# Mathematical learning theory through time

CrossMark

## Marc W. Howard

*Department of Psychology and Center for Memory and Brain, Boston University, United States*

## HIGHLIGHTS

- Traces themes established by stimulus sampling theory through subsequent memory models.
- Categorizes memory models according to the properties of the memory representations they generate.
- Describes recent neuroscientific results that place constraints on memory models.

## ARTICLE INFO

## ABSTRACT

Stimulus sampling theory (SST: Estes, 1950, 1955a,b, 1959) was the first rigorous mathematical model of learning that posited a central role for an abstract cognitive representation distinct from the stimulus or the response. SST posited that (a) conditioning takes place not on the nominal stimulus presented to the learner, but on a cognitive representation caused by the nominal stimulus, and (b) the cognitive representation caused by a nominal stimulus changes gradually across presentations of that stimulus. Retrieved temporal context models assume that (a) a distributed representation of temporal context changes gradually over time in response to the studied stimuli, and (b) repeating a stimulus can recover a prior state of temporal context. We trace the evolution of these ideas from the early work on SST, and argue that recent neuroscientific evidence provides a physical basis for the abstract models that Estes envisioned more than a half-century ago.

© 2013 Elsevier Inc. All rights reserved.

## 1. Introduction

Scientists working in mathematical learning theory wrote down equations implementing elementary psychological mechanisms. These mechanisms were then treated analytically to generate precise behavioral predictions for a variety of experimental settings. Critically, the equations were not an exercise in simple curve-fitting of behavioral data, but a concrete hypothesis about how the mind learns. In retrospect, given what was known about systems neurobiology in the 1950s, this was an audacious research program. The brain has, in principle, a huge number of degrees of freedom at its disposal to generate behavior. Writing down correct expressions for the actual physical process supporting memory, given only constraints from behavioral data, seems impossible. In this paper, we follow the implications of two key insights introduced and formalized in stimulus sampling theory (SST) through decades of subsequent memory modeling to contemporary findings from cognitive neuroscience. Even though it must have seemed impossible in the 1950s, we argue that the research program of mathematical learning theory has been largely successful in describing essential features of neural data. Moreover,

the key insights of SST were essential in setting the agenda for these developments.

One key insight of SST is that the nominal stimulus – the light or tone presented to the subject – is not isomorphic to the functional stimulus. In Estes (1950), the nominal stimulus evokes a set of "conditioning elements" that can be conditioned to a particular response. In contemporary terms, we might say that the current set of active conditioning elements is the state of a "memory representation" at the time of the presentation of the nominal stimulus. At each moment, the currently active memory representation is conditioned to a response. At later times, the degree to which a particular response will be evoked is determined by the overlap between the currently active memory representation and the stored memory representation in which the response was learned.

The second key insight of SST is the concept that the memory representation following one presentation of a stimulus changes across different presentations of the stimulus. In much the same way that one cannot step into the same river twice, in SST the functional stimulus caused by different presentations of the same nominal stimulus need not be identical. Moreover, in SST, the functional stimulus caused by a particular nominal stimulus changes gradually across multiple presentations of the nominal stimulus (Estes, 1955a,b). This property enabled a treatment of a variety of phenomena that involve sensitivity to temporal variables, such as forgetting, spontaneous recovery, and the spacing effect.

*E-mail address:* MarcWHoward777@gmail.com.
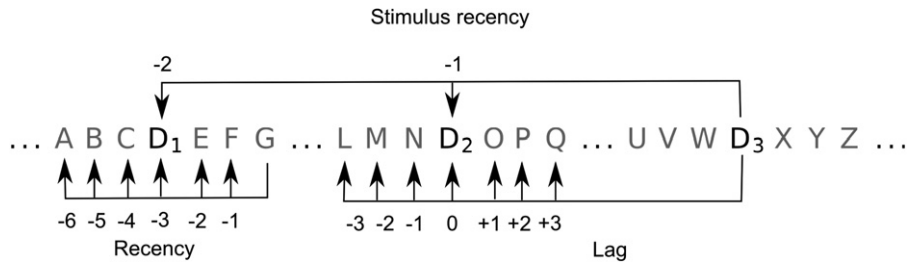*URL:* http://people.bu.edu/marc777/.

**Fig. 1.** Schematic for illustrating temporal structure. Models of memory can be distinguished by the similarity of the memory representations across three variables. First, do the states after different presentations of $D$ change over presentations such that the state at $D_3$ is more similar to the state at $D_2$ than to $D_1$? We define the stimulus recency between $D_3$ and $D_2$ to be $-1$; the stimulus recency between $D_3$ and $D_1$ is $-2$. Second, how does the state of memory vary across time around the presentation of a stimulus. That is, is the set of cells active after presentation of $G$ more similar to the representation after $F$ than it is to the representation after $E$? We define recency as the difference in serial position between two events. The recency between $G$ and $F$ is $-1$; the recency between $G$ and $E$ is $-2$. Third, how does repeating a stimulus affect the relationships in the memory representation? This can be assessed by comparing the memory representation after $D_3$ to the neighbors of a prior presentation of $D$, here $D_2$. We refer to this variable as lag. The lag between $D_3$ and $D_2$ is defined to be 0. The lag between $D_3$ and $P$ is $+2$; the lag between $D_3$ and $N$ is $-1$.

In the simple conditioning experiments primarily considered by SST, it was only necessary to consider one nominal stimulus. In subsequent years, memory researchers considered more elaborate verbal learning experiments in which many stimuli are experienced and the categorical distinction between stimulus and response is blurred. For instance, in a free recall experiment, the subject might be presented with a list of 20 words presented one at a time. After a delay, the subject's task is to recall the words from the list in the order they come to mind. The nominal stimuli in this experiment are the sequence of words. But the concept of the response is more ambiguous. Associations between strings of recalls (see, e.g., Bousfield, 1953; Kahana, 1996; Pollio, Kasschau, & DeNise, 1968) suggest that memory must include a network of evolving associations between many stimuli that double as their own response. These associations could be mediated by the functional stimuli caused by each nominal stimulus.

SST specified how the memory representation following a stimulus changes over time, but it did not specify how the relationships between memory representations following different stimuli change as a function of the structure of experience. We will see that subsequent mathematical models of memory distinguish themselves from each other largely by how they respond to this structure. We will review these models in Section 2, making explicit their concrete hypotheses about how memory representations change over time. If we could directly measure the similarity between memory representations at various times, these hypotheses could be directly evaluated.

It is now possible to directly measure the similarity between brain states at different times using a variety of methods. We will discuss three such techniques. Functional magnetic resonance imaging (fMRI) provides an estimate of the oxygenation of blood, believed to be a correlate of neural function, at the spatial scale of millimeters. The pattern of activation across many individual voxels at different points in time can be compared to one another. In human epilepsy patients, electrodes are often placed below the skull for clinical reasons. In many cases these electrodes are too large to record the activity of single neurons, but they can nonetheless record meaningful signals believed to be associated with aspects of cognition. When individual neurons cannot be resolved, oscillatory fluctuations in voltage can be recorded at different anatomical locations. Finally, it is possible to record from many individual neurons using extracellular recording techniques. While it is relatively rare to record at the level of resolution necessary to identify individual neurons in humans, these methods are routinely applied in animal preparations. Extracellular recording can be used to generate a vector of firing rate across neurons, either simultaneously measured or inferred from many single neurons recorded in identical experimental preparations. In each case these methods give rise to a distributed pattern of activity across voxels, or electrodes, or neurons. Each pattern of activity can be compared to the pattern of activity at another point in time; one can construct a scalar measure to characterize the similarity between states. The similarity can be aggregated as a function of behaviorally relevant variables and compared to predictions from mathematical models describing cognition (see, e.g., Kriegeskorte, Mur, & Bandettini, 2008). In Section 3, we review recent neuroscientific work that attempts to address empirical questions about the nature of memory representations raised by SST.

## 2. Dynamic memory representations in mathematical memory models

Prior to SST, many models of memory simply described the strength of direct atomic associations between stimuli and responses. Modern memory models construct a description of a memory representation that changes dynamically in response to stimuli. This representation can be quite abstract (as in the SIMPLE model Brown, Steyvers, & Hemmer, 2007) or considerably more concrete (as, for instance, in TODAM2 Murdock, 1997). In this section, we describe how the memory representations developed by various mathematical memory models evolve over experience with different stimuli and how these choices endow the models with power to explain various behavioral phenomena. Although these models are in all cases quantitatively implemented, we will not focus on their precise mathematical form, focusing instead on the qualitative changes in the memory representation caused by different kinds of experience. So, for instance, we will not focus on the difference between the context representation in the Mensink and Raaijmakers (1988) model of interference and the context representation in the Murdock (1997) TODAM2 model. Although these representations change over time according to different equations, they share the property that they change gradually over time and are independent of the stimuli presented.

Fig. 1 provides a schematic that enables us to illustrate three distinguishable types of temporal relationship. Let us denote the state of the memory representation when, say, stimulus $A$ is presented as $\mathbf{s}_A$. First, we can consider how the state changes across different presentations of a particular nominal stimulus. Consider the three occurrences of $D$ in Fig. 1. Indexing the three presentations by a subscript, we can ask whether these representations change gradually over time, or if they are independent of one another. That is, if the memory states are independent, then $\mathbf{s}_{D_1} \cdot \mathbf{s}_{D_2} = \mathbf{s}_{D_1} \cdot \mathbf{s}_{D_3}$. In contrast, if the memory representation after presentation of the nominal stimulus $D$ changes gradually over time, then we would expect that $\mathbf{s}_{D_1} \cdot \mathbf{s}_{D_2} > \mathbf{s}_{D_1} \cdot \mathbf{s}_{D_3}$. We refer to the variable describing the number of presentations of the same stimulus as *stimulus recency* (Fig. 1). For instance, the stimulus recency