There also are suggestions that specific conceptual, linguistic, or mnemonic changes play a role in the explanation of childhood amnesia, rather than global cognitive change. By some accounts, adults have few memories from infancy because for the first 2 years, infants lack the concept of a self around which memories can be organized (e.g., see Howe & Courage, 1993, 1997, for reviews). By other accounts, beyond a physical sense of self, development of autobiographical memory awaits a subjective self who evaluates and takes personal perspective on life events (e.g., Fivush, 2014). Absent these developments, there is no *auto* to lend the *auto*biographical character to episodic memories. By other accounts, for the first 5 to 7 years of their lives, children lack autonoetic consciousness, rendering it impossible for them to recognize that the source of their mental experience is a representation of a past event (e.g., Perner & Ruffman, 1995), or to engage in the subjective mental time travel that accompanies episodic and autobiographical memory retrieval (e.g., Suddendorf, Nielsen, & van Gehlen, 2011; Tulving, 2005; Wheeler, 2000). Each of these explanations implicates a different specific component ability. However, what the suggestions have in common is the perspective that as a result of some deficit, although children may remember past events, their memories are lacking in the qualities that typify the autobiographical memories formed by older children and adults. It is only once the general or specific cognitive or conceptual ingredients that are missing from early memories become

available that children begin to form, retain, and later retrieve memories that are autobiographical.

A well-articulated example of this perspective was provided by Nelson and Fivush (2004). As depicted in Figure 2 (reproduced from Nelson & Fivush, 2004), they suggested that it is not until 5 years of age that the many cognitive dimensions required for encoding, retention, and later retrieval of memories have reached a sufficient level of development to support autobiographical memory. They noted the necessity for autobiographical memory of developments in self concept, language and narrative, theory of mind, understanding of time and place, subjective sense of self, mental time travel, and autonoetic awareness, and others, all of which undergo development from infancy through early childhood, culminating in the capacity to form autobiographical memories. Until that time, children may have semantic and perhaps even episodic memories, but their memories are not autobiographical. In summary, what these perspectives have in common is the assumption that the reason adults are unable to recollect early childhood is because the memory system available to them as children was lacking in ingredients essential for formation, retention, and later retrieval of autobiographical memories. As a consequence, not only adults—but also children—lack autobiographical memories from the first years of life.

Emphasis on Functional Disappearance: Memories Are Formed but Become Inaccessible

The second major category of perspectives on the source of childhood amnesia is characterized by the assumption that young children and perhaps even infants form memories of the events of their lives, but that the memories subsequently become inaccessible. Perhaps the best known (and most infamous) of such accounts is Freud's (1905/1953) psychodynamic theory. He remarked that "We forget of what great intellectual accomplishments and of what complicated emotions a child of four years is capable . . . Yet, in spite of this unparalleled effectiveness they (memories of early life events) were forgotten!" (Freud, 1905/1953, p. 64). Freud suggested that the memories became inaccessible as a result of repression of inappropriate or disturbing content of early, often traumatic (because of their sexual nature) experiences. Memories of events that were not repressed were altered to remove the offending content. Freud suggested that the negative emotion in these memories was screened off, leaving only bland skeletons of once-significant experiences (Freud, 1916/1966).

More contemporary accounts also makes the assumption that memories of early life events are formed but become inaccessible, but for cognitive or linguistic rather than emotional reasons. These perspectives have in common the suggestion that different times or phases of life are experienced through different cognitive structures or "lenses." The structures of one life period are considered sufficiently different from those for another that memories created with one set of structures are inaccessible once new structures become dominant. By some accounts the structures differ in the extent to which they are reliant on language (e.g., Neisser, 1962). Because infants lack language and very young children lack many nuances of language, they encode memories visually or imaginally, but not symbolically. The suggestion is that with the advent of language skills, exclusively nonverbal encoding gives way to primarily verbal encoding. As the system becomes more and more saturated with language, it becomes increasingly difficult to gain access to memories encoded without language (Neisser, 1962). The result is that early memories become inaccessible.

Different lenses or cognitive structures may result not only from the linguistic revolution, but from changes over life periods, each of which has a distinctive sense of self, with different hopes, fears, and challenges, for example. Life periods may correspond to elementary versus secondary school versus college, or before
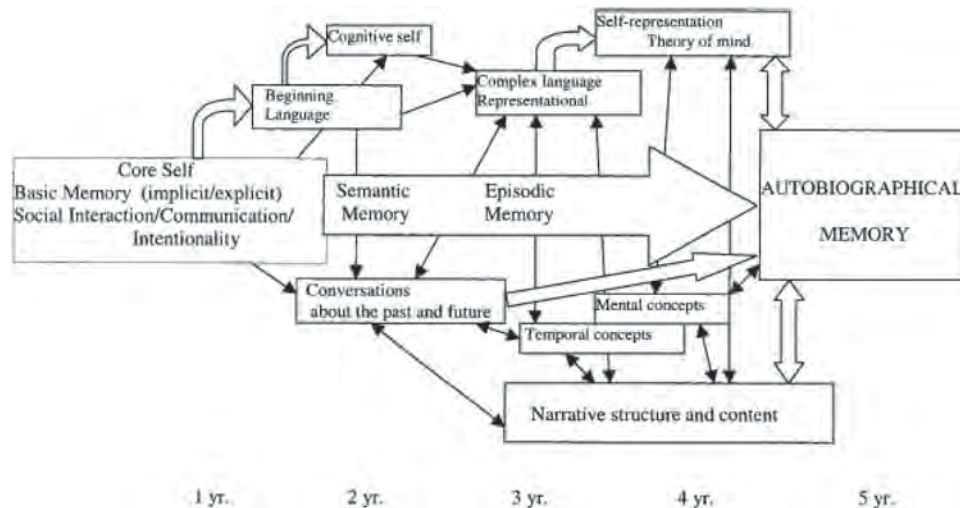
Figure 2. Depiction of contributors to and course of development of autobiographical memory provided in Nelson and Fivush (2004). Autobiographical memory is characterized as emerging at ~5 years of age, when developments in requisite contributing domains reach criterial levels. From "The Emergence of Autobiographical Memory: A Social Cultural Developmental Theory," by K. Nelson and R. Fivush, 2004, *Psychological Review, 111,* p. 490. Copyright 2004 by the American Psychological Association.

versus after marriage, or before versus after retirement (Conway, 1996; Conway & Pleydell-Pearce, 2000). Memories from prior lifetime periods may differ from those from the current period not only because time has passed, but because of the new sense of self that is associated with changes in thinking or world view. Though these perspectives differ in the specific causes of change in accessibility that they invoke, they have in common the assumption that the capacity to form memories of personally relevant events is present, even early in development. Some emotional, cognitive, or linguistic force renders the early memories inaccessible and functionally forgotten. Thus, unlike the category of explanations discussed above, in these explanations, there is nothing lacking in early memory. Instead, something happens later in development that renders early memories inaccessible.

Explanation of Childhood Amnesia in Terms of Complementary Processes

As just described, traditional accounts of childhood amnesia make one of two assumptions—either that early life is devoid of the ability to form autobiographical memories or that memories are formed in childhood but later become inaccessible to recollection. Thus, they emphasize either positive changes in memory (emergence of a new memory system) or negative changes (loss of accessibility), but neglect the complementary process. The result is that each faces challenges to the adequacy of the explanation of childhood amnesia that it offers. The category of explanations that emphasizes positive changes in memory adopts the stance that autobiographical memory is a developmentally later achievement, one that emerges only with either general or specific cognitive developmental changes. A corollary of the assumption is that memories formed before the target development(s) are not "autobiographical." Rather, they lack a feature or features considered defining of the category. At issue for these accounts is the fact that, as outlined below, features associated with autobiographical memories are apparent in memory behavior before the end of the second year of life; they become more and more prominent over the course of the preschool years. The challenge then becomes how to explain why memory representations that look and feel autobiographical, are not part of the autobiographical record.

The category of explanations that emphasizes functional disappearance of memories makes the assumption—either implicitly or explicitly—that autobiographical memories are formed, even early in life. Its challenge is to explain why autobiographical memories of early childhood become inaccessible to later recollection. Typical forgetting processes cannot be the answer because, as described earlier, the distribution of memories from early childhood is not well characterized by normal forgetting (Wetzler & Sweeney, 1986). Rather, the rate of forgetting is accelerated. Based on adult data alone—which is the empirical foundation from which these accounts were advanced—there is not a ready explanation for accelerated forgetting.

The perspective I advance in the balance of this review is that elements of both of these perspectives are part of the explanation of childhood amnesia. The amnesia can be understood in terms of the complementary processes that improve memory traces and that degrade them. The perspective highlights developments that result in formation of memory traces that bear more, better elaborated, and more tightly integrated autobiographical features. It also high-

lights developmental changes in the rate of forgetting associated with normative neural, cognitive, and mnemonic processes that result in decreases in the vulnerability of memory traces (i.e., at least through young adulthood). When considered together, the complementary processes—and their developmental dynamics—provide a ready explanation for the characteristic distribution of autobiographical memories across the life span and a more compelling explanation of the phenomenon of childhood amnesia (see also Bauer, 2007, 2008, for previews of this perspective).

The Quality of Memory Traces Increases Over Development

There are pronounced changes in memory behavior over the course of early childhood. There also are different perspectives on the implications of the changes. In traditional accounts of childhood amnesia, one or more of the changes is considered criterial for the emergence of autobiographical memory. A major source of evidence as to whether the criterial element (or elements) is present is children's verbal accounts or narrative descriptions of past events (e.g., Nelson & Fivush, 2004). For much of early development, children's memory reports omit some of the elements that are associated with autobiographical memory, leading to the conclusion that the elements also are missing from the underlying memory representations, thus rendering them nonautobiographical. The complementary processes perspective takes exception to these arguments on the bases that (a) verbal behavior alone is an insufficient source of evidence as to whether children "have" autobiographical memory; and (b) salient elements of autobiographical memory are apparent early in development—in verbal as well as nonverbal behavior—and that the capacity does not await an hypothesized criterial feature that is late to emerge. Rather than of late emergence of autobiographical memory, in the complementary processes perspective, changes in both verbal and nonverbal memory behavior are interpreted as evidence of gradual increases in the quality of memory traces—including autobiographical ones—over development, such that with development, memory traces bear more, better elaborated, and more tightly integrated personal-episodic or autobiographical features. These perspectives are presented in turn.

Developmental change in children's verbal behavior. A full autobiographical report features a number of elements, including *who* participated in the event, *what* happened, *where* and *when* the event took place, and *why* the sequence of actions unfolded as it did. It also features information about *how* the participants in the event reacted to it in terms of their emotions, thoughts, or evaluations of the event. The latter element is especially important to establishing the self-referential nature of autobiographical memories. Furthermore, the elements are presented in a coherent manner, allowing the listener (or reader) to understand the theme of the event, the context in which it took place, and the chronology of actions (Reese, Haden, Baker-Ward, Bauer, Fivush, & Ornstein, 2011).

There are marked changes in children's verbal behavior throughout early childhood. In their earliest verbal reports of past events, young children frequently omit one or more elements of a complete and coherent story. Instead, they typically merely confirm or deny information provided by another. For example, an adult mentions a recent visit to the zoo and offers the observation that the child

enjoyed the animals, and the child responds with an enthusiastic "Yes!" At around the age of 3 years, children begin contributing memory content. However, they frequently include only the most crucial elements such as who and what ("I played"). They omit many of the elements that make for a good story, such as where and when the event occurred, and why it happened as it did (see Bauer, 2013, 2014; Nelson & Fivush, 2004, for reviews). Over the course of the preschool and early school years, children take on increasingly active roles in conversations. They contribute more of the elements of a complete verbal report (i.e., the *who*, *what*, *where*, *when*, *why*, and *how* of events), more descriptive details, and more evaluative information, thereby adding texture and obvious self-relevance to their narratives (e.g., Bauer & Larkina, 2014; Haden, Haine, & Fivush, 1997).

Traditional interpretation of omissions from children's verbal behavior. By traditional criterial accounts, because children's early memory reports lack some of the features associated with autobiographical memories, their memories may be considered episodic (or even semantic, Nelson, 1993), but not autobiographical. This provides a ready account for childhood amnesia: memories formed in the first 5 to 7 years are not autobiographical, thus explaining the relative paucity among adults of autobiographical memories from this life period. Autobiographical memory is recognized only with verbal evidence that the event being recalled is located in specific time and place and is self-referential, as indicated by personal or evaluative perspective, or even that the event is part of an extended life narrative (e.g., Fivush, 2012; Fivush & Zaman, 2014).

Exception to the traditional criterial view. As previewed above, the complementary processes account developed here takes exception to the traditional criterial view on the grounds that absence of evidence does not constitute evidence of absence. In other words, the fact that young children's verbal reports do not feature all of the elements that are associated with autobiographical memories does not mean that the elements are missing from the memory representations and does not license the conclusion that the memory is not autobiographical. Moreover, though one or more elements characteristic of autobiographical memory may be missing from *any given* report that a young child provides about a past event, there does not seem to be a single element or feature that is missing from *all* reports. This calls into question the suggestion that young children lack a criterial or defining feature of autobiographical memory. In addition, as reviewed below, features characteristic of autobiographical memory are apparent at least by the end of the second year of life; they become increasingly obvious over childhood, such that reports feature more personal-episodic or autobiographical elements and the elements are better elaborated and more tightly integrated with one another. Based on this evidence, I suggest that autobiographical memory is best conceived not as a classical concept with defining features, but in terms of a prototype or family resemblance with characteristic features (see also Bauer, 2007, 2008, 2012, 2014). The essence of the argument is that, as more and more of the features associated with autobiographical memory are included in the memory trace, memories become increasingly autobiographical.

Absence of features from verbal reports does not imply absence from memory. As just discussed, the fact that early verbal accounts of past events typically omit features associated with autobiographical memories—especially location in time and place

and a subjective personal perspective—has been interpreted to suggest that these elements are missing from the representation and thus that early memories are not autobiographical. Verbal reports and memory representations are not isomorphic, however. When elements of autobiographical memory are featured in verbal reports, it is reasonable to conclude that the elements also are part of the representation. However, memory representations may include features not expressed in verbal form. I offer two sources of evidence to make the case—one from children and the other from adults—as well as an argument regarding sufficiency.

One source of evidence that memory representations may include mnemonic features that are not expressed 2012 verbally comes from a prospective study of children's recall of early life events (Bauer & Larkina, 2013). At the age of 3 to 4 years, children were asked to recall a number of events from the recent past. They featured an average of 3.59 narrative elements in their reports (out of a possible of eight narrative features: *who*, *what-object*, *what-action*, *where*, *when*, *why*, *how-description*, and *how-explanation*). Seemingly consistent with the interpretation that their memories were not autobiographical, they included only 1.10 mentions of where events occurred, 0.16 temporal markers, and 0.35 subjective evaluations. However, when the same children recalled the same events at 9 years of age (6 years later)—to the extent that they remembered the events (more about this in a later section)—they included the features so saliently omitted from their early verbal accounts. At 9 years of age, they more than doubled the number of indicators of where events took place (to 2.29), showed a fourfold increase in the number of temporal markers included (to 0.72), and featured more than twice the number of subjective evaluations (to 0.78). Thus, later in development (with additional verbal competence), memories from the age of 3 years were expressed with strong autobiographical flavor. This suggests that the memory representations formed early in life included the mnemonic features that render memories autobiographical, even though evidence of them was missing from the relatively impoverished verbal reports produced earlier in development (see Tustin & Hayne, 2010, for a consistent discussion of the episodic nature of children's early memories).

A second, and complementary, source of evidence that memory representations may include mnemonic features that are not expressed verbally comes from adults' narratives, which not infrequently *lack* the features of autobiographical reports. For example, in Bauer and Larkina (2014), college students and middle-age adults included an average of only 5.32 and 5.88 of eight narrative features (specific features mentioned above), respectively. They frequently omitted information about where events took place and why they happened as they did. Strikingly, college students and middle-age adults provided a subjective perspective on the events only 46% and 53% of the time, respectively. The fact that the verbal reports of adults frequently omit information that clearly marks events as autobiographical makes it difficult to justify an argument that because children's reports lack these features, the memory representations that gave rise to them are not autobiographical. Indeed, were we take the argument to its logical conclusion, then we would "disqualify" many of the memory reports that adults provide as well. We do not make this claim because—when adults are the subjects—we recognize that verbal reports do not necessarily convey the full richness of memories.

