From Universal Laws of Cognition to Specific Cognitive Models

Nick Chater,[a] Gordon D. A. Brown[b]

[a]*Department of Psychology, University College London*
[b]*Department of Psychology, University of Warwick*

Abstract

The remarkable successes of the physical sciences have been built on highly general quantitative laws, which serve as the basis for understanding an enormous variety of specific physical systems. How far is it possible to construct universal principles in the cognitive sciences, in terms of which specific aspects of perception, memory, or decision making might be modelled? Following Shepard (e.g., 1987), it is argued that some universal principles may be attainable in cognitive science. Here, 2 examples are proposed: the simplicity principle (which states that the cognitive system prefers patterns that provide simpler explanations of available data); and the scale-invariance principle, which states that many cognitive phenomena are independent of the scale of relevant underlying physical variables, such as time, space, luminance, or sound pressure. This article illustrates how principles may be combined to explain specific cognitive processes by using these principles to derive SIMPLE, a formal model of memory for serial order (Brown, Neath, & Chater, 2007), and briefly mentions some extensions to models of identification and categorization. This article also considers the scope and limitations of universal laws in cognitive science.

Keywords: Scale; Scale-invariance; Memory; Free recall; Serial recall; Kolmogorov complexity; Similarity; Generalization; Universal laws; Categorization

A central question in cognitive science is the extent to which mental phenomena are subject to law-like regularities such as those observed in many aspects of the physical world. In addition, how far might such regularities be general or, indeed, universal, in the same sense as, for example, the Newtonian principles that govern the movements of the planets, the behavior of falling bodies, and the properties of pendulums? The search for such laws has a long history. The German psychophysicists, for example, explored whether there were law-like dependencies in the process of sensory transduction (most notably leading to Weber's Law; and, more controversially, Fechner's Law). In the domain of learning, the programme of behaviorism took a strong universalist stance, aiming to formulate laws of learning (albeit typically in qualitative rather than quantitative form) concerning relationships between stimuli

Correspondence should be addressed to Nick Chater, Department of Psychology, University College London, Gower Street, London WC1E 6BT. E-mail: n.chater@ucl.ac.uk

and responses that were intended to underpin the acquisition of all forms of behavior, in any species. Although the scope of this program now appears to have been radically overambitious, it is nonetheless still true that theories of associative learning (e.g., Shanks, 1995) and their rivals (e.g., Gallistel & Gibbon, 2000) provide quantitative and, potentially, general accounts of many learning phenomena.

The cognitive revolution, however, can be viewed as largely focussing attention away from the attempt to build universal laws. By viewing the mind as a highly complex computational device, it becomes natural to think of cognitive science as a process of "reverse engineering"—or more specifically, "reverse computer science"—rather than following in the mould of physics. Computer science does not seem to be full of quantitative universal laws—instead, its focus is on representations and algorithms operating over those representations. Cognitive science has taken the same tack, exploring the multiple mental representations appropriate for perception, motor control, language, and common-sense reasoning, and considering computational frameworks, and specific processing algorithms in which calculations over these representations can be defined; thus, the projects of discovering how people recognize speech (e.g., Norris, McQueen, & Cutler, 2000), compute depth from stereo (e.g., Marr & Poggio, 1976), or control the motor system (e.g., Wolpert & Ghahramani, 2004). It is tempting to suspect that each such problem is *sui generis*; and that the mechanisms involved will have no more in common that the properties of the liver and the properties of the heart (e.g., J. A. Fodor, 1983). From this perspective, the quest for universal laws might seem inappropriate in the context of cognitive science, and as resulting from a false comparison with the physical sciences.

The correct locus for general principles in cognitive science, if any, might naturally be viewed as the computational architecture. Attempts to provide architectural frameworks that might provide a universal, or near-universal, structure into which specific cognitive theories may be implemented, have been central to the development of the field (e.g., production system architectures, e.g., Anderson, 1983; Newell, 1990; and various kinds of connectionist architectures, e.g., Rumelhart & McClelland, 1986; as well as symbolic–connectionist hybrids of various kinds, e.g., Smolensky & Legendre, 2006).

Roger Shepard (1957, 1958a, 1958b, 1962a, 1962b, 1964, 1965, 1966, 1980, 1982, 1987, 1994), in a remarkable sequence of publications, argued that, nonetheless, quantitative universal principles are possible in cognitive science—and indeed that such principles may serve as crucial building blocks for the construction of cognitive theories in specific domains. Two proposals have been particularly influential: Shepard's Universal Law of Generalization, and the *geodesic* account of mental transformations (e.g, Carlton & Shepard, 1990a, 1990b). These proposals seek to drive quantitative laws into the very heart of central cognitive phenomena—which might seem to be governed by endlessly capricious representations and algorithms. In this article, we focus on the Universal Law, which will be central to the later discussion in this article.