A final source of concern with reliance on verbal reports as the index of when event memories become autobiographical stems from the fact that there are continuous changes in verbal and narrative behavior throughout childhood and into adolescence. This makes it challenging to identify a time in development when verbal reports have "enough" autobiographical features to consider them indicative of the memory type. The point here is that developmental changes in verbal behavior do not end at 5 to 7 years of age. Throughout the school years there are changes in the breadth of reports that children tell and in the coherence of their accounts. For example, between the ages of 7 and 11 years, there are increases in the length and complexity of children's autobiographical reports (e.g., Habermas, Negele, & Mayer, 2010). The amount of information that children include nearly doubles over this period (Van Abbema & Bauer, 2005), as does the temporal organization of the reports that children produce (Morris, Baker-Ward, & Bauer, 2010). Ten- to 12-year-old children also produce verbal reports that more effectively orient the listener to the time and place of the event, and they maintain and elaborate on topics more effectively than 7- to 9-year-old children (e.g., O'Kearney, Speyer, & Kenardy, 2007; Reese et al., 2011). However, even at age 11 to 12 years, children's reports still are lacking in the causal connections (e.g., *because*, *so that*) that characterize older adolescents' and adults' narrative accounts (e.g., Bauer, Stark, Lukowski, Rademacher, Van Abbema, & Ackil, 2005; Habermas et al., 2010). In adolescence, individuals use their autobiographical memories to construct an extended life story or personal history (e.g., Bohn & Berntsen, 2008; Fivush & Zaman, 2014; Habermas & Bluck, 2000; Thomsen, 2009; see Bohn & Berntsen, 2014). The point of summarizing these changes is to illustrate the seeming arbitrariness of selecting any single development in verbal narrative behavior as indicative of the "onset" of autobiographical memory. The evidence is more consistent with characterization of autobiographical memory as developing gradually and continuously.

Features of autobiographical memory are apparent early in development. The inadvisability of relying exclusively on verbal data as the source of evidence as to whether memories are autobiographical compels examination of other expressions of memory by children, with special emphasis on whether they remember specific past events located in place and time, and whether they show evidence of the personal relevance of the events. As will become apparent, they do; both behaviorally and verbally, the features become more and more prominent over the course of childhood. This pattern is part of the foundation upon which rests the suggestion that rather than as a classical concept with defining features, autobiographical memory is better characterized as a family resemblance concept with characteristic features (e.g., Bauer, 2007). Another component of the foundation is the observation that although any single expression of memory may be lacking in some features associated with autobiographical memory, there is no one element consistently missing from all expressions of memory. In other words, though there is evidence that some features are less frequently expressed, there is a lack of evidence that any given criterial feature is missing from early memory representations.

The ability to recall specific past events is readily apparent before the end of the second year of life—well before children provide verbal evidence of autobiographical memory. Some of the strongest evidence of the capacity comes from studies using non-

verbal imitation-based tasks in which props are used to produce novel actions or sequences of actions that infants are invited to imitate (e.g., Bauer & Mandler, 1989; Bauer & Shore, 1987; Meltzoff, 1985). As discussed in detail elsewhere (Bauer, 2007, 2013; Bauer, Wenner, Dropik, & Wewerka, 2000; Carver & Bauer, 1999; McDonough, Mandler, McKee, & Squire, 1995; Squire et al., 1993), the task is an accepted analogue to verbal report. Using this technique, researchers have found evidence of memory for unique events even in the first year of life. Infants as young as 6 months of age reliably imitate novel actions they observe produced by an experimenter (see Lukowski & Bauer, 2014, for a review). By 9 months of age, they remember unique actions and sequences of action over delays of at least 1 month (Carver & Bauer, 1999, 2001). By 20 months of age, the length of time over which recall is apparent has increased to 12 months (Bauer et al., 2000). In the same period, the robustness of memory increases such that infants remember more, based on fewer experiences of events (Bauer & Leventon, 2013; see Bauer, 2007, 2013, for reviews). In addition, recall over long delays is more reliably observed. Whereas at 9 months of age only roughly 50% of infants show evidence of long-term recall (e.g., Carver & Bauer, 1999), by 20 months, individual differences in whether or not infants recall are the exception rather than the rule (though there remain individual differences in how much is remembered; Bauer et al., 2000).

Because the actions and sequences on which infants are tested are novel to them, their behavior provides evidence that they remember unique events. Moreover, several other features associated with autobiographical memories also are apparent in nonverbal behavior and even in early verbal behavior; the features are evident well before children provide narrative evidence of autobiographical memory. For example, because infants recall both the individual target actions (*what-action*) and the temporal order in which they occurred (*when*; e.g., Bauer et al., 2000), there is evidence that they have some capacity for organization of event representations. Infants under 1 year of age demonstrate temporal organization for events that are logically (or causally) ordered (e.g., Carver & Bauer, 1999, 2001); by 20 to 24 months of age, they also reliably order events without this inherent structure (e.g., Bauer, Hertsgaard, Dropik, & Daly, 1998). One- to 2-year-olds also remember the specific locations in which events occurred (*where*), even over substantial delays (Lukowski, Lechuga, & Bauer, 2011). Infants under 2 years of age also demonstrate that they remember specific features of events, in that they reliably select the correct objects from arrays including objects that are different from, yet perceptually similar to, those used to produce event sequences (i.e., *what-object*, *how-description*; Bauer & Dow, 1994; Lechuga, Marcos-Ruiz, & Bauer, 2001; see also Wiebe & Bauer, 2005). At least by 20 months of age, memory for the specific props used to produce an event is related to memory for the event itself (Bauer & Lukowski, 2010). Infants under 2 years of age also evidence behavior that indicates that they have some understanding of *why* events unfold as they do (e.g., Bauer, 1992)—from their reproductions of event sequences they exclude actions that are irrelevant to the outcome. Finally, as they approach and enter the third year of life and gain the fluency to provide verbal descriptions of events experienced in imitation-based tasks, children spontaneously verbalize about who took part in the events (*who*) and they provide evaluative comments on the activities in which they engaged (*how-evaluation*; Bauer & Wewerka, 1997).

These behaviors make clear that sometimes well before they provide linguistic evidence, children encode, retain, and later retrieve memory representations that feature each of the individual elements associated with autobiographical memory (see Bauer, 2007; Bauer & Leventon, 2013, for discussions).

Over the preschool and early school years, memory processes improve to the point that children remember unique experiences, even over substantial delays. Children also provide more frequent and consistent evidence that they remember the events from their own personal, self-referential perspective. For example, Hamond and Fivush (1991) found that children who experienced a trip to Disneyworld when they were 36 or 48 months of age remembered the event even 18 months later. In Bauer and Larkina (2013), children 3 years of age at the time of events remembered in excess of 60% of them over delays of as many as 3 years. The preschool and early school years also are marked by developments in the ability to locate events in a particular time and place. Children become increasingly accurate and reliable in determining which of two events occurred earlier and in justifying their choices (Pathman, Larkina, Burch, & Bauer, 2013). They also show growing command of the use of conventional indices of time, such as calendars (e.g., Friedman, Reese, & Dai, 2011) and seasons (Bauer, Burch, Scholin, & Güler, 2007; Bauer & Larkina, 2014), to locate when personally relevant event occurred. Such markers serve as a timeline along which records of events can be ordered (see Friedman, 2014; Pathman & St. Jacques, 2014, for reviews). Children also become increasingly proficient at remembering the location in which they experienced specific past events (Bauer, Doydum, et al., 2012, Bauer et al., in press). These changes mean that more events are stored with more, better elaborated, and more tightly integrated elements of autobiographical memories: unique events, with distinctive features, accurately located in time and place.

For memories to be considered autobiographical, they also must be self-referential. As such, a self concept is a necessary ingredient for an autobiography (see Fivush & Zaman, 2014; Howe, 2014). Children first begin to make reference to themselves in past events at about the same time as they begin to recognize themselves in a mirror, namely, between 18 and 24 months (Howe & Courage, 1993, 1997, for discussions). Children who recognize themselves in the mirror have more robust event memories and over subsequent months, they make faster progress in independent autobiographical reports, relative to children who do not yet exhibit self recognition (Harley & Reese, 1999; see Reese, 2014). Throughout the preschool years, children develop a more self-oriented or subjective perspective on experience, as evidenced by increasingly frequent references to their own (and others') emotional and cognitive states (see Fivush & Zaman, 2014). References to the emotional and cognitive states of the experiencer indicate the sense of personal ownership and unique perspective that is characteristic of autobiographical memories.
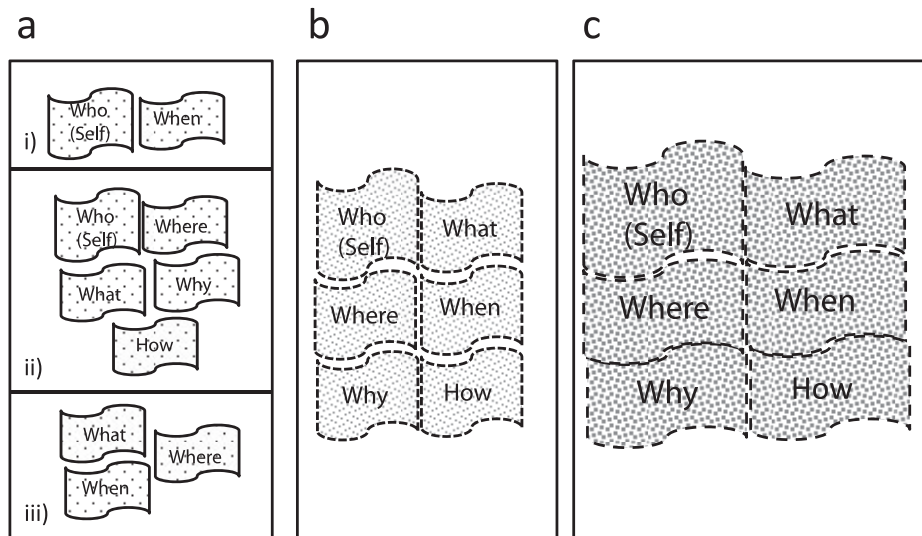
Memory traces are increasingly autobiographical with development. As this literature review makes clear, elements of autobiographical memory are apparent before the end of the second year of life. Indeed, review of the corpus of studies with infant participants reveals evidence of memory for all of the features or elements characteristic of autobiographical memory: *who, what-action, what-object, where, when, why, how-description,* and *how-evaluation*. In other words, even before the end of the second year

of life, memory behavior does not seem to be lacking any single defining or criterial element of autobiographical memory, even though, as represented in Figure 3, Panel a, no one memory may feature all of the elements (subpanels i, ii, and iii, represent separate memory traces, none of which features all of the elements characteristic of autobiographical memory). As infants become children, more and more representations feature all of the elements associated with autobiographical memory (Figure 3, Panel b); the features become better elaborated and more tightly integrated with one another (Figure 3, Panel c). As discussed elsewhere (Bauer, 2007, 2012, 2014), the net effect is that over the course of development, memory traces of past events become more and more prototypical of the category—they become more and more autobiographical. Expressions of memory that feature many of the elements we associate with autobiography (Figure 3, Panel c)— and that are highly prototypical of the category—will be readily recognized as members of the class (the "robins" of autobiographical memory). Expressions of memory that feature fewer of the elements (Figure 3, Panel a)—and that are less prototypical of the category—are less readily recognized as members of the class (the "ostriches" of autobiographical memory). The argument put forth here is that, just as ostriches are birds even though they cannot fly, many of the memories formed early in life have a sufficient number of autobiographical features to merit recognition of them as members of the autobiographical class. From this perspective, autobiographical memory does not emerge at 5 years of age (Nelson & Fivush, 2004; or even later, as suggested by, e.g., Fivush, 2012; Fivush & Zaman, 2014). Rather, over the course of development, memories become more and more typical exemplars of the category—they feature more and more of the attributes we associate with the category and the attributes become better elaborated and more tightly integrated with one another.

Summary. The developmental literature features extensive evidence of changes in memory for personally relevant past events over the course of early childhood. In traditional criterial accounts, children's early memories lack one or more defining features of autobiographical memory and thus are not considered to be autobiographical. In a prominent traditional account, for example (Nelson & Fivush, 2004), autobiographical memory emerges at age 5 years, coincident with development of a number of features considered defining of autobiographical memory. The complementary processes perspective developed in this review recognizes elements of autobiographical memory in the behavior of infants even before the end of the second year of life; the number and variety of elements increases over early childhood (and beyond) and the elements become better elaborated and more tightly integrated with one another. As a result, it becomes easier and easier to "see" autobiographical memory.

The Paradox of Childhood Amnesia

If memory just gets better and better and more and more autobiographical, why is it that so few memories from the first years of life survive (i.e., memory seems to get worse and worse)? I suggest that the solution to this paradox can be found by considering the complementary function of increases in the quality of memories, namely, the vulnerability of memories, especially those formed early in life (when the memories are of lower quality, as just discussed). In effect, events are remembered, but they also are

Memories become increasingly "auto**biographical**"

Figure 3. Depiction of course of development of autobiographical memory from the complementary processes perspective. The elements characteristic of autobiographical memory include who (self), what, where, when, why, and how (evaluative or subjective perspective). Early in development, individual memory representations (represented in Panel a, subpanels i, ii, and iii) may feature only a subset of the elements characteristic of autobiographical memories, yet no one element is missing from all memory traces. With development, memory traces feature more elements (Panel b). The elements also become more elaborated (depicted with increases in size and texture), and the elements become more tightly integration with one another (depicted with reduction in the spacing between elements, Panel c). As a result, over the course of development, memories of personally relevant specific past events take on more and more of the features of autobiographical memory.

forgotten (even by adults), and children forget at a faster rate, relative to adults. Essentially, youth is a risk factor for autobiographical memories. More important, events are not functionally forgotten because they are repressed (Freud, 1916/1966). Nor do they become inaccessible because of the onset of language (e.g., Neisser, 1962), or different senses of self associated with different life periods (e.g., Conway, 1996). Rather, early memories are forgotten because of normative processes involved in transformation of labile representations of experience into enduring memory traces. These suggestions are supported in the next section.

The Vulnerability of Memory Traces Declines Over Development

The suggestion that patterns of what we remember can be explained in part by patterns of what we forget is not new. There is a long tradition of work on forgetting functions in the adult literature (e.g., Ebbinghaus, 1885; Rubin & Wenzel, 1996; Wixted & Ebbesen, 1991, 1997) and in the developmental literature there have been examinations of the variance in long-term recall that can be explained by forgetting shortly after experience of an event (e.g., Bauer, 2005; Bauer, Van Abbema, & de Haan, 1999; Howe & O'Sullivan, 1997; discussed in more detail below). However, until the beginning of the 21st century, there were virtually no data that directly addressed the suggestion that the characteristic distribution of autobiographical memories across the life span—including the phenomenon of childhood amnesia—could be explained in

part by the relative vulnerability of memory traces formed early in life. To do so requires documenting memories created in the period eventually obscured by childhood amnesia and then prospectively tracking them across the boundary of the amnesia, to determine whether they are still remembered. However, throughout most of the century-plus of research on childhood amnesia, virtually all of the studies on the phenomenon were with adults; none of the studies was prospective. Early reports involving child subjects were retrospective—they asked what was retained from childhood, but not what was lost. As a result, though the possibility that memory trace vulnerability was part of the explanation of childhood amnesia had been implicated theoretically (Bauer, 2007, 2008; Olson & Newcombe, 2014; Peterson, Warren, & Short, 2011; Tustin & Hayne, 2010), there has been little opportunity to evaluate its role empirically.

One of the small number of prospective investigations of childhood amnesia in childhood was conducted by Cleveland and Reese (2008). To investigate the process of loss of (or loss of access to) memories for early life events, they recorded conversations between mothers and their children about past events at each of ages 19, 25, 32, 40, and 65 months. The fact that, during these interviews, the children provided unique information about the events—information that had not been given by their mothers— provided evidence that they had formed memories of them (a criterion of at least two unique pieces of information is typical in this literature). When the same children were 66 months of age,

they were tested for memory for events from each of the prior data collection points. Thus, at 66 months, children were asked for their memories of events from 1, 26, 34, 41, and 47 months in the past. The number of events that the 5.5-year-olds remembered decreased steadily as the retention interval increased. They recalled roughly 80% of events from only 1 month in the past but fewer than 40% of events from 47 months in the past. Fivush and Schwarzmueller (1998) reported a similar trend, albeit higher rates of retention, from 8-year-olds interviewed about events from ages 3.5 and 4 (77% of events remembered) versus 5 and 6 (92% of events remembered) years of age. Both studies provide evidence that as time goes by, forgetting (or loss of access) becomes more pronounced. However, in both studies, events with the longest delays between the initial and later tests also were events with the earliest age of encoding (i.e., 19 months in Cleveland & Reese, 2008). Thus, it is not possible to determine whether forgetting was a result of the length of the delay or the age of the children at the time of experience of the events. Furthermore, only Fivush and Schwarzmueller (1998) documented the fates of memories over the boundary of childhood amnesia (i.e., beyond age 7 years).

In Bauer and Larkina (2013) and Van Abbema and Bauer (2005), we held the age at encoding constant and varied the retention interval, thereby allowing for examination of fates of early memories over time. Specifically, we recorded conversations of dyads of 3-year-old children and their mothers as they discussed a number of events from the recent past. As for the studies just discussed (Cleveland & Reese, 2008; Fivush & Schwarzmueller, 1998), children's own unique contributions to the conversations made clear that they had formed memories of the events. Thus, we had documentation of memories from the age period corresponding to that from which adults report their earliest memories. We then tested different subgroups of the children again roughly 2, 3, 4, 5, or 6 years later, at the ages of 5, 6, 7, 8, or 9 years of age—ages at which, based on adult data, we would expect to see evidence of childhood amnesia. The later interviews were conducted by experimenters (rather than the children's mothers). The data are strongly suggestive of a role for forgetting in explanation of the onset of the amnesia. Whereas the children 5 to 7 years of age remembered more than 60% of the events from age 3 years, the 8- and 9-year-olds remembered fewer than 40% of the events. The difference in levels of recall was apparent even though all children were provided with prompts and cues to aid their memories, as determined to be necessary when free-recall failed. Moreover, the number of children who recalled none of the events from age 3 years also suggested a change in the accessibility of early memories after age 7 years. Whereas a maximum of 6% of children ages 5, 6, and 7 years recalled none of the events from age 3 years, 37% of 8-year-olds and 25% of 9-year-olds recalled none of the early life events. Again, the difference in levels of recall was observed even though all children received prompts and cues, as necessary.

There also is evidence that younger children forget more rapidly than older children. Morris et al. (2010) examined recall after a 1-year delay of events originally experienced at ages 4, 6, and 8 years. Children's contributions to the experimenter-conducted interviews made clear that they remembered the events. One year later, when the children were 5, 7, and 9 years of age, children's recall was tested again. As in the studies just discussed (Bauer & Larkina, 2013; Van Abbema & Bauer, 2005), children were given prompts and cues to aid their memories, as determined to be

necessary when free-recall failed. The children who had been the youngest at the time of the events remembered ~70% of them 1 year later. In contrast, the children who had been the oldest at the time of the events remembered 90% of them 1 year later. This pattern is strong evidence that within the period eventually obscured by childhood amnesia, the rate of forgetting is more accelerated among younger relative to older children.

Accelerated rate of forgetting. The facts that (a) even young children form memories of early life events but then seemingly forget them over time, and (b) younger children forget more rapidly than older children, demand an explanation of childhood amnesia that recognizes accelerated forgetting in childhood. Wetzler and Sweeney (1986) provided suggestive evidence of this phenomenon based on adult data. As noted earlier, they fitted a power function to data obtained by Rubin (1982) using cue word elicitation. The power function was a good fit to data from age 8 to adulthood. In contrast, it was a poor fit to data from birth to age 6 years, implying accelerated forgetting of memories from ages 6 and below.

In two studies, my colleagues and I have provided direct evidence of accelerated forgetting in childhood; the pace of forgetting is accelerated well beyond age 6 years. In Bauer et al. (2007), we used the cue word technique to examine the distribution of autobiographical memories in children 7 to 10 years of age. The children successfully generated memories in response to the cue words and accurately dated them, based on parental report. The distribution of memories produced by the children was better fit by the exponential than by the power function (see Table 1). The same pattern was obtained in an independent study by Bauer and Larkina (2014). We tested 20 children at each of the ages of 7, 8, 9, 10, and 11 years (100 children total), as well as two groups of adults: college students and middle-aged adults. As reflected in Table 1, the data from the children provided a replication of the results of Bauer et al. (2007). For the entire sample of children and for each group of children (7-, 8-, 9-, 10-, and 11-year-olds) separately, the best fitting function to the distribution was the exponential. In contrast, for both adult samples, the best fitting function was the power. The relative fits are illustrated in Figure 4. These data clearly suggest that as old as age 11 years, the distribution of children's autobiographical memories is not adult-like. In contrast to adults, children experience exponential forgetting.

Functional outcome of exponential forgetting. The exponential function implies a constant half-life. That is, over each unit of time (e.g., a month) the number of memories in the corpus decreases by one half. To use the earlier example, if Time 1 recall was of 100 memories, then recall at Times 2, 3, and 4 would be of 50, 25, and 12.5 memories, respectively. This pattern implies that the pool of memories available for recollection is ever-shrinking, suggesting that memories do not consolidate (see Bauer, 2012; Bauer et al., 2007; and Bauer & Larkina, 2014, for discussions). The contrast between a distribution of memories characterized by the exponential function relative to the power function is provided in Figure 5. The distributions differ both in terms of the initial rate of forgetting (also apparent in Figure 4), and in terms of the number of memories lost from the corpus with each unit of time. Consider that for both adults and children, many events are lost from memory virtually immediately after experience of them (see T2 in Figure 5). More important, for adults, the rate of forgetting slows over time, with individual memories becoming less vulner-

Table 1
Fit Indices for Power and Exponential Functions of the Distribution of Autobiographical Memories Elicited by Cue Words for Children 7 To 11 Years of Age and Adults

| Study | Age group | | Fit by function | |
	Overall	Individual age groups	Power	Exponential
Bauer et al. (2007)	Children 7 to 10 years		.95	.98
Bauer & Larkina (2014a)	Children 7 to 11 years		.82	.94
		7-year-olds	.94	.97
		8-year-olds	.87	.92
		9-year-olds	.84	.89
		10-year-olds	.82	.88
		11-year-olds	.72	.86
	Adults		.91	.65
		College students	.84	.61
		Middle-age adults	.93	.70

Note. For each age group, the best fit function is highlighted by a box.

able to disruption and interference, resulting in a relatively stable corpus (e.g., Wixted, 2004, for discussion). In contrast, children experience a sharp initial decline in the number of memories in the corpus (T2) and unlike for adults, for them, the rate of forgetting does not slow down—it is exponential.

The apparent fact of the exponential form of forgetting in childhood has two important implications for the fates of memories formed in the first decade of life. First, over a given unit of time, children lose more memories than adults do. Second, because the rate of forgetting is constant, the memories that survive the initial ravages of time may nevertheless eventually succumb to forgetting. Over time, the corpus or pool of memories of early life events shrinks. Moreover, because the rate of forgetting does not slow down, the pool of memories is ever-shrinking, contributing to the appearance of a "childhood amnesia component" (Pillemer & White, 1989)—a smaller number of memories than expected by normal forgetting (i.e., with "normal" forgetting equated with an adult rate, characterized by the power function). We may think of the process as one that reduces "pools" of memories to isolated "puddles" of memories, resulting in a sparse representation. More isolated representational structures are more difficult to retrieve.

Understanding children's "earliest memory" data. Thinking in terms of "pools to puddles" of memories over the course of the first decade of life aids in understanding of the patterns of recall of early autobiographical memories by adults that is associated with childhood amnesia. It also aids in understanding of the emerging body of data on "earliest memories" of children. As noted above, studies of children's recall of their earliest memories are a relatively new addition to the literature and the number of studies is small. They reveal that by the end of the first decade of life, children's earliest memories show the same distribution as adults. Specifically, queries about the age of earliest memory among children as young as 6 years of age and as old as 19 years have produced estimates of the average age of earliest memory at 38 months, with a range of 28 months to 45 months (Jack, MacDonald, Reese, & Hayne, 2009; Larkina, Merrill, Fivush, & Bauer, 2009; Peterson, Grant, & Boland, 2005; Reese, Jack, & White, 2010). The estimates fit comfortably around the 3- to 4-year (36 to 48 month) range obtained from adults.

In contrast to the adult-like distribution of memories among older children, younger children's recall of their earliest memories differs from the patterns seen in adulthood in at least two ways: age-cohort effects and instability in the memory identified as "earliest." First, age-cohort effects are apparent in children but not in adults. Two studies have revealed age-cohort effects within childhood, such that the age of earliest memory is earlier for younger children relative to older children. For example, in Tustin and Hayne (2010) the average age of earliest memory among 5-year-old children was 1.7 years, whereas the average age of earliest memory among 12- to 13-year-old children was 2.5 years. Similarly, in Peterson et al. (2005), the average age of earliest memory among 6- to 9-year-old children was 3 years, whereas the average age of earliest memory among 10-year-olds was 3.5 years. In contrast, as noted earlier, among adults, there are not age-cohort effects. Whether tested at 20 years of age or 70 years of age, the average age of earliest memory among adults is 3 to 4 years (Rubin & Schulkind, 1997).

There also is evidence of less stability in the "earliest memory" among children relative to adults, both in terms of the memory identified as the earliest and in terms of the age of earliest memory. Peterson et al. (2011) interviewed children 4 to 13 years of age about their earliest memories. Two years later, they asked the same children to once again report their earliest memories. Strikingly, among children 4 to 7 years of age at the first interview, there was little overlap in the memories nominated at the two time points: only 7% of 4- to 5-year-olds and 13% of 6- to 7-year-olds nominated the same earliest memory, whereas 12- to 13-year-olds were consistent 39% of the time. When the events were the same, the children were inconsistent in their estimates of their ages at the time of the events. Between queries, the estimated age of earliest memories increased from 32 to 39.6 months. In contrast, over a 3-year period, 82% of adult women identified the same memories as their earliest and they varied by only 0.3 months in dating the event that gave rise to the memory (38.9 vs. 38.6 months; Bauer et al., in press). Thus, there is substantially less consistency in the corpus of earliest memories among children relative to adults. The pattern is to be expected of a pool of memories that is ever-shrinking and ever-changing.

Summary. Throughout most of the history of research on childhood amnesia, the sole participants were adults. Without exception, the studies were retrospective. As a result, the field lacked the most relevant data on the source of the phenomenon, namely, documentation of memories formed in the period eventually obscured by the amnesia, and tracking of the memories across the boundary. With the

a

b
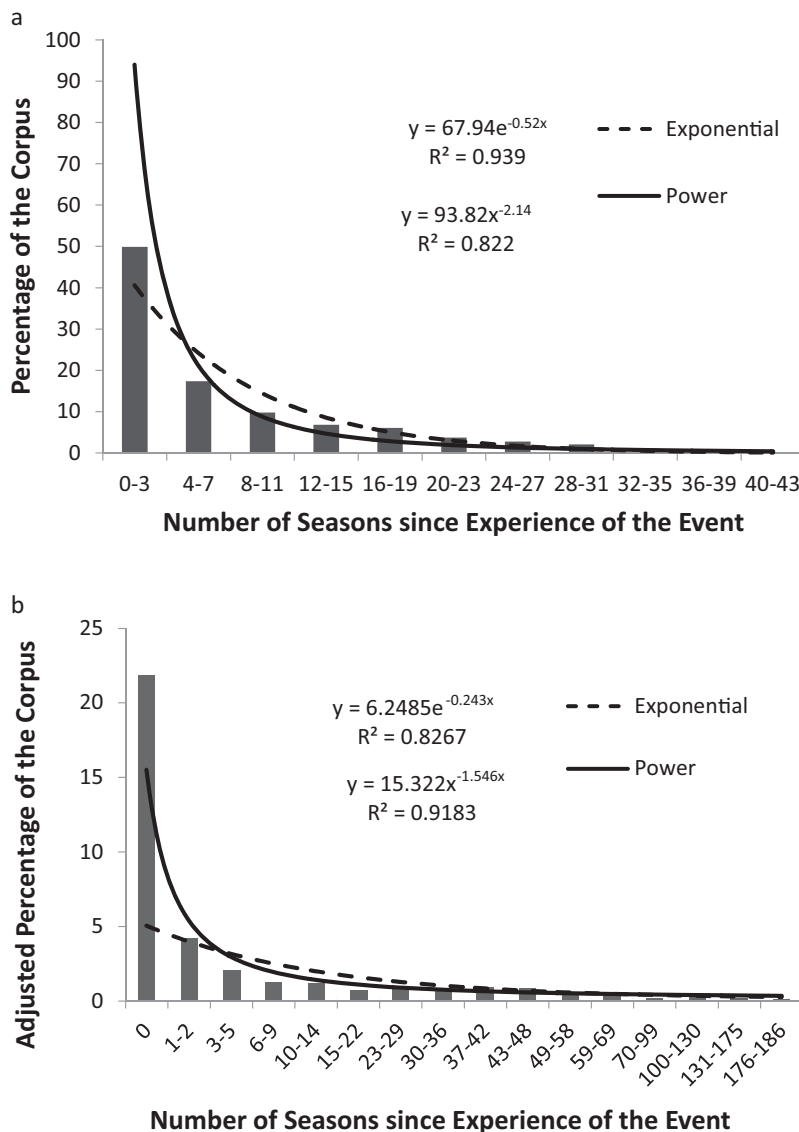
Figure 4. The power function and exponential function fitted to the distribution of memories generated by children 7 to 11 years of age as a function of the number of seasons since experience of the event (Panel a, based on 1,995 memories). The power function and exponential function fitted to the distribution of memories generated by college-age and middle-age adults (Panel b, based on 800 memories). For adults, the figure represents adjusted data such that the horizontal axis is the time (the mean number of seasons to the middle of the bin) since experience of the event, and the vertical axis is the percentage of data averaged across the seasons included in the bin (see Bauer & Larkina, 2014, for discussion). Adapted from "Childhood Amnesia in the Making: Different Distributions of Autobiographical Memories in Children and Adults," by P. J. Bauer and M. Larkina, 2014, *Journal of Experimental Psychology: General, 143,* p. 605. Copyright 2014 by the American Psychological Association.

advent of such data, it has become clear that childhood amnesia emerges by middle childhood. Specifically, by the time children are 8 to 9 years of age, they have forgotten a substantial proportion of early childhood events they once remembered (Bauer & Larkina, 2013; Van Abbema & Bauer, 2005). The available data strongly suggest that the amnesia emerges as a result of exponential forgetting in childhood, relative to adulthood. A consequence of exponential forgetting by children is that the pool of autobiographical memories they have

formed eventually diminishes to isolated puddles of memories; isolation makes the remaining memories even more difficult to retrieve. Findings of exponential forgetting in childhood explain the emergence of childhood amnesia in childhood. They also bring order to an array of findings on children's earliest memories, and help to explain why in some cases they are different, relative to adults. In the next section, I take up the question that now demands to be addressed, namely, why childhood is a period of accelerated forgetting, in the
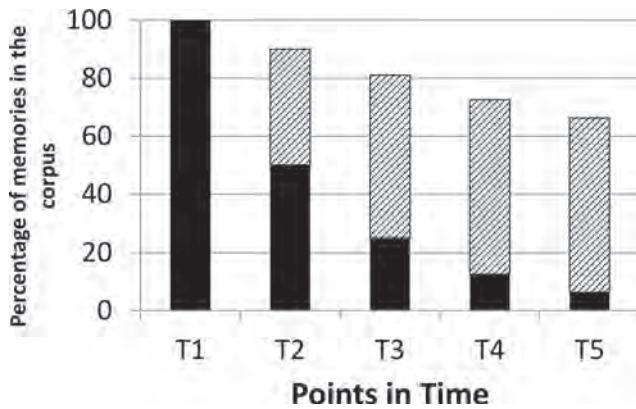
Figure 5. Schematic depiction of the number of memories in the corpus that survive over each hypothetical unit of time (T1-T5) in distributions characterized by the exponential function (dark bars) and the power function (hashed bars).

face of salient increases in the quality of memory representations that are formed.

The Dynamics of Increases in the Quality and Decreases in the Vulnerability of Memory Traces and Mechanisms of Developmental Change

The theoretical argument I have advanced recognizes the complementary processes involved in the creation, retention, and later retrieval of memories (glossed as increases in the quality of memory traces) and in the loss of representational traces from memory (glossed as vulnerabilities in memory traces). In this section, I make explicit the interaction of these processes. Specifically, I highlight improvements in the quality of the representations that are formed. Over developmental time, memory representations include more of the features that characterize autobiographical memories, and the features are better elaborated and more tightly integrated. The result is mnemonic materials that are of higher quality. At the same time, there are developments in the neural substrate operating on the available representations, both in terms of the structures involved and in the level of connectivity of the network of structures. The developments herald more efficient and effective cognitive and mnemonic processing, resulting in decreases in the vulnerability of memory traces to forgetting. Put another way, over developmental time, less and less negatively impacts the mnemonic representations that are formed. The net effect is that with development, more and more memories of higher and higher quality survive to be recalled at later points in time, producing the characteristic distribution of autobiographical memories across the life span. I first summarize the processes involved in the formation of memory representations, before discussing the developmental interactions.

Processes Involved in the Formation of Memory Representations

The layperson's view of memory is of a file cabinet stuffed full of file folders, the contents of each of which is a memory representation. To recall, one finds the right folder, pulls it out of the cabinet, opens it, and reads off what happened to whom and when. In actuality, of course, memory is nothing like a file cabinet or folder. Rather, memory representations are made up of individual bits and pieces of experience that are encoded in synaptic connections between individual neurons distributed across the cortex. They begin their lives as patterns of neural activity that give rise to conscious experience of events. Their continued existence as "memories" depends on subsequent processing carried out by a multicomponent neural network that includes structures in the medial-temporal lobe, as well as the neocortex. Their subsequent retrieval entails recreation of the pattern of neural activity that gave rise to the event in the first place. Each of these steps is elaborated below (see Eichenbaum & Cohen, 2001; Kandel & Squire, 2000; Manns & Eichenbaum, 2006; Nadel, Samsonovich, Ryan, & Moscovitch, 2000; Rubin, 2005, 2006; Winocur & Moscovitch, 2011; Zola & Squire, 2000; for reviews of these processes).

Memory begins with encoding of experience into a memory trace. Encoding, in turn, begins with the initial registration of information in the brain. The whole of experience does not impinge upon the brain at once in the same time and place, but is distributed across multiple cortical areas. For example, cell fields in primary somatosensory cortex register object or event-related tactile information from the skin and proprioceptive inputs from the muscles and joints. Simultaneously, fields in primary visual cortex register the form, color, and motion of the object or event, and fields in primary auditory cortex respond to the various attributes of the sounds associated with the object or event. Inputs from these primary sensory cortices are sent (projected) to unimodal sensory association areas where they are integrated into whole percepts of what the object or event feels like, looks like, and sounds like, respectively. Unimodal association areas in turn project the information to polymodal (also termed multimodal) posterior-parietal, anterior-prefrontal, and limbic-temporal association areas. The coordinated activity of these cortical areas gives rise to experience of a coherent event.

For the experience of an event to endure as a memory, the pattern of neural activity giving rise to the experience must undergo a process of stabilization into a memory trace and integration of the trace into long-term storage. This process—known as *consolidation*—depends on neurochemical and neuroanatomical changes that create a physical record of the experience (McGaugh, 2000). It results from the coordinated actions of structures in the medial-temporal lobes and cortical association areas. Specifically, inputs from the association areas are projected to structures in the medial-temporal lobes, with inputs that specify the nonspatial or "object" features of experience projected to perirhinal cortex, and inputs that specify the spatial or "contextual" features of experience projected to parahippocampal cortex (see, e.g., Manns & Eichenbaum, 2006, for a review). These cortices are thought to hold the still-segregated streams of experience (i.e., nonspatial and spatial) in an "intermediate-term memory" on which the hippocampus proper operates. The information is held in the medial-temporal cortices temporarily (over periods of at least several minutes), presumably as a result of prolonged neuronal firing (e.g., Suzuki, Miller, & Desimone, 1997).

To be stabilized into a coherent memory trace, information must make its way into the hippocampus proper. It does so via the connecting link of the entorhinal cortex. Specifically, the perirhi-

nal and parahippocampal cortices project their highly processed sensory inputs to the lateral and medial aspects of the entorhinal cortex, respectively. The entorhinal cortex projects these inputs into the hippocampus proper where all of the different components of the event are bound into a single representation (see Manns & Eichenbaum, 2006, for additional discussion). This "binding" of the elements of experience depends upon iterative processing of the conjunctions and relations among the stimuli that gave rise to the event. That is, the pattern is regularly "refreshed" by additional neural signaling among the hippocampus, the surrounding medial-temporal cortices, and the association areas. It also maintains and strengthens the linkages between the distributed cortical representations that make up the entire event. As it does so, it strengthens the intracortical connections between the different elements of the event representation. By some accounts, representations may strengthen to the point that, eventually, they no longer require the activity of the hippocampus for their maintenance (Alvarez & Squire, 1994; McClelland, McNaughton, & O'Reilly, 1995; Reed & Squire, 1998; Squire, 1992; Zola & Squire, 2000). By other accounts, memory traces remain dependent on the hippocampus, especially for retrieval (e.g., Moscovitch & Nadel, 1998; Nadel et al., 2000; Winocur & Moscovitch, 2011).

Iterative processing of the conjunctions and relations among event-related stimuli in the medial-temporal structures not only serves to stabilize new memory representations—it also supports the integration of the new information with that previously stored (e.g., McKenzie & Eichenbaum, 2011). The basis for integration is overlapping or shared elements. To the extent that a new event memory shares elements with memories already in storage, the representations will be simultaneously activated. Neurons that are repeatedly activated together (synchronous convergence) tend to become associated. The result is an entire pattern of interconnection of new information with old.

Finally, the *raison d'etre* for the consolidation and storage of memories is so that they can be retrieved at some later time. Retrieval is, in essence, a reactivation of the neural network that represents the event. Reactivation occurs because "An internal or external stimulus, whose cortical representation is part of the network by prior association, will reactivate that representation and, again by association, the rest of the network" (Fuster, 1997, p. 455). Specifically, retrieval of information from long-term stores is accomplished by the same circuits as were involved in initial registration of the experience. In the case of autobiographical memory, neuroimaging studies have revealed that retrieval is carried out by a distributed network involving the hippocampus and surrounding cortices, amygdala (for emotional events), retrosplenial cortex, posterior parietal regions (including precuneus), visual cortex, and lateral and medial prefrontal cortex (Addis, McIntosh, Moscovitch, Crawley, & McAndrews, 2004; Cabeza et al., 2004; Greenberg, Rice, Cooper, Cabeza, Rubin, & LaBar, 2005; see Gilboa, 2004; Rubin, 2005, 2006; Shimamura, 2011, for reviews; see Maguire, 2001; Svoboda, McKinnon, & Levine, 2006, for meta-analyses). During the initial phases of memory search, the more anterior, frontal-temporal components of this network are especially active, whereas in the later phases of retrieval—as the trace is elaborated—the more posterior, occipital, and parietal regions are especially active (e.g., Daselaar, Rice, Greenberg, Cabeza, LaBar, & Rubin, 2008; McCormick, St-

Laurent, Ty, Valiante, & McAndrews, 2013; St. Jacques, Kragel, & Rubin, 2011;).

Developmental Interactions Involved in Increases in the Quality and Decreases in the Vulnerability of Memory Traces

Recognition that memory traces are distributed representations of loosely affiliated elements awaiting consolidating "glue" to hold them together, aids in understanding of the importance of considering both the number and the richness of the elements of the experience that are available for inclusion in the memory trace and the efficiency and efficacy of the neural processes implicated in memory formation—as well as their interaction—to explanation of childhood amnesia. To the extent that the distributed representations become more complete, elaborated, and more intimately tied to related constructs and representations, the larger and more numerous the surfaces upon which to apply the consolidating glue. To the extent that the glue that works to hold the representations together become more efficient and effective, fewer of the elements of experience will escape the bonds and be lost from the memory trace. I discuss each element of the equation, beginning with the glue.

Development of the glue: The neural substrate of memory. The glue is a metaphor for the stabilizing "efforts" of the medial-temporal and cortical structures that work to transform transient, labile representations of experience into enduring traces. The efficiency and efficacy of those processes is intimately tied to the developmental status of the neural substrate responsible for them. As explained in detail elsewhere (see, e.g., Bachevalier, 2014; Bauer, 2007, 2009a, 2013; Ghetti & Bunge, 2012; Ghetti & Lee, 2014; Nelson, de Haan, & Thomas, 2006, for reviews), portions of the neural network implicated in episodic and thus autobiographical memory develop early, whereas many of the structures (or aspects thereof), as well as the connections between and among the structures, undergo a protracted developmental course that continues well into adolescence. This leads to expectations of changes in the way in which the network functions throughout infancy and childhood (and even beyond). Early changes support the emergence of the capacity to form, retain, and later retrieve self-relevant memories of past events; later changes herald more efficient and effective function and a concomitant reduction in the rate of forgetting.

Early changes. As outlined above, over the first months of life there are salient changes in the robustness, reliability, and temporal extend of infants' memories. The changes likely are related to postnatal changes in brain that take place over the course of infancy (as well as other influences, such as the social environment in which early development takes place; e.g., Fivush, 2014; Reese, 2014). Whereas much of brain development occurs prenatally, there also are pronounced changes in the first years of postnatal life.

With regard to the structures implicated in memory for specific past events, as summarized by Seress and Ábrahám (2008), in the medial-temporal lobes, hippocampal cells are generated during the first half of prenatal development and have migrated to their final destinations by birth. Synapses are apparent by about 15 weeks gestation. The number of hippocampal synapses and synaptic density increases until about 6 months of age, at which time adult

levels are reached. At this same time, glucose utilization (an indicator of energy use) also reaches adult levels, likely in relation to the increased number of synapses (Chugani, 1994; Chugani & Phelps, 1986). By the end of the first year of life, the volume of the hippocampus has doubled (Gilmore et al., 2012).

Within the hippocampus, the development of the specific region of the dentate gyrus is more protracted (Seress & Ábrahám, 2008). This area of the brain includes about 70% of the adult complement of cells at birth; the remaining cells are produced postnatally (neurogenesis in this region has been confirmed in childhood and beyond; Tanapat, Hastings, & Gould, 2001). Morphologically the structure is adult-like around 12 to 15 months after birth. Increases in synaptic density also are somewhat protracted relative to what is observed in other regions of the hippocampus; synaptic density in this region increases starting around 8 to 12 months after birth and peaks around 18 to 20 months (Eckenhoff & Rakic, 1991). Prefrontal cortex undergoes its early development at a similar pace. Synaptic density in this region increases beginning around 8 months after birth and reaches its peak between 15 and 24 months (Huttenlocher, 1979; Huttenlocher & Dabholkar, 1997). Synapses appear adult-like in their morphology by 24 months (Huttenlocher, 1979).

Achievement of the peak number of synapses in both the hippocampus and prefrontal cortex has important implications for behavior. As argued by Goldman-Rakic (1987), with this development, we should expect to see the emergence of what she termed the "signatory" (or characteristic) functions of the neural substrate; attainment of mature levels of function would coincide with the period of synapse elimination in the network. Though Goldman-Rakic made this argument with respect to cortical regions, extension of the suggestion to the entire temporal-cortical network leads to the prediction of emergence of the signatory function of long-term memory for specific past events by the end of the second year of life, with continued development for years thereafter. Specifically, the cortical components of the network, and the connections both within the medial-temporal lobe (i.e., those involving the dentate gyrus of the hippocampus) and between the cortex and the medial-temporal components, would be expected to reach functional maturity over the course of the second year of life, coincident with estimates of achievement of the peak of synaptogenesis by 20 months in the dentate gyrus, and by 24 months in the prefrontal cortex. The expectation of developmental changes for months and years thereafter stems from the schedule of protracted pruning both in the dentate gyrus (until 4 to 5 years; e.g., Eckenhoff & Rakic, 1991) and in the prefrontal cortex (throughout adolescence; e.g., Huttenlocher & Dabholkar, 1997). These estimates are consistent with the behavioral data summarized above, which indicate the ability to recall multiple episodic features (e.g., *what*, *where*, and *when*) at least by the end of the second year of life.

Later changes. Postnatal changes in the neural substrate that supports memory for specific past events and thus autobiographical memory continue well beyond the first 2 years of life; the changes have implications for memory behavior. At the level of the brain as a whole, children's brains have reached roughly 90% of adult volume by the time they are 5 years of age (Kennedy, Makris, Herbert, Takahashi, & Caviness, 2002). There are further volume increases through puberty (Caviness, Kennedy, Richelme, Rademacher, & Filipek, 1996). Beyond puberty, there is actually a reduction in gray matter volume, likely associated with synaptic pruning and other regressive processes (Gogtay et al., 2004; Huttenlocher, 1990; Jernigan, Trauner, Hesselink, & Tallal, 1991). In contrast, white matter volume increases linearly with age (Giedd et al., 1999). The increases are associated with greater connectivity between brain regions and with myelination processes that continue into young adulthood (e.g., Johnson, 1997; Klingberg, Vaidya, Gabrielli, Moseley, & Hedehus, 1999; Schneider, Il'yasov, Hennig, & Martin, 2004).

Different cortical structures undergo changes in gray matter volume at different rates. Specifically, longitudinal data (e.g., Østby, Tamnes, Fjell, & Walhovd, 2011) indicate that the nonlinear changes in cortical gray matter occur earlier in the frontal and occipital poles, relative to the rest of the cortex, which matures in a parietal-to-frontal direction. The superior temporal cortex is last to mature (though the temporal poles mature early; Gogtay et al., 2004). Prefrontal cortex undergoes an especially protracted development, with adult levels of synapses not reached until late adolescence or even early adulthood (Huttenlocher, 1979; Huttenlocher & Dabholkar, 1997), and myelination processes continuing well into adolescence or young adulthood (e.g., Johnson, 1997; Klingberg et al., 1999; Schneider et al., 2004). It is not until adolescence that neurotransmitters such as acetylcholine reach adult levels (discussed in Benes, 2001).

The volume of the hippocampus also undergoes changes, with gradual increases into adolescence (e.g., Gogtay et al., 2004; Østby, Tamnes, Fjell, Westlye, Due-Tønnessen, & Walhovd, 2009; Pfluger et al., 1999; Utsunomiya, Takano, Okazaki, & Mistudome, 1999). In addition, myelination in the hippocampal region continues throughout childhood and adolescence (Arnold & Trojanowski, 1996; Benes, Turtle, Khan, & Farol, 1994; Schneider et al., 2004). It also is subregion specific, with some subregions achieving mature patterns of myelination in infancy (i.e., in the fimbria), childhood (i.e., CA1 and CA3 subfields), and after puberty (i.e., hilus of the dentate gyrus; Ábrahám et al., 2010). There also are marked developmental changes in connectivity between the hippocampus and other neural regions (see, e.g., Schahmann & Pandya, 2006, for a review).

Implications for behavior. Developmental changes in the structures that subserve memory as well as in the connections among them can be expected to have implications for memory behavior (see Bauer, 2007, 2009a, 2009b, 2013, for other reviews). Late development of aspects of the hippocampal formation may be especially critical because of their role in stabilization and integration of memory traces into long-term storage. The immaturity of these structures and connections between them would present challenges to these processes. The challenges may be most pronounced until 4 to 5 years of age, when the schedule of protracted pruning of synapses in the dentate gyrus of the hippocampus has largely been reached (e.g., Eckenhoff & Rakic, 1991).

The suggestion that the processes involved in stabilization and integration of memory traces into long-term storage are a source of variance in recall in infancy and early childhood is supported by behavioral and electrophysiological data. Specifically, using both behavioral measures (e.g., Bauer, 2005; Bauer, Güler, Starr, & Pathman, 2011; Bauer et al., 1999) and measures of recognition obtained from event-related potentials (ERPs; Bauer et al., 2006; Bauer, Wiebe, Carver, Waters, & Nelson, 2003), my colleagues and I have probed the integrity of memory traces at various points

post encoding, ranging from minutes after experience of an event to 2 weeks later. We then examined the proportion of variance in long-term recall explained by the measures obtained shortly after encoding, as memory traces presumably are undergoing consolidation. In these cases, the measure is not of forgetting, per se, but of what is retained. However, because what is retained hours to days after experience of an event is inversely related to the amount forgotten, the logic holds. We have found that estimates of memory obtained hours to days after experience of events are predictive of infants' recall after subsequent delays of weeks to a month. The measures account for significant unique variance above and beyond that explained by measures of encoding, and in some cases, render encoding-related variance nonsignificant (Pathman & Bauer, 2013). Similar effects are observed for children as old as 3 and 4 years of age (Bauer, Larkina, & Doydum, 2012). Indeed, as summarized in Howe and O'Sullivan (1997), throughout childhood, memory failure at the level of consolidation and/or storage accounts for the largest proportion of age-related variance in children's recall of laboratory stimuli.

The data discussed earlier on the distribution of autobiographical memories across the first decade of life (Bauer et al., 2007; Bauer & Larkina, 2014) indicate that failure at the level of consolidation and/or storage extends beyond laboratory stimuli to personally relevant events. They suggest that at least until the age of 11 years, memories may be lost because they fail to consolidate. Failed consolidation in turn provides an explanation for the accelerated rate of forgetting of autobiographical experiences in childhood (Bauer et al., 2007; Bauer & Larkina, 2014). Failed consolidation and exponential forgetting may stem not only from the relative immaturity of the neural network involved, but also may be linked with hippocampal neurogenesis, which is argued to play a role in forgetting (e.g., Frankland, Köhler, & Josselyn, 2013). Rates of neurogenesis decline with age (e.g., Kuhn, Dickinson-Anson, & Gage, 1996), raising the possibility that higher rates in the first years of life may contribute to exponential forgetting in childhood. Consistent with this suggestion, in infancy, when rates of neurogenesis are high, decreasing neurogenesis (either genetically or pharmacologically) after new memory formation mitigates forgetting in the mouse model (Akers et al., 2014). We also may speculate that consolidation processes would be impacted by sleep parameters that change over the course of development (e.g., Ohayon, Carskadon, Guilleminault, & Vitiello, 2004). Changes in slow-wave sleep, in particular, may have implications for memory consolidation and the form of forgetting (e.g., Rasch, Büchel, Gais, & Born, 2007). Conversely, data from Østby et al. (2011) and DeMaster, Pathman, Lee, and Ghetti (2012) suggest that increases in the reliability of consolidation processes may be related to changes in hippocampal volume (see also Ghetti & Lee, 2014 and Sowell, Delis, Stiles, & Jernigan, 2001, for discussion). Region-specific activation in the hippocampus—specifically, in the anterior region—has been linked with better memory for the types of associations among stimuli that are part-and-parcel of episodic memory (Paz-Alonso, Ghetti, Donohue, Goodman, & Bunge, 2008; Ghetti, DeMaster, Yonelinas, & Bunge, 2010).

Cortical structures also are implicated in the consolidation processes that effectively stabilize memory traces and accomplish their integration into long-term storage, as well as in their initial encoding and later retrieval. As such, protracted development of cortical structures can be expected to be related to memory behav-

ior. Consistent with this suggestion, Østby and colleagues (2011) found correlations between the thickness of left orbitofrontal cortex and encoding success in children 8 to 19 years of age. Studies of encoding-related activation indicate age-related increases in recruitment of prefrontal cortical regions across childhood. For instance, Wendelken, Baym, Gazzaley, and Bunge (2011) found that among children 8 to 13 years of age, older children had higher levels of recruitment of the dorsolateral prefrontal cortex during encoding of faces and scenes, which in turn modulated subsequent memory (see also Ofen, Kao, Sokol-Hessner, Kim, Whitfield-Gabrieli, & Gabrieli, 2007). Finally, though the necessary studies have not been done, it is logical to assume that developmental changes in medial prefrontal cortex, posterior parietal cortex, and precuneus would impact the efficiency with which autobiographical memories are accessed and elaborated (e.g., Cabeza et al., 2004; Daselaar et al., 2008; McCormick et al., 2013).

Summary. The efficiency and efficacy of the processes by which transient, labile representations of experience are transformed into enduring memory traces is closely tied to the developmental status of the neural substrate responsible for them. In the first several months of life, the substrate develops rapidly such that as infants approach the end of second year, the neural structures and network connections seemingly have reached a level of maturity sufficient to support their signatory function. Though the function is apparent, it is far from fully mature. Over the subsequent period of childhood and early adolescence, the structures and network continue to develop, heralding increases in the efficiency and thus the efficacy with which it glues or binds together the elements of experience into enduring traces. The rate of progress is relatively slow, such that at least for the first full decade of life, effective loss from memory is faster than the rate in adolescence and adulthood.

Development of the "surface": The raw materials of memory. Of course, the power of the consolidating glue is only as good as the raw materials to which it is applied. In terms of memory representations, simply put, if there is less available to start with, less will survive the ensuing encoding and consolidation processes, and less will be available for retrieval. As discussed above, the elements associated with autobiographical memory are apparent in memory behavior even early in childhood. That is, there does not seem to be any one criterial feature "missing" from young children's representations of past events that would preclude them from forming autobiographical memories. However, young children's memories nevertheless seem rather atypical of the category—they are more ostrich-like than robin-like. They have this quality in part because any given memory trace may lack information that is prototypical of the category of autobiographical memories, such as specific location in place and time, for example (Bauer, Doydum, et al. 2012; Pathman et al., 2013, respectively). The net effect is that within memory traces, there are fewer features to be associated and between and among which conjunctions are formed. There also are fewer features to differentiate one event or episode from another. Seemingly paradoxically, there also are fewer opportunities for integration of new representations with memory traces already in long-term storage. As memory representations become more and more populated by temporal and spatial features, for example, the number of distinct events and experiences represented in memory increases. As the memory traces become inte-

grated with one another, they enable formation of a timeline of past events.

Early in development, opportunities for integration of new memory traces with those already in long-term storage also are lessened by the relative immaturity—and lack of elaboration—of the concepts upon which autobiographical memory depends. For example, the self to which early formed memories are referenced is not as elaborated or stable a construct as it will be later in development (Fivush & Zaman, 2014; Howe, 2014). As such, it provides a less coherent reference point for memories about the self. As well, children's relatively immature understanding of the representational nature of the human mind may make it more challenging for them to adopt a unique or subjective perspective on events. As these concepts and understandings are elaborated, they become ever-more integral elements of the memory representations that are formed of the distinct events and experiences that make up the life timeline of events. Greater elaboration of the concepts upon which autobiographical memory depends may even operate to facilitate consolidation and reduce the vulnerability of new memories to forgetting. In a mouse model, Tse and colleagues (Tse et al., 2007) found that the rate of systems-level consolidation of new learning was accelerated in mice that had previously acquired a relevant "schema" to which the new information could be assimilated. These findings indicate that the contents of long-term memory may have functional consequences for the incorporation of new memories. Together, these suggestions imply that as the concepts upon which autobiographical memory depends are more fully elaborated and stabilized, they become more integral to mnemonic traces of personally relevant events and simultaneously may work to accelerate the incorporation of the new traces into long-term stores.

Finally, children's less well developed verbal and narrative skills also may negatively impact the quality of the raw materials for memory. As discussed above, younger children's verbal descriptions of events feature fewer of the elements that make for a good personal story, relative to older children. Their narratives also are less well organized and coherent. Again, as discussed above, this does not necessarily imply that their memory representations suffer these same deficiencies. However, it does mean that younger children miss out on the opportunities afforded by verbal retelling of an event. Incomplete narratives also tend to be less well organized (Reese et al., 2011), and retelling of them affords less rehearsal benefit (Bauer & Larkina, 2013). Incomplete narrative rehearsal may even result in retrieval-induced forgetting, the phenomenon by which retrieving some elements of a memory trace may result in forgetting of nonretrieved elements (e.g., Anderson, Bjork, & Bjork, 1994), with resultant further trace degradation. Thus, developments in verbal and narrative skill herald cognitive benefits of rehearsal and organization, which are advantageous to memory. Beyond individual memories, developments in narrative skills also make possible construction of a life story or autobiography that weaves together the individual experiences of one's life into a sequence of temporally linked events (e.g., Brewer, 1980; Fivush et al., 2011; McAdams, 2001). Narrative evidence of this development becomes apparent only in adolescence (e.g., Bohn & Berntsen, 2008, 2014; Fivush & Zaman, 2014; Habermas & Bluck, 2000).

Summary and implications. Memory representations are the products of the activity of large populations of individual neurons that are widely distributed across large tracts of neural tissue. Both the number of elements of experience that are featured in the representation and the extent to which the elements are successfully integrated with one another and with existing memory representations influence the integrity of the traces and their longevity and later accessibility. Early in development, any given memory representation is less populated by the features that typify autobiographical memories. The features that are represented are less fully elaborated. Moreover, the representations that already exist in memory and are available for integration with newly formed traces are themselves less well populated and elaborated. The result is that the raw materials for formation of new memories are suboptimal.

Adding insult to injury is the fact that the less-than-optimal raw mnemonic materials are operated upon by less-than-optimal neural, cognitive, and mnemonic processes. Many of the structures as well as the network of structures that supports memory undergo a protracted course of development lasting well into adolescence. For much of the first 2 years of life, encoding, consolidation, and subsequent retrieval of long-term memories of specific past events are fragile processes owing to immaturity of the neural substrate that supports these signatory functions. For years thereafter, the substrate operates relatively inefficiently and ineffectively. The result is that elements of experience that are the raw materials for memory are not effectively stabilized or integrated into long-term memory stores. Together, these forces create a dynamic such that memories of the first years of life are formed and may survive over some period of time, but they are challenged to survive for the long term.

Over developmental time, individual memory representations include more and more of the features that characterize autobiographical memories and the features themselves are better elaborated and more tightly integrated with one another. The result is mnemonic materials that are of higher quality. At the same time, there are developments in the neural substrate operating on the available representations, both in terms of the structures involved and in the level of connectivity of the network of structures. The developments herald more efficient and effective cognitive and mnemonic processing, resulting in decreases in the vulnerability of memory traces and, thus, in the rate of forgetting. The net effect is that more and more memories of higher and higher quality survive to be recalled at later points in time, producing the characteristic distribution of autobiographical memories across the life span.

Summary and Conclusions

Autobiographical memory is an important source of "evidence" for continuity of self over time. It is with us virtually for our lifetimes, even though, by late childhood, many memories of early life events are obscured by childhood amnesia. Numerous studies conducted with adults have revealed a relative paucity of memories of events from the first 3 to 4 years of life, with a seemingly gradual increase in the number of memories until approximately age 7 years, after which an adult distribution has been assumed. Historically, the amnesia was attributed to late development of the autobiographical memory system or to inaccessibility resulting from repression or emergence of new cognitive structures. Because they emphasized one or the other—but not both—sides of the mnemonic coin, traditional theories had difficulty accounting for all of the data. Theories that postulate late development of the autobiographical memory system are challenged by data that in-

dicate autobiographical memories within the period eventually obscured by childhood amnesia. Theories that postulate an event or change that renders early memories inaccessible to later recollection are challenged to explain why memories from the first years of life are differentially forgotten.

The Complementary Processes Involved in Increases in the Quality of and Decreases in the Vulnerability of Memory Traces

The alternative complementary processes account advanced in this review recognizes both sides of the mnemonic coin: processes that contribute to increases in the quality of memory traces and to decreases in their vulnerability to forgetting. Rather than defining autobiographical memory in terms of criterial features, which necessitates exclusion of seemingly appropriate members of the category, it defines autobiographical memory in terms of characteristic features. Over the course of development, memories bear more and more of the features that are characteristic of the class (see Figure 3). The perspective also recognizes that memory traces weaken to the point of forgetting. However, forgetting is not an event that renders early memories inaccessible to later recollection. Rather, forgetting is the result of inefficient and ineffective cognitive and mnemonic processes performed by an immature neural substrate. The net effect is a perfect storm: the raw materials are less-than-optimal and they are operated upon by a system that itself is less than efficient and effective. The result is an accelerated rate of forgetting relative to that which characterizes adulthood; the specific form of forgetting is exponential. Exponential forgetting reduces the pool of early memories children have formed to a few isolated puddles, making them even more difficult to retrieve. Eventually, with concomitant increases in the quality of the raw materials and the operating system, forgetting slows to the rate observed in adulthood.

The complementary processes conceptualization prompts a change in perspective on the distribution of autobiographical memories across the life span. In Figure 1 is the typical representation
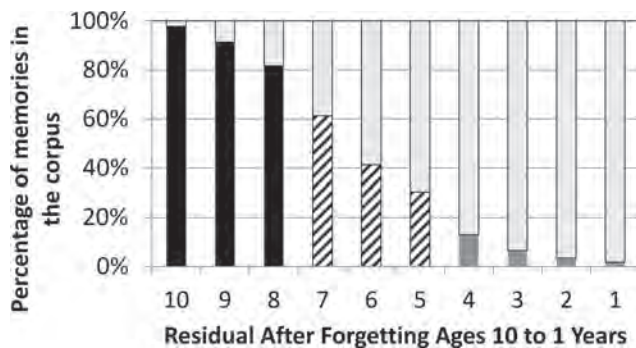
of the distribution of memories over the first decade of life. It lends itself to description in terms of a gradual increase in the density of memories over the first 7 to 10 years of age. However, recognition of early development of autobiographical memory (albeit in a less prototyical, ostrich form) coupled with early vulnerability of those memories in terms of exponential forgetting, prompts a change in the perspective. A more appropriate characterization may be in terms of a decrease in the number of memories as a result of exponential forgetting associated with failed consolidation. This perspective is depicted schematically in Figure 6, which is the mirror image of Figure 1. For each year of life from ages 1 to 10 years, what is represented in the figure is the total corpus of memories formed (100%, represented in light-gray bars) and the percentage of the corpus that remains accessible to recollection. The very short dark gray bar on the far right of the figure indicates that only a very small percentage of memories formed in the first year of life remains in the corpus—most of these very early memories have been lost as a result of failed consolidation and exponential forgetting. In contrast, the tall solid black bar on the far left represents the large percentage of memories formed in the 10th year of life that remain accessible. From this perspective, the tail of the distribution apparent on the right of the figure is the small residual after forgetting, not the start of remembering, as suggested by the corresponding bars on the left side of Figure 1.

Thinking of the tail of the distribution in terms of the residual after forgetting also helps to make sense of the small—yet critically informative—body of literature on childhood amnesia in childhood. By the end of the first decade of life, the average age of earliest memory among children is roughly on par with that among adults—age 3 to 4 years. However, within the preschool years, the average age of earliest memory is substantially earlier, with one estimate as early as 18 months (Tustin & Hayne, 2010). Moreover, children's earliest memory also seems to be inconsistent both in terms of the event nominated as the source of the memory and in terms of the date of the event, which increases systematicly over time (Peterson et al., 2011). These patterns are different from those observed among adults (Bauer et al., in press). They are precisely what is to be expected from a pool of memories that is ever shrinking, eventually diminishing to isolated puddles of memory traces.

Directions for Future Research

By the measure of the ability of the perspective to bring order to the existing literature, the value of the complementary processes account outlined here is clear. Another index of heuristic value is the extent to which a perspective guides future research. In this regard, the message also is clear—future research should feature measures of increases in the quality of memory traces over the course of development, measures of the rate at which memory traces are weakened and lost, and how they relate with one another.

There have been numerous studies of increases in the quality of memory traces over the course of development. Indeed, virtually all of the research on the early development of autobiographical memory involves assessment of changes in how adequately children recall past events. However, most of the studies have asked how the verbal reports or narrative descriptions of memories of past events change with development. They have not asked

Figure 6. Schematic depiction of the distribution of memories across the first decade of life from the complementary processes perspective, suggesting a residual number of early memories remaining after forgetting. For each year of life from ages 1 to 10 years, what is represented is the total corpus of memories formed (100%, represented in light gray bars) and the percentage of the corpus that remains accessible to recollection (dark gray bars for ages 1–4 years, striped bars for ages 5–7 years, and solid black bars for ages 8–10 years).

whether with development, children's memory representations tend to include a larger number of event features, better elaborated event features, more tightly related or integrated features, and so forth. These two questions are related, to be sure. However, as argued elsewhere in this article, verbal reports and memory representations are not one in the same. Examinations of the integrity of memory traces that do not rely on verbal expression would inform the question of whether the strength of memory representations changes in the manner implied in this review. An example of this approach from the infancy literature is available in Bauer and Lukowski (2010). We tested infants' recognition of the specific features of objects used to produce multistep sequences in an imitation-based task and how the specificity of feature memory related to recall of the events 1 month later. Infants who had higher levels of accurate recognition of the particular objects actually used—from among highly perceptually similar distractor objects—had higher levels of long-term recall. Analogous tests with older children would provide valuable information on the nature and pace of change in what is represented in memory—in this case, the specificity of the representation—and how it relates to recall of personally relevant past events over long delays.

Another promising line of research is suggested by studies of developmental changes in children's abilities to create conjunctions between items and their locations (e.g., Bauer, Doydum, et al., 2012; Bauer et al., in press; Sluzenski et al., 2006), or information and the source from which it was acquired (e.g., Cycowicz, Friedman, Snodgrass, & Duff, 2001; Drummey & Newcombe, 2002; Riggins, 2014). The ability to form these conjunctions is a crucial step in the process of transition of labile patterns of neural representation into enduring memory traces. Studies of various conjunctions in memory find age-related differences across early childhood (see Olson & Newcombe, 2014, for a review), suggesting that memory representations strengthen in this manner. Because these studies typically are conducted with controlled laboratory stimuli (e.g., pictures of common objects such as animals and kitchen utensils; Cycowicz et al., 2001), the developmental course for conjunctions among the elements of naturally occurring, personally relevant past events is not known. Moreover, how the ability to create and maintain conjunctions among the elements of experience relates to developments in autobiographical memory has yet to be examined.

Future research on developmental increases in the quality of memory representations will tell only part of the story. There also is need for future research on how developmental changes in the rate of forgetting relate to autobiographical memory. To date, relevant evidence comes primarily from the cue word technique and findings that in childhood, forgetting is better characterized by the exponential than by the power function (Bauer et al., 2007; Bauer & Larkina, 2014). Such studies provide an estimate of forgetting in a population; they are not revealing of the rate of forgetting among individuals. Once again, a model for future investigation is available in the infancy literature. As discussed above, my colleagues and I have examined the rate of forgetting in infancy and the preschool years by probing the integrity of memory traces at various points post encoding, ranging from minutes after experience of an event to 2 weeks later. We then examine the proportion of variance in subsequent long-term recall explained by the measures obtained shortly after encoding, as memory traces presumably are undergoing consolidation. We have found that

measures of memory obtained hours to days after experience of events are predictive of recall after delays of weeks to months (e.g., Bauer, Larkina, et al., 2012; Pathman & Bauer, 2013). Similar designs could be used in tests of older children's memories of specific past events to determine how the rate of forgetting relates to long-term recall.

Future research also is needed to more accurately pin-point the transition from a rate of forgetting that it better characterized by the exponential function to a rate that is better characterized by the power function observed in adults. In Bauer and Larkina (2014), we found that the power function provided a good fit to data obtained from college students as well as middle age adults. In contrast, the exponential function was a better fit to the data obtained from children 7 to 11 years of age as a whole (see also Bauer et al., 2007) and for each of the age groups separately. However, there also was evidence of a deceleration in the rate of forgetting even within childhood. Specifically, between 7 and 11 years, the rate of forgetting, represented by the b parameter of the power function (e.g., Wixted & Ebbesen, 1997), decreased from 2.21 to 1.62; there was a further decrease to 1.01 by the college years. A productive direction for future research would be to obtain estimates of forgetting across the childhood years and use the estimates to predict the rate of forgetting of events experienced at different points in development.

Another direction for future research is to examine factors that have been found to relate to increases in the quality of memory traces and to declines in their vulnerability to forgetting, and examine their influence on the complementary process. For example, one of the most robust determinants of developmental improvements in autobiographical memory—as measured by the quality of narratives about past events—is the conversational style used by children's parents. As summarized by Fivush (2014), adults exhibit two different styles when they engage in memory talk with their children. Parents who frequently engage in conversations about the past, provide rich descriptive information about previous experiences, and invite their children to "join in" on memory conversations, are said to use an *elaborative* style. In contrast, parents who provide fewer details and instead pose specific questions to their children (e.g., "What was the name of the restaurant where we had breakfast?"), are said to use a *repetitive* or *low elaborative* style. Numerous studies have found that children of parents using the elaborative style report more about past events than children of parents using the repetitive or low elaborative style (see Fivush, 2014, for a review). Relations are observed both concurrently and over time (e.g., Reese, Haden, & Fivush, 1993). In future research it would be informative to test whether maternal narrative style also is predictive of changes in rates of the forgetting of memory traces.

Research on the complementary approach of examining determinants of "forgetting" to see how they relate to developmental improvements in autobiographical memory is underway. As outlined in this review, an important determinant of forgetting is the integrity of the neural substrate responsible for memory trace formation, consolidation, and subsequent retrieval. In the adult literature, there have been numerous tests of relations between neural function and episodic encoding and retrieval (e.g., Nyberg et al., 2000) as well as autobiographical memory retrieval (e.g., Cabeza et al., 2004; Daselaar et al., 2008; McCormick et al., 2013). There are even studies that test relations between the integrity of

memory traces during the period of consolidation and subsequent recall success (Bosshardt et al., 2005). The developmental literature on relations between neural structures and neural function and episodic memory behavior is substantially smaller than that on relations in adulthood (see Ghetti & Lee, 2014, for a review). The literature on relations to recall of personally relevant past events is nonexistent. Clearly, further research is needed.

Conclusion

The distribution of autobiographical memories across the life span has been a matter of considerable scientific curiosity and debate since the end of the 19th century. Typical depictions of the distribution (see Figure 1) plot the event of birth on the left and show what appears to be a gradually increasing number of memories from birth through the first 7 to 10 years of life, after which an adult-like distribution is assumed. In the context of this article, I have argued that from the perspective of development, a better characterization is achieved by "flipping" the chart and plotting the event of birth on the right of the distribution (see Figure 6). The plot then shows the large number of autobiographical memories of events experienced in older childhood (and adulthood, on the left of the graph), and the relatively small residual number of memories of early life events that remain in the corpus. The residual are the puddles of memories that are spared from the constant rate of forgetting that characterizes mnemonic life during childhood. This perspective assumes relatively early development of the ability to form, retain, and later retrieve memories of personally relevant past events. Early in life, the event memories bear relatively fewer of the features that we associate with adult-like autobiographical memory. Over the course of development, more and more memories take on more and more of the features of autobiographical memory, rendering them both more obvious members of the class, and more impervious to the ravages of forgetting. This perspective recognizes the complementary processes involved in increases in the quality of memory traces and decreases in their vulnerability to forgetting, and the developmental dynamics of both. It accounts for the distribution of autobiographical memories across the life span, which forms the basis for construction of a personal past.

References

Ábrahám, H., Vincze, A., Jewgenow, I., Veszprémi, B., Kravják, A., Gömöri, E., & Seress, L. (2010). Myelination in the human hippocampal formation from midgestation to adulthood. *International Journal of Developmental Neuroscience, 28,* 401–410. http://dx.doi.org/10.1016/j.ijdevneu.2010.03.004

Addis, D. R., McIntosh, A. R., Moscovitch, M., Crawley, A. P., & McAndrews, M. P. (2004). Characterizing spatial and temporal features of autobiographical memory retrieval networks: A partial least squares approach. *NeuroImage, 23,* 1460–1471. http://dx.doi.org/10.1016/j.neuroimage.2004.08.007

Akers, K. G., Martinez-Canabal, A., Restivo, L., Yiu, A. P., De Cristofaro, A., Hsiang, H.-L., . . . Frankland, P. W. (2014). Hippocampal neurogenesis regulates forgetting during adulthood and infancy. *Science, 344,* 598–602. http://dx.doi.org/10.1126/science.1248903

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences of the United States of America, 91,* 7041–7045. http://dx.doi.org/10.1073/pnas.91.15.7041

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087. http://dx.doi.org/10.1037/0278-7393.20.5.1063

Arnold, S. E., & Trojanowski, J. Q. (1996). Human fetal hippocampal development: I. Cytoarchitecture, myeloarchitecture, and neuronal morphologic features. *The Journal of Comparative Neurology, 367,* 274–292. http://dx.doi.org/10.1002/(SICI)1096-9861(19960401)367:2<274::AID-CNE9>3.0.CO;2-2

Bachevalier, J. (2014). The development of memory from a neurocognitive and comparative perspective. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 109–125). West Sussex, United Kingdom: Wiley-Blackwell.

Baddeley, A., Eysenck, M. W., & Anderson, M. C. (2009). *Memory.* New York, NY: Psychology Press.

Bauer, P. J. (1992). Holding it all together: How enabling relations facilitate young children's event recall. *Cognitive Development, 7,* 1–28. http://dx.doi.org/10.1016/0885-2014(92)90002-9

Bauer, P. J. (2005). Developments in declarative memory: Decreasing susceptibility to storage failure over the second year of life. *Psychological Science, 16,* 41–47. http://dx.doi.org/10.1111/j.0956-7976.2005.00778.x

Bauer, P. J. (2007). *Remembering the times of our lives: Memory in infancy and beyond.* Mahwah, NJ: Erlbaum.

Bauer, P. J. (2008). Infantile amnesia. In M. M. Haith & J. B. Benson (Eds.), *Encyclopedia of infant and early childhood development* (pp. 51–62). San Diego, CA: Academic Press. http://dx.doi.org/10.1016/B978-012370877-9.00007-4

Bauer, P. J. (2009a). The cognitive neuroscience of the development of memory. In M. L. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (2nd ed., pp. 115–144). New York, NY: Psychology Press.

Bauer, P. J. (2009b). Neurodevelopmental changes in infancy and beyond: Implications for learning and memory. In O. A. Barbarin & B. H. Wasik (Eds.), *Handbook of child development and early education: Research to practice* (pp. 78–102). New York, NY: Guilford Press.

Bauer, P. J. (2012). The life I once remembered: The waxing and waning of early memories. In D. Berntsen & D. C. Rubin (Eds.), *Understanding autobiographical memory: Theories and approaches* (pp. 205–225). Cambridge, United Kingdom: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139021937.016

Bauer, P. J. (2013). Memory. In P. D. Zelazo (Ed.), *Oxford handbook of developmental psychology* (Vol. 1, pp. 505–541). New York, NY: Oxford University Press.

Bauer, P. J. (2014). The development of forgetting: Childhood amnesia. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 519–544). West Sussex, United Kingdom: Wiley-Blackwell.

Bauer, P. J., Burch, M. M., Scholin, S. E., & Güler, O. E. (2007). Using cue words to investigate the distribution of autobiographical memories in childhood. *Psychological Science, 18,* 910–916. http://dx.doi.org/10.1111/j.1467-9280.2007.01999.x

Bauer, P. J., & Dow, G. A. A. (1994). Episodic memory in 16- and 20-month-old children: Specifics are generalized, but not forgotten. *Developmental Psychology, 30,* 403–417. http://dx.doi.org/10.1037/0012-1649.30.3.403

Bauer, P. J., Doydum, A. O., Pathman, T., Larkina, M., Güler, O. E., & Burch, M. (2012). It's all about location, location, location: Children's memory for the "where" of personally experienced events. *Journal of Experimental Child Psychology, 113,* 510–522. http://dx.doi.org/10.1016/j.jecp.2012.06.007

Bauer, P. J., Güler, O. E., Starr, R. M., & Pathman, T. (2011). Equal learning does not result in equal remembering: The importance of post-encoding processes. *Infancy, 16,* 557–586. http://dx.doi.org/10.1111/j.1532-7078.2010.00057.x

Bauer, P. J., Hertsgaard, L. A., Dropik, P., & Daly, B. P. (1998). When even arbitrary order becomes important: Developments in reliable temporal sequencing of arbitrarily ordered events. *Memory, 6,* 165–198. http://dx.doi.org/10.1080/741942074

Bauer, P. J., & Larkina, M. (2013). The onset of childhood amnesia in childhood: A prospective investigation of the course and determinants of forgetting of early-life events. *Memory, 22,* 907–924. http://dx.doi.org/10.1080/09658211.2013.854806

Bauer, P. J., & Larkina, M. (2014). Childhood amnesia in the making: Different distributions of autobiographical memories in children and adults. *Journal of Experimental Psychology: General, 143,* 597–611. http://dx.doi.org/10.1037/a0033307

Bauer, P. J., Larkina, M., & Doydum, A. O. (2012). Explaining variance in long-term recall in 3- and 4-year-old children: The importance of post-encoding processes. *Journal of Experimental Child Psychology, 113,* 195–210. http://dx.doi.org/10.1016/j.jecp.2012.05.006

Bauer, P. J., & Leventon, J. S. (2013). Memory for one-time experiences in the second year of life: Implications for the status of episodic memory. *Infancy, 18,* 755–781.

Bauer, P. J., & Lukowski, A. F. (2010). The memory is in the details: Relations between memory for the specific features of events and long-term recall in infancy. *Journal of Experimental Child Psychology, 107,* 1–14. (ID# NIHMS196588)

Bauer, P. J., & Mandler, J. M. (1989). One thing follows another: Effects of temporal structure on one- to two-year-olds' recall of events. *Developmental Psychology, 25,* 197–206. http://dx.doi.org/10.1037/0012-1649.25.2.197

Bauer, P. J., & Shore, C. M. (1987). Making a memorable event: Effects of familiarity and organization on young children's recall of action sequences. *Cognitive Development, 2,* 327–338. http://dx.doi.org/10.1016/S0885-2014(87)80011-4

Bauer, P. J., Stark, E. N., Lukowski, A. F., Rademacher, J., Van Abbema, D. L., & Ackil, J. K. (2005). Working together to make sense of the past: Mothers' and children's use of internal states language in conversations about traumatic and non-traumatic events. *Journal of Cognition and Development, 6,* 463–488. http://dx.doi.org/10.1207/s15327647jcd0604_2

Bauer, P. J., Stennes, L., & Haight, J. C. (2003). Representation of the inner self in autobiography: Women's and men's use of internal states language in personal narratives. *Memory, 11,* 27–42.

Bauer, P. J., Stewart, R., White, E. A., & Larkina, M. (in press). *A place for every event and every event in its place: Memory for locations and activities by 4-year-old children.* Manuscript under review. http://dx.doi.org/10.1080/15248372.2014.949521

Bauer, P. J., Tasdemir-Ozdes, A., & Larkina, M. (2014). Adults' reports of their earliest memories: Consistency in events, ages, and narrative characteristics over time. *Consciousness and Cognition, 27,* 76–88. http://dx.doi.org/10.1016/j.concog.2014.04.008

Bauer, P. J., Van Abbema, D. L., & de Haan, M. (1999). In for the short haul: Immediate and short-term remembering and forgetting by 20-month-old children. *Infant Behavior and Development, 22,* 321–343. http://dx.doi.org/10.1016/S0163-6383(99)00014-4

Bauer, P. J., Wenner, J. A., Dropik, P. L., & Wewerka, S. S. (2000). Parameters of remembering and forgetting in the transition from infancy to early childhood. *Monographs of the Society for Research in Child Development, 65* (4, Serial No. 263).

Bauer, P. J., & Wewerka, S. S. (1997). Saying is revealing: Verbal expression of event memory in the transition from infancy to early childhood. In P. van den Broek, P. J. Bauer, & T. Bourg (Eds.), *Developmental spans in event representation and comprehension: Bridging fictional and actual events* (pp. 139–168). Mahwah, NJ: Erlbaum.

Bauer, P. J., Wiebe, S. A., Carver, L. J., Lukowski, A. F., Haight, J. C., Waters, J. M., & Nelson, C. A. (2006). Electrophysiological indexes of encoding and behavioral indexes of recall: Examining relations and developmental change late in the first year of life. *Developmental Neuropsychology, 29,* 293–320. http://dx.doi.org/10.1207/s15326942dn2902_2

Bauer, P. J., Wiebe, S. A., Carver, L. J., Waters, J. M., & Nelson, C. A. (2003). Developments in long-term explicit memory late in the first year of life: Behavioral and electrophysiological indices. *Psychological Science, 14,* 629–635. http://dx.doi.org/10.1046/j.0956-7976.2003.psci_1476.x

Benes, F. M. (2001). The development of prefrontal cortex: The maturation of neurotransmitter systems and their interaction. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 79–92). Cambridge, MA: The MIT Press.

Benes, F. M., Turtle, M., Khan, Y., & Farol, P. (1994). Myelination of a key relay zone in the hippocampal formation occurs in the human brain during childhood, adolescence, and adulthood. *Archives of General Psychiatry, 51,* 477–484. http://dx.doi.org/10.1001/archpsyc.1994.03950060041004

Bluck, S., & Alea, N. (2008). Remembering being me: The self-continuity function of autobiographical memory in younger and older adults. In F. Sani (Ed.), *Self-continuity: Individual and collective perspectives* (pp. 55–70). New York, NY: Erlbaum.

Bohn, A., & Berntsen, D. (2008). Life story development in childhood: The development of life story abilities and the acquisition of cultural life scripts from late middle childhood to adolescence. *Developmental Psychology, 44,* 1135–1147. http://dx.doi.org/10.1037/0012-1649.44.4.1135

Bohn, A., & Berntsen, D. (2014). Cultural life scripts and the development of personal memories. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 626–644). West Sussex, United Kingdom: Wiley-Blackwell.

Bosshardt, S., Degonda, N., Schmidt, C. F., Boesiger, P., Nitsch, R. M., Hock, C., & Henke, K. (2005). One month of human memory consolidation enhances retrieval-related hippocampal activity. *Hippocampus, 15,* 1026–1040. http://dx.doi.org/10.1002/hipo.20105

Brewer, W. F. (1980). Literary theory, rhetoric, stylistics: Implications for psychology. In R. J. Spiro, B. C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 221–239). Hillsdale, NJ: Erlbaum.

Brewer, W. F. (1986). What is autobiographical memory? In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 25–49). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511558313.006

Cabeza, R., Prince, S. E., Daselaar, S. M., Greenberg, D. L., Budde, M., Dolcos, F., . . . Rubin, D. C. (2004). Brain activity during episodic retrieval of autobiographical and laboratory events: An fMRI study using a novel photo paradigm. *Journal of Cognitive Neuroscience, 16,* 1583–1594.

Carver, L. J., & Bauer, P. J. (1999). When the event is more than the sum of its parts: 9-month-olds' long-term ordered recall. *Memory, 7,* 147–174. http://dx.doi.org/10.1080/741944070

Carver, L. J., & Bauer, P. J. (2001). The dawning of a past: The emergence of long-term explicit memory in infancy. *Journal of Experimental Psychology: General, 130,* 726–745.

Caviness, V. S., Jr., Kennedy, D. N., Richelme, C., Rademacher, J., & Filipek, P. A. (1996). The human brain age 7–11 years: A volumetric analysis based on magnetic resonance images. *Cerebral Cortex, 6,* 726–736. http://dx.doi.org/10.1093/cercor/6.5.726

Chugani, H. T. (1994). Development of regional blood glucose metabolism in relation to behavior and plasticity. In G. Dawson & K. Fischer (Eds.), *Human behavior and the developing brain* (pp. 153–175). New York, NY: Guilford Press.

Chugani, H. T., & Phelps, M. E. (1986). Maturational changes in cerebral function in infants determined by 18FDG positron emission tomography. *Science, 231,* 840–843. http://dx.doi.org/10.1126/science.3945811

Cleveland, E. S., & Reese, E. (2008). Children remember early childhood: Long-term recall across the offset of childhood amnesia. *Applied Cognitive Psychology, 22,* 127–142. http://dx.doi.org/10.1002/acp.1359

Conway, M. A. (1996). Autobiographical knowledge and autobiographical memories. In D. C. Rubin (Ed.), *Remembering our past: Studies in autobiographical memory* (pp. 67–93). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511527913.003

Conway, M. A. (2005). Memory and the self. *Journal of Memory and Language, 53,* 594–628. http://dx.doi.org/10.1016/j.jml.2005.08.005

Conway, M. A., & Pleydell-Pearce, C. W. (2000). The construction of autobiographical memories in the self-memory system. *Psychological Review, 107,* 261–288. http://dx.doi.org/10.1037/0033-295X.107.2.261

Crovitz, H. F., & Schiffman, H. (1974). Frequency of episodic memories as a function of their age. *Bulletin of the Psychonomic Society, 4,* 517–518. http://dx.doi.org/10.3758/BF03334277

Cycowicz, Y. M., Friedman, D., Snodgrass, J. G., & Duff, M. (2001). Recognition and source memory for pictures in children and adults. *Neuropsychologia, 39,* 255–267. http://dx.doi.org/10.1016/S0028-3932(00)00108-1

Daselaar, S. M., Rice, H. J., Greenberg, D. L., Cabeza, R., LaBar, K. S., & Rubin, D. C. (2008). The spatiotemporal dynamics of autobiographical memory: Neural correlates of recall, emotional intensity, and reliving. *Cerebral Cortex, 18,* 217–229. http://dx.doi.org/10.1093/cercor/bhm048

DeMaster, D., Pathman, T., Lee, J. K., & Ghetti, S. (2012). *Structural development of the hippocampus and episodic memory: Developmental dissociations along the anterior/posterior axis*. Manuscript under review.

Drummey, A. B., & Newcombe, N. S. (2002). Developmental changes in source memory. *Developmental Science, 5,* 502–513. http://dx.doi.org/10.1111/1467-7687.00243

Dudycha, G. J., & Dudycha, M. M. (1933a). Adolescents' memories of preschool experiences. *The Journal of Genetic Psychology, 42,* 468–480.

Dudycha, G. J., & Dudycha, M. M. (1933b). Some factors and characteristics of childhood memories. *Child Development, 4,* 265–278. http://dx.doi.org/10.2307/1125689

Eacott, M. J., & Crawley, R. A. (1998). The offset of childhood amnesia: Memory for events that occurred before age 3. *Journal of Experimental Psychology: General, 127,* 22–33. http://dx.doi.org/10.1037/0096-3445.127.1.22

Ebbinghaus, H. (1885). *On memory* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Teachers' College, 1913. Paperback ed., New York: Dover, 1964.

Eckenhoff, M. F., & Rakic, P. (1991). A quantitative analysis of synaptogenesis in the molecular layer of the dentate gyrus in the rhesus monkey. *Brain Research Developmental Brain Research, 64,* 129–135. http://dx.doi.org/10.1016/0165-3806(91)90216-6

Eichenbaum, H., & Cohen, N. J. (2001). *From conditioning to conscious recollection: Memory systems of the brain*. New York, NY: Oxford University Press.

Fivush, R. (2012). Subjective perspective and personal timeline in the development of autobiographical memory. In D. Berntsen & D. C. Rubin (Eds.), *Understanding autobiographical memory: Theories and approaches* (pp. 226–245). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139021937.017

Fivush, R. (2014). Maternal reminiscing style: The sociocultural construction of autobiographical memory across childhood and adolescence. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 568–585). West Sussex, United Kingdom: Wiley-Blackwell.

Fivush, R., Habermas, T., Waters, T. E. A., & Zaman, W. (2011). The making of autobiographical memory: Intersections of culture, narratives and identity. *International Journal of Psychology, 46,* 321–345. http://dx.doi.org/10.1080/00207594.2011.596541

Fivush, R., & Schwarzmueller, A. (1998). Children remember childhood: Implications for childhood amnesia. *Applied Cognitive Psychology, 12,* 455–473. http://dx.doi.org/10.1002/(SICI)1099-0720(199810)12:5<455::AID-ACP534>3.0.CO;2-H

Fivush, R., & Zaman, W. (2014). Gender, subjective perspective, and autobiographical consciousness. In P. J. Bauer & R. Fivush (Eds.), *The Wiley handbook on the development of children's memory* (Vol. II, pp. 586–604). Chinchester: Wiley.

Frankland, P. W., Köhler, S., & Josselyn, S. A. (2013). Hippocampal neurogenesis and forgetting. *Trends in Neurosciences, 36,* 497–503. http://dx.doi.org/10.1016/j.tins.2013.05.002

Freud, S. (1905/1953). Childhood and concealing memories. In A. A. Brill (Trans. & Ed.), *The basic writings of Sigmund Freud* (pp. 62–68). New York, NY: The Modern Library.

Freud, S. (1916/1966). The archaic features and infantilism of dreams. In J. Strachey (Trans. & Ed.), *Introductory lectures on psychoanalysis* (pp. 199–212). New York, NY: Norton.

Freud, S. (1920/1935). *A general introduction to psycho-analysis* (J. Riviere, Trans.). Garden City, NY: Garden City Publishing Company, Inc. http://dx.doi.org/10.1037/10667-000

Friedman, W. J. (2014). The development of memory for the times of past events. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 394–407). West Sussex, United Kingdom: Wiley-Blackwell.

Friedman, W. J., Reese, E., & Dai, J. (2011). Children's memory for the times of events from the past years. *Applied Cognitive Psychology, 25,* 156–165. http://dx.doi.org/10.1002/acp.1656

Fuster, J. M. (1997). Network memory. *Trends in Neurosciences, 20,* 451–459. http://dx.doi.org/10.1016/S0166-2236(97)01128-4

Ghetti, S., & Bunge, S. A. (2012). Neural changes underlying the development of episodic memory during middle childhood. *Developmental Cognitive Neuroscience, 2,* 381–395. http://dx.doi.org/10.1016/j.dcn.2012.05.002

Ghetti, S., DeMaster, D. M., Yonelinas, A. P., & Bunge, S. A. (2010). Developmental differences in medial temporal lobe function during memory encoding. *The Journal of Neuroscience, 30,* 9548–9556. http://dx.doi.org/10.1523/JNEUROSCI.3500-09.2010

Ghetti, S., & Lee, J. K. (2014). The development of recollection and familiarity during childhood: Insight from studies of behavior and brain. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 309–335). West Sussex, United Kingdom: Wiley-Blackwell.

Giedd, J. N., Blumenthal, J., Jeffries, N. O., Castellanos, F. X., Liu, H., Zijdenbos, A., . . . Rapoport, J. L. (1999). Brain development during childhood and adolescence: A longitudinal MRI study. *Nature Neuroscience, 2,* 861–863. http://dx.doi.org/10.1038/13158

Gilboa, A. (2004). Autobiographical and episodic memory—One and the same? Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia, 42,* 1336–1349. http://dx.doi.org/10.1016/j.neuropsychologia.2004.02.014

Gilmore, J. H., Shi, F., Woolson, S. L., Knickmeyer, R. C., Short, S. J., Lin, W., . . . Shen, D. (2012). Longitudinal Development of Cortical and Subcortical Gray Matter from Birth to 2 Years. *Cerebral Cortex, 22,* 2478–2485.

Gogtay, N., Giedd, J. N., Lusk, L., Hayashi, K. M., Greenstein, D., Vaituzis, A. C., . . . Thompson, P. M. (2004). Dynamic mapping of human cortical development during childhood through early adulthood. *Proceedings of the National Academy of Sciences of the United States of America, 101,* 8174–8179. http://dx.doi.org/10.1073/pnas.0402680101

Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. In F. Plum (Ed.), *Handbook of physiology, the nervous system, higher functions of the brain* (Vol. 5, pp. 373–417). Bethesda, MD: American Physiological Society.

Greenberg, D. L., Rice, H. J., Cooper, J. J., Cabeza, R., Rubin, D. C., & Labar, K. S. (2005). Co-activation of the amygdala, hippocampus and inferior frontal gyrus during autobiographical memory retrieval. *Neuropsychologia, 43,* 659–674. http://dx.doi.org/10.1016/j.neuropsychologia.2004.09.002

Habermas, T., & Bluck, S. (2000). Getting a life: The emergence of the life story in adolescence. *Psychological Bulletin, 126,* 748–769.

Habermas, T., & Köber, C. (2014). Autobiographical reasoning is constitutive for narrative identity: The role of the life story for personal continuity. In K. C. McLean & M. Syed (Eds.), *The Oxford handbook of identity development* (pp. 267–299). Oxford, United Kingdom: Oxford University Press.

Habermas, T., Negele, A., & Mayer, F. B. (2010). Honey, you're jumping about: Mothers' scaffolding of their children's and adolescents' life narration. *Cognitive Development, 25,* 339–351. http://dx.doi.org/10.1016/j.cogdev.2010.08.004

Haden, C., Haine, R., & Fivush, R. (1997). Development narrative structure in parent-child reminiscing across the preschool years. *Developmental Psychology, 33,* 295–307. http://dx.doi.org/10.1037/0012-1649.33.2.295

Hamond, N. R., & Fivush, R. (1991). Memories of Mickey Mouse: Young children recount their trip to Disneyworld. *Cognitive Development, 6,* 433–448. http://dx.doi.org/10.1016/0885-2014(91)90048-I

Harley, K., & Reese, E. (1999). Origins of autobiographical memory. *Developmental Psychology, 35,* 1338–1348. http://dx.doi.org/10.1037/0012-1649.35.5.1338

Henri, V., & Henri, C. (1895). On our earliest recollections of childhood. *Psychological Review, 2,* 215–216.

Henri, V., & Henri, C. (1896). Enquete sur les premiers souvenirs de l'enfance. *L'Année Psychologique, 3,* 184–198. http://dx.doi.org/10.3406/psy.1896.1831

Henri, V., & Henri, C. (1898). Earliest recollections. *Popular Science Monthly, 53,* 108–115.

Howe, M. L. (2014). The co-emergence of the self and autobiographical memory: An adaptive view of early memory. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 545–567). West Sussex, United Kingdom: Wiley-Blackwell.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin, 113,* 305–326. http://dx.doi.org/10.1037/0033-2909.113.2.305

Howe, M. L., & Courage, M. L. (1997). The emergence and early development of autobiographical memory. *Psychological Review, 104,* 499–523. http://dx.doi.org/10.1037/0033-295X.104.3.499

Howe, M. L., & O'Sullivan, J. T. (1997). What children's memoriestell us about recalling our childhoods: A review of storage and retrieval processes in the development of long-term retention. *Developmental Review, 17,* 148–204. http://dx.doi.org/10.1006/drev.1996.0428

Huttenlocher, P. R. (1979). Synaptic density in human frontal cortex - developmental changes and effects of aging. *Brain Research, 163,* 195–205. http://dx.doi.org/10.1016/0006-8993(79)90349-4

Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia, 28,* 517–527. http://dx.doi.org/10.1016/0028-3932(90)90031-I

Huttenlocher, P. R., & Dabholkar, A. S. (1997). Regional differences in synaptogenesis in human cerebral cortex. *The Journal of Comparative Neurology, 387,* 167–178. http://dx.doi.org/10.1002/(SICI)1096-9861(19971020)387:2<167::AID-CNE1>3.0.CO;2-Z

Inhelder, B., & Piaget, J. (1958). *The Growth of Logical Thinking from Childhood to Adolescence*. New York, NY: Basic Books. http://dx.doi.org/10.1037/10034-000

Jack, F., & Hayne, H. (2010). Childhood amnesia: Empirical evidence for a two-stage phenomenon. *Memory, 18,* 831–844.

Jack, F., MacDonald, S., Reese, E., & Hayne, H. (2009). Maternal reminiscing style during early childhood predicts the age of adolescents' earliest memories. *Child Development, 80,* 496–505. http://dx.doi.org/10.1111/j.1467-8624.2009.01274.x

Jernigan, T. L., Trauner, D. A., Hesselink, J. R., & Tallal, P. A. (1991). Maturation of human cerebrum observed in vivo during adolescence. *Brain: A Journal of Neurology, 114,* 2037–2049. http://dx.doi.org/10.1093/brain/114.5.2037

Johnson, M. H. (1997). *Developmental cognitive neuroscience*. Oxford, United Kingdom: Blackwell Publishers, Inc.

Kandel, E. R., & Squire, L. R. (2000). Neuroscience: Breaking down scientific barriers to the study of brain and mind. *Science, 290,* 1113–1120. http://dx.doi.org/10.1126/science.290.5494.1113

Kennedy, D. N., Makris, N., Herbert, M. R., Takahashi, T., & Caviness, V. S., Jr. (2002). Basic principles of MRI and morphometry studies of human brain development. *Developmental Science, 5,* 268–278. http://dx.doi.org/10.1111/1467-7687.00366

Kihlstrom, J. F., & Harackiewicz, J. M. (1982). The earliest recollection: A new survey. *Journal of Personality, 50,* 134–148. http://dx.doi.org/10.1111/j.1467-6494.1982.tb01019.x

Klingberg, T., Vaidya, C. J., Gabrieli, J. D., Moseley, M. E., & Hedehus, M. (1999). Myelination and organization of the frontal white matter in children: A diffusion tensor MRI study. *NeuroReport, 10,* 2817–2821. http://dx.doi.org/10.1097/00001756-199909090-00022

Kuhn, H. G., Dickinson-Anson, H., & Gage, F. H. (1996). Neurogenesis in the dentate gyrus of the adult rat: Age-related decrease of neuronal progenitor proliferation. *The Journal of Neuroscience, 16,* 2027–2033.

Larkina, M., Merrill, N., Fivush, R., & Bauer, P. J. (2009, October). *Linking children's earliest memories and maternal reminiscing style*. Poster presented to the Cognitive Development Society, San Antonio, TX.

Lechuga, M. T., Marcos-Ruiz, R., & Bauer, P. J. (2001). Episodic recall of specifics and generalisation coexist in 25-month-old children. *Memory, 9,* 117–132. http://dx.doi.org/10.1080/09658210042000111

Lukowski, A. F., & Bauer, P. J. (2014). Long-term memory in infancy and early childhood. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 230–254). West Sussex, United Kingdom: Wiley-Blackwell.

Lukowski, A. F., Garcia, M. T., & Bauer, P. J. (2011). Memory for events and locations obtained in the context of elicited imitation: Evidence for differential retention in the second year of life. *Infant Behavior and Development, 34,* 55–62. http://dx.doi.org/10.1016/j.infbeh.2010.09.006

Maguire, E. A. (2001). Neuroimaging studies of autobiographical event memory. *Philosophical Transactions of the Royal Society of London, 356,* 1441–1451. http://dx.doi.org/10.1098/rstb.2001.0944

Manns, J. R., & Eichenbaum, H. (2006). Evolution of declarative memory. *Hippocampus, 16,* 795–808. http://dx.doi.org/10.1002/hipo.20205

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality, 63,* 365–396. http://dx.doi.org/10.1111/j.1467-6494.1995.tb00500.x

McAdams, D. P. (2001). The psychology of life stories. *Review of General Psychology, 5,* 100–122. http://dx.doi.org/10.1037/1089-2680.5.2.100

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457. http://dx.doi.org/10.1037/0033-295X.102.3.419

McCormick, C., St-Laurent, M., Ty, A., Valiante, T. A., & McAndrews, M. P. (2013). Functional and effective hippocampal-neocortical connectivity during construction and elaboration of autobiographical memory retrieval. *Cerebral Cortex*. [Advance online publication.] http://dx.doi.org/10.1093/cercor/bht324

McDonough, L., Mandler, J. M., McKee, R. D., & Squire, L. R. (1995). The deferred imitation task as a nonverbal measure of declarative

memory. *Proceedings of the National Academy of Sciences of the United States of America, 92,* 7580–7584. http://dx.doi.org/10.1073/pnas.92.16.7580

McGaugh, J. L. (2000). Memory: A century of consolidation. *Science, 287,* 248–251. http://dx.doi.org/10.1126/science.287.5451.248

McKenzie, S., & Eichenbaum, H. (2011). Consolidation and reconsolidation: Two lives of memories? *Neuron, 71,* 224–233. http://dx.doi.org/10.1016/j.neuron.2011.06.037

Meltzoff, A. N. (1985). Immediate and deferred imitation in fourteen- and twenty-four-month-old infants. *Child Development, 56,* 62–72.

Miles, C. (1895). A study of individual psychology. *The American Journal of Psychology, 6,* 534–558. http://dx.doi.org/10.2307/1411191

Morris, G., Baker-Ward, L., & Bauer, P. J. (2010). What remains of that day: The survival of children's autobiographical memories across time. *Applied Cognitive Psychology, 24,* 527–544.

Moscovitch, M., & Nadel, L. (1998). Consolidation and the hippocampal complex revisited: In defense of the multiple-trace model. *Current Opinion in Neurobiology, 8,* 297–300. http://dx.doi.org/10.1016/S0959-4388(98)80155-4

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging, and neuropsychological results. *Hippocampus, 10,* 352–368. http://dx.doi.org/10.1002/1098-1063(2000)10:4<352::AID-HIPO2>3.0.CO;2-D

Neisser, U. (1962). Cultural and cognitive discontinuity. In T. E. Gladwin & W. Sturtevant (Eds.), *Anthropology and human behavior* (pp. 54–71). Washington, DC: Anthropological Society of Washington DC.

Nelson, C. A., de Haan, M., & Thomas, K. (2006). Neural bases of cognitive development. In D. Kuhn & R. Siegler (Volume Eds.: *Volume 2–Cognition, perception, and language*), W. Damon & R. M. Lerner (Eds.-in-Chief), *Handbook of child psychology* (6th ed., pp. 3–57). Hoboken, NJ: Wiley

Nelson, K. (1993). Events, narratives, memory: What develops? In C. A. Nelson (Ed.), *The Minnesota symposium on child psychology: Volume 26. Memory and affect in development* (pp. 1–24). Hillsdale, NJ: Erlbaum.

Nelson, K., & Fivush, R. (2004). The emergence of autobiographical memory: A social cultural developmental theory. *Psychological Review, 111,* 486–511. http://dx.doi.org/10.1037/0033-295X.111.2.486

Nyberg, L., Persson, J., Habib, R., Tulving, E., McIntosh, A. R., Cabeza, R., & Houle, S. (2000). Large scale neurocognitive networks underlying episodic memory. *Journal of Cognitive Neuroscience, 12,* 163–173. http://dx.doi.org/10.1162/089892900561805

Ofen, N., Kao, Y. C., Sokol-Hessner, P., Kim, H., Whitfield-Gabrieli, S., & Gabrieli, J. D. (2007). Development of the declarative memory system in the human brain. *Nature Neuroscience, 10,* 1198–1205. http://dx.doi.org/10.1038/nn1950

Ohayon, M. M., Carskadon, M. A., Guilleminault, C., & Vitiello, M. V. (2004). Meta-analysis of quantitative sleep parameters from childhood to old age in healthy individuals: Developing normative sleep values across the human lifespan. *Sleep, 27,* 1255–1273.

O'Kearney, R., Speyer, J., & Kenardy, J. (2007). Children's narrative memory for accidents and their post-traumatic distress. *Applied Cognitive Psychology, 21,* 821–838. http://dx.doi.org/10.1002/acp.1294

Olson, I. R., & Newcombe, N. S. (2014). Binding together the elements of episodes: Relational memory and the developmental trajectory of the hippocampus. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 285–308). West Sussex, United Kingdom: Wiley-Blackwell.

Østby, Y., Tamnes, C. K., Fjell, A. M., & Walhovd, K. B. (2011). Morphometry and connectivity of the fronto-parietal verbal working memory network in development. *Neuropsychologia, 49,* 3854–3862. http://dx.doi.org/10.1016/j.neuropsychologia.2011.10.001

Østby, Y., Tamnes, C. K., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., & Walhovd, K. B. (2009). Heterogeneity in subcortical brain develop-

ment: A structural magnetic resonance imaging study of brain maturation from 8 to 30 years. *The Journal of Neuroscience, 29,* 11772–11782. http://dx.doi.org/10.1523/JNEUROSCI.1242-09.2009

Pathman, T., & Bauer, P. J. (2013). Beyond initial encoding: Measures of the post-encoding status of memory traces predict long-term recall during infancy. *Journal of Experimental Child Psychology, 114,* 321–338. http://dx.doi.org/10.1016/j.jecp.2012.10.004

Pathman, T., Larkina, M., Burch, M., & Bauer, P. J. (2013). Young children's memory for the times of personal past events. *Journal of Cognition and Development, 14,* 120–140. http://dx.doi.org/10.1080/15248372.2011.641185

Pathman, T., & St. Jacques, P. L. (2014). Locating events in personal time: Time in autobiography. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 408–426). West Sussex, United Kingdom: Wiley-Blackwell.

Paz-Alonso, P. M., Ghetti, S., Donohue, S. E., Goodman, G. S., & Bunge, S. A. (2008). Neurodevelopmental correlates of true and false recognition. *Cerebral Cortex, 18,* 2208–2216. http://dx.doi.org/10.1093/cercor/bhm246

Perner, J., & Ruffman, T. (1995). Episodic memory and autonoetic consciousness: Developmental evidence and a theory of childhood amnesia. *Journal of Experimental Child Psychology, 59,* 516–548. http://dx.doi.org/10.1006/jecp.1995.1024

Peterson, C., Grant, V. V., & Boland, L. D. (2005). Childhood amnesia in children and adolescents: Their earliest memories. *Memory, 13,* 622–637. http://dx.doi.org/10.1080/09658210444000278

Peterson, C., Warren, K. L., & Short, M. M. (2011). Infantile amnesia across the years: A 2-year follow-up of children's earliest memories. *Child Development, 82,* 1092–1105. http://dx.doi.org/10.1111/j.1467-8624.2011.01597.x

Pfluger, T., Weil, S., Weis, S., Vollmar, C., Heiss, D., Egger, J., . . . Hahn, K. (1999). Normative volumetric data of the developing hippocampus in children based on magnetic resonance imaging. *Epilepsia, 40,* 414–423. http://dx.doi.org/10.1111/j.1528-1157.1999.tb00735.x

Piaget, J. (1962). *Play, dreams and imitation in childhood.* New York, NY: Norton & Co.

Pillemer, D. B., & White, S. H. (1989). Childhood events recalled by children and adults. In H. W. Reese (Ed.), *Advances in child development and behavior* (Vol. 21, pp. 297–340). Orlando, FL: Academic Press. http://dx.doi.org/10.1016/S0065-2407(08)60291-8

Prebble, S. C., Addis, D. R., & Tippett, L. J. (2013). Autobiographical memory and sense of self. *Psychological Bulletin, 139,* 815–840. http://dx.doi.org/10.1037/a0030146

Rasch, B., Büchel, C., Gais, S., & Born, J. (2007). Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science, 315,* 1426–1429. http://dx.doi.org/10.1126/science.1138581

Reed, J. M., & Squire, L. R. (1998). Retrograde amnesia for facts and events: Findings from four new cases. *The Journal of Neuroscience, 18,* 3943–3954.

Reese, E. (2014). Taking the long way: Longitudinal approaches to autobiographical memory development. In P. J. Bauer & R. Fivush (Eds.), *The Wiley-Blackwell handbook on the development of children's memory* (pp. 972–995). West Sussex, United Kingdom: Wiley-Blackwell.

Reese, E., Haden, C. A., Baker-Ward, L., Bauer, P. J., Fivush, R., & Ornstein, P. A. (2011). Coherence of personal narratives across the lifespan: A multidimensional model and coding method. *Journal of Cognition and Development, 12,* 424–462. http://dx.doi.org/10.1080/15248372.2011.587854

Reese, E., Haden, C. A., & Fivush, R. (1993). Mother-child conversations about the past: Relationships of style and memory over time. *Cognitive Development, 8,* 403–430. http://dx.doi.org/10.1016/S0885-2014(05)80002-4

Reese, E., Jack, F., & White, N. (2010). Origins of adolescents' autobiographical memories. *Cognitive Development, 25,* 352–367. http://dx.doi.org/10.1016/j.cogdev.2010.08.006

Riggins, T. (2014). Longitudinal investigation of source memory reveals different developmental trajectories for item memory and binding. *Developmental Psychology, 50,* 449–459. http://dx.doi.org/10.1037/a0033622

Rubin, D. C. (1982). On the retention function for autobiographical memory. *Journal of Verbal Learning and Verbal Behavior, 21,* 21–38. http://dx.doi.org/10.1016/S0022-5371(82)90423-6

Rubin, D. C. (2000). The distribution of early childhood memories. *Memory, 8,* 265–269. http://dx.doi.org/10.1080/096582100406810

Rubin, D. C. (2005). A basis systems approach to autobiographical memory. *Current Directions in Psychological Science, 14,* 79–83. http://dx.doi.org/10.1111/j.0963-7214.2005.00339.x

Rubin, D. C. (2006). The basic-systems model of episodic memory. *Perspectives on Psychological Science, 1,* 277–311. http://dx.doi.org/10.1111/j.1745-6916.2006.00017.x

Rubin, D. C., & Schulkind, M. D. (1997). Distribution of important and word-cued autobiographical memories in 20-, 35-, and 70-year-old adults. *Psychology and Aging, 12,* 524–535. http://dx.doi.org/10.1037/0882-7974.12.3.524

Rubin, D. C., & Umanath, S. (2015). Event memory: A theory of memory for laboratory, autobiographical, and fictional events. *Psychological Review, 122,* 1–23. http://dx.doi.org/10.1037/a0037907

Rubin, D. C., & Wenzel, A. E. (1996). One hundred years of forgetting: A quantitative description of retention. *Psychological Review, 103,* 734–760. http://dx.doi.org/10.1037/0033-295X.103.4.734

Rubin, D. C., Wetzler, S. E., & Nebes, R. D. (1986). Autobiographical memory across the adult lifespan. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 202–222). Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511558313.018

Schahmann, J. D., & Pandya, D. N. (2006). *Fiber pathways in the brain.* New York, NY: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780195104233.001.0001

Schneider, J. F. L., Il'yasov, K. A., Hennig, J., & Martin, E. (2004). Fast quantitative diffusion-tensor imaging of cerebral white matter from the neonatal period to adolescence. *Neuroradiology, 46,* 258–266. http://dx.doi.org/10.1007/s00234-003-1154-2

Seress, L., & Ábrahám, H. (2008). Pre- and postnatal morphological development of the human hippocampal formation. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (2nd ed., pp. 187–212). Cambridge, MA: The MIT Press.

Sheingold, K., & Tenney, Y. J. (1982). Memory for a salient childhood event. In U. Neisser (Ed.), *Memory observed: Remembering in natural contexts* (pp. 201–212). New York, NY: Freeman and Company.

Shimamura, A. P. (2011). Episodic retrieval and the cortical binding of relational activity. *Cognitive, Affective & Behavioral Neuroscience, 11,* 277–291. http://dx.doi.org/10.3758/s13415-011-0031-4

Sluzenski, J., Newcombe, N. S., & Kovacs, S. L. (2006). Binding, relational memory, and recall of naturalistic events: A developmental perspective. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 89–100. http://dx.doi.org/10.1037/0278-7393.32.1.89

Sowell, E. R., Delis, D., Stiles, J., & Jernigan, T. L. (2001). Improved memory functioning and frontal lobe maturation between childhood and adolescence: A structural MRI study. *Journal of the International Neuropsychological Society, 7,* 312–322. http://dx.doi.org/10.1017/S135561770173305X

Squire, L. R. (1987). *Memory and brain.* New York, NY: Oxford University Press.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231. http://dx.doi.org/10.1037/0033-295X.99.2.195

Squire, L. R., Knowlton, B., & Musen, G. (1993). The structure and organization of memory. *Annual Review of Psychology, 44,* 453–495. http://dx.doi.org/10.1146/annurev.ps.44.020193.002321

St. Jacques, P. L., Kragel, P. A., & Rubin, D. C. (2011). Dynamic neural networks supporting memory retrieval. *NeuroImage, 57,* 608–616. http://dx.doi.org/10.1016/j.neuroimage.2011.04.039

Suddendorf, T., Nielsen, M., & von Gehlen, R. (2011). Children's capacity to remember a novel problem and to secure its future solution. *Developmental Science, 14,* 26–33. http://dx.doi.org/10.1111/j.1467-7687.2010.00950.x

Suzuki, W. A., Miller, E. K., & Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *Journal of Neurophysiology, 78,* 1062–1081.

Svoboda, E., McKinnon, M. C., & Levine, B. (2006). The functional neuroanatomy of autobiographical memory: A meta-analysis. *Neuropsychologia, 44,* 2189–2208. http://dx.doi.org/10.1016/j.neuropsychologia.2006.05.023

Tanapat, P., Hastings, N. B., & Gould, E. (2001). Adult neurogenesis in the hippocampal formation. In C. A. Nelson & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience* (pp. 93–105). Cambridge, MA: The MIT Press.

Thomsen, D. K. (2009). There is more to life stories than memories. *Memory, 17,* 445–457. http://dx.doi.org/10.1080/09658210902740878

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., . . . Morris, R. G. M. (2007). Schemas and memory consolidation. *Science, 316,* 76–82. http://dx.doi.org/10.1126/science.1135935

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). New York, NY: Academic Press.

Tulving, E. (1983). *Elements of episodic memory.* Oxford: Oxford University Press.

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53,* 1–25. http://dx.doi.org/10.1146/annurev.psych.53.100901.135114

Tulving, E. (2005). Episodic memory and autonoesis: Uniquely human? In H. S. Terrace & J. Metcalfe (Eds.), *The missing link in cognition* (pp. 3–56). Oxford: Oxford University Press.

Tustin, K., & Hayne, H. (2010). Defining the boundary: Age-related changes in childhood amnesia. *Developmental Psychology, 46,* 1049–1061. http://dx.doi.org/10.1037/a0020105

Usher, J., & Neisser, U. (1993). Childhood amnesia and the beginnings of memory for four early life events. *Journal of Experimental Psychology: General, 122,* 155–165.

Utsunomiya, H., Takano, K., Okazaki, M., & Mitsudome, A. (1999). Development of the temporal lobe in infants and children: Analysis by MR-based volumetry. *American Journal of Neuroradiology, 20,* 717–723.

Van Abbema, D. L., & Bauer, P. J. (2005). Autobiographical memory in middle childhood: Recollections of the recent and distant past. *Memory, 13,* 829–845. http://dx.doi.org/10.1080/09658210444000430

Waldfogel, S. (1948). The frequency and affective character of childhood memories. *Psychological Monographs, 62* (Whole No. 291).

Wang, Q. (2006). Earliest recollections of self and others in European American and Taiwanese young adults. *Psychological Science, 17,* 708–714. http://dx.doi.org/10.1111/j.1467-9280.2006.01770.x

Wang, Q. (2014). The cultured self and remembering. In P. J. Bauer & R. Fivush (Eds.), *Wiley-Blackwell handbook on the development of children's memory* (pp. 605–625). New York, NY: Wiley-Blackwell.

Wang, Q., Conway, M., & Hou, Y. (2004). Infantile amnesia: A crosscultural investigation. *Cognitive Science, 1,* 123–135.

Waters, T. E. A., Bauer, P. J., & Fivush, R. (2014). Autobiographical memory functions served by multiple event types. *Applied Cognitive Psychology, 28,* 185–195.

Weigle, T. W., & Bauer, P. J. (2000). Deaf and hearing adults' recollections of childhood and beyond. *Memory, 8,* 293–309. http://dx.doi.org/10.1080/09658210050117726

Wendelken, C., Baym, C. L., Gazzaley, A., & Bunge, S. A. (2011). Neural indices of improved attentional modulation over middle childhood. *Developmental Cognitive Neuroscience, 1,* 175–186. http://dx.doi.org/10.1016/j.dcn.2010.11.001

West, T. A., & Bauer, P. J. (1999). Assumptions of infantile amnesia: Are there differences between early and later memories? *Memory, 7,* 257–278. http://dx.doi.org/10.1080/096582199387913

Wetzler, S. E., & Sweeney, J. A. (1986). Childhood amnesia: An empirical demonstration. In D. C. Rubin (Ed.), *Autobiographical memory* (pp. 191–201). New York, NY: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511558313.017

Wheeler, M. A. (2000). Episodic memory and autonoetic awareness. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 597–608). New York, NY: Oxford University Press.

Wickelgren, W. A. (1974). Single-trace fragility theory of memory dynamics. *Memory & Cognition, 2,* 775–780. http://dx.doi.org/10.3758/BF03198154

Wickelgren, W. A. (1975). Alcoholic intoxication and memory storage dynamics. *Memory & Cognition, 3,* 385–389. http://dx.doi.org/10.3758/BF03212929

Wiebe, S. A., & Bauer, P. J. (2005). Interference from additional props in an elicited imitation task: When in sight, firmly in mind. *Journal of Cognition and Development, 6,* 325–363. http://dx.doi.org/10.1207/s15327647jcd0603_2

Wilson, A. E., & Ross, M. (2003). The identity function of autobiographical memory: Time is on our side. *Memory, 11,* 137–149. http://dx.doi.org/10.1080/741938210

Winocur, G., & Moscovitch, M. (2011). Memory transformation and systems consolidation. *Journal of the International Neuropsychological Society, 17,* 766–780. http://dx.doi.org/10.1017/S1355617711000683

Wixted, J. T. (2004). On Common Ground: Jost's (1897) law of forgetting and Ribot's (1881) law of retrograde amnesia. *Psychological Review, 111,* 864–879. http://dx.doi.org/10.1037/0033-295X.111.4.864

Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological Science, 2,* 409–415. http://dx.doi.org/10.1111/j.1467-9280.1991.tb00175.x

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory & Cognition, 25,* 731–739. http://dx.doi.org/10.3758/BF03211316

Zola, S. M., & Squire, L. R. (2000). The medial temporal lobe and the hippocampus. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 485–500). New York, NY: Oxford University Press.