1. Candidate principles 1: Scale-invariance

One of Shepard's most distinctive contributions to psychological theory has been an emphasis on the importance of symmetries in cognitive processes (e.g., Carlton & Shepard, 1990a,

1990b; Farrell & Shepard, 1981; Shepard & Cooper, 1982; Shepard & Zare, 1983). One of the most basic symmetries, and one that plays a central in the physical sciences, concerns scale. Scaling laws reveal so-called self-similarities: patterns that repeat in time, space, or other dimensions. From planetary motion, to shockwaves and fluid flow, self-similarity is a ubiquitous feature of the physical world. Such patterns of scaling are so familiar that they may escape our notice. For example, as a pendulum is lengthened by a factor f, its dynamics are invariant, except that its period increases by a factor $\sqrt{f}$. Suppose that we record the pendulum's behavior; but we forget to record either spatial or temporal scale. Self-similarity implies that we can never recover this scale information. If our data are consistent with one combination of spatial and temporal scales, then they will be consistent with a combination in which the spatial scale is f times larger; and time is flowing $\sqrt{f}$ times more slowly. Thus, the motion of the pendulum is *scale-invariant*. Remarkably, it turns out that many subtle problems in physics can be solved by making scale-invariance assumptions alone (Barenblatt, 1996).[1]

A particularly notable feature of self-similar phenomena is that the relevant descriptive laws are *power laws* (i.e., laws of the form):

$$y = ax_1^{n_1} \ldots x_i^{n_i} \ldots x_m^{n_m} \tag{1}$$

where a, and $n_1, \ldots n_i, \ldots n_m$ are arbitrary constants. Power laws embody self-similarity— their structure is the same, whatever scale we consider; that is, merely by looking at the form of the data, it is impossible to discover the absolute scale (and hence the units of measurement) of the variables y, and $x_1, \ldots x_i, \ldots x_m$. Fig. 1 illustrates this point with a simple example $y = 1/x^3$. Note that the shape of the function is identical when we "zoom in" on it, in this case by a factor of two on the x axis, and by a factor $2^3 = 8$ on the y axis. Note that any other functional relations between variables are *not* scale-invariant. For example, exponential decay has a specific scale (the "half life"); and the Gaussian has a specific scale (i.e., standard deviation). Although scale-invariance is well-enough understood to be widely practically
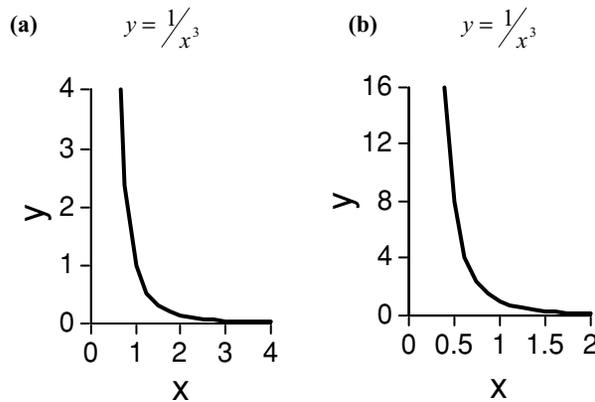

Fig. 1. Self-similarity of power laws. *Note*: Changes of scale on both x and y axes simultaneously leave the function invariant. Thus, the form of the function cannot provide information about absolute scale. Power laws exhibit scale-invariance.

applied throughout the physical sciences, nonetheless, its precise conceptual foundations are still a question for active research (e.g., Barenblatt, 1996).

Physics aside, scale-invariance applies across a wide range of social and economic (e.g., Gabaix, 1999; Ijiri & Simon, 1977; Mantegna & Stanley, 1995), biological (e.g., Gisiger, 2001; Goldberger et al., 2002; West & Brown, 2005), and cognitive phenomena. In cognitive science, scale-invariance applies to many of the best claims to be psychological laws (Table 1). Thus, the power laws of forgetting, the power law of practice, and Stevens's law (that perceptual inputs and judgments are related by a power law) are instances of scale-invariance. So, too, is Weber's Law, which states that the precision of the encoding of a stimulus is

Table 1
Scale-invariance and psychological "laws"

Regularity	Form	Comment	References
Weber's Law	$\Delta I \propto I$	$\Delta I/I =$ constant, if independent of units	Weber (1834/1996)
Stevens's Law	$I^\alpha \propto S$ (power law)	$\Delta I/I \propto \Delta S/S$ (i.e., increasing I by any factor leads to an increase in S by a, typically different, constant factor	Stevens (1957)
Power law of forgetting	$m(t) = m(0)t^{-\alpha}$	Ratio preserving: Memory–time	Wixted and Ebbesen (1997); although see, e.g., Chechile (2006)
Power law of practice	$RT(N) = a + b.N^{-c}$	Ratio preserving: Trials–speed	Newell and Rosenbloom (1981); although see Heathcote, Brown, and Mewhort (2000)
Time-scale invariance of animal learning	Slowing or speeding up learning schedules does not affect level of learning	Information (in bits) between stimuli and reward determines learning	Gallistel and Gibbon (2000); Balsam, Fairhurst, and Gallistel (2006)
Fitts's Law	$T = a + b.\log(\Delta D/D)$	*nb* non-invariant *a* for initiating movement	Fitts (1954)
Space–time invariance in motor skills	Spatial structure and kinematics of handwriting, reaching, and so on independent of time–space	Arising from principles of nonlinear dynamics/self-organization?	Viviani and Terzuolo (1980); Kelso (2000); Wing (2000)

Note. Scale-invariance provides an explanation for a wide range of psychological regularities, including many that aspire to the status of psychological laws. We suggest that scale-invariance should be expected, as a null hypothesis in cognitive science, as it is in the natural and social sciences. Indeed, any mental process in which no scale plays a privileged role will have scale-invariant behavior. Thus, we suggest that it is not appropriate for focus on piecemeal attempts to explain the particular form of these regularities—instead, we suggest that *departures* from scale-invariance are likely to be much more informative concerning the nature of cognition.
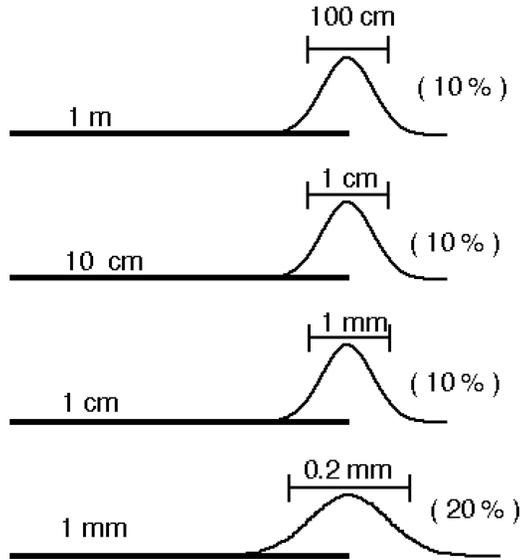
Fig. 2. Scale-invariance and Weber's Law. *Note*: If we assume that sensory magnitudes (here, judgments of distance) have scale-invariant sensory noise, then this implies that the ratio of magnitude to error will be constant. This is Weber's Law. At the limits of the sensory system (e.g., when magnitudes become extremely small), error will increase disproportionately. Thus, the departure from scale-invariance reveals the mechanistic properties of the sensory system. By contrast, where scale-invariance holds, this provides little diagnostic information concerning cognitive mechanisms (see Chater & Brown, 1999).

proportion to the magnitude of a stimulus. Weber's Law is an example of scale-invariance because the *ratio* of precision and absolute magnitude is constant—and hence this ratio is the same at all scales (Fig. 2). Weber's Law is, of course, violated for extremely stimulus values, where the sensory system is either overloaded, or at the limits of stimulus detection. In both cases, it is the departures from scale-invariance—cognitive *bumps*—that are revealing about underlying cognitive mechanisms.

There is also a wide range of phenomena that have a complex pattern that is invariant over changes of scale. For example, patterns of recall often appear invariant to changes in time scale, in a range of paradigms. Maylor, Chater, and Brown (2001) asked people to recall events that had happened in the last day, week, or year, and found that the speed at which they were able to retrieve these memories was the same for each condition. Of course, the type of event recalled at each time scale was very different (e.g., we might recall a holiday, in the "year" condition; or eating breakfast in the "day" condition). Had people not made this "significance" adjustment, participants could clearly have retrieved more items for the longer time periods, as any event recalled in the last day is *a fortiori* a memory from the last week and year. In fact, people adjusted what counts as significant in a way that maintained scale-invariance—the rate of retrieving items was independent of time scale (Fig. 3). As also shown in Fig. 3, the same effect occurred with prospective memories (i.e., memories for things that will happen in the future).

Another example concerns the shape of serial position curves in serial and free recall, which may exhibit scale-invariance with respect to time. Scale-invariance is consistent, in particular,
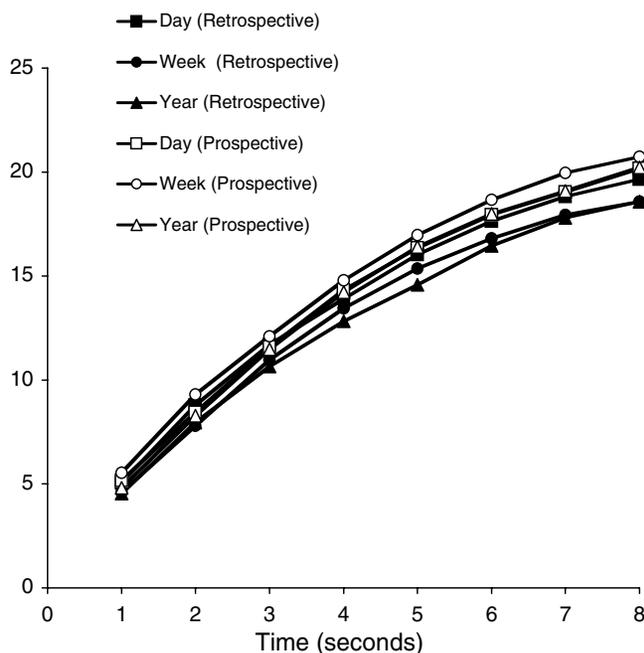
Fig. 3. Scale-invariance in recall over different time intervals. *Note*: Maylor, Chater, and Brown (2001) asked people to recalls events that had occurred (a) or were going to occur (b) within 1 day, week, or year. The number of items retrieved was invariant to the time interval probed. This implies that the "salience" of items considered worth reporting for each of these intervals scales with the size of that interval. In the absence of such scaling of salience, it would be natural to assume that more items should be retrieved for longer intervals because all items that occurred over a short interval necessarily also occurred during the longer interval. Data are re-plotted from Maylor et al.

with the much-discussed "ratio rule": that the slope of the recency curve is determined by the ratio of the rate at which items are presented and the time since the last item (e.g., Bjork & Whitten, 1974; Glenberg, Bradley, Kraus, & Renzaglia, 1983; but see Davelaar, Goshen-Gottstein, Ashkenazi, Haarmann, & Usher, 2005). Thus, if the entire schedule of learning and test is scaled multiplicatively (see Fig. 4), then the slope of the recency curve should be unchanged. Similar effects are seen in memory for temporal order. Neath and Brown (2006) noted that when the position of items separated by 50 msec in a list must be recalled, or the day of the week on which an event occurred must be remembered, similar serial position curves are observed over timescales that vary by six orders of magnitude.

Finally, note that scale-invariance is ubiquitous in many aspects of the perceptual and motor system. In the context of perception, scale-invariance is sometimes so natural that it is scarcely noticed—for example, object, text, or face recognition is roughly invariant to the retinal size of the stimulus; recognition of melodies is roughly invariant both to the volume and pitch, and so on. These invariance of scale are, of course, part of a wider set of symmetries (e.g., in the visual world, involving translations and rotations), which Shepard and colleagues have explored, both in perception and imagery (e.g., Farrell & Shepard, 1981; Shepard & Cooper, 1982). Similarly, scale-invariance is widespread in the motor system. For example, Schmidt,
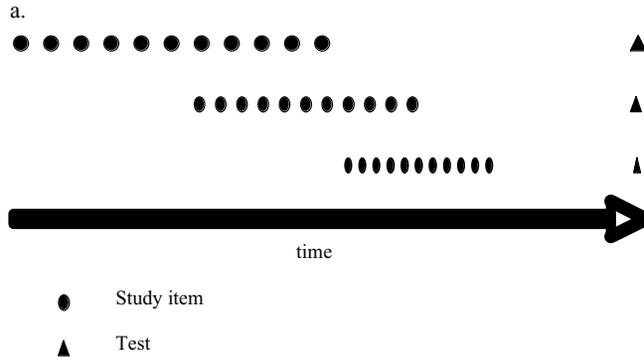
a.

Fig. 4. Scale-invariance, the ratio rule, and serial position. *Note*: Three learning schedules differ purely by temporal scale. Thus, the ratio between the intervals between items, and the interval from the final study item to test, is held constant. The ratio rule (Bjork & Whitten, 1974; Glenberg, Bradley, Kraus, & Renzaglia, 1983) states that the slope of the recency curve should depend only on this ratio (i.e., should be invariant across these cases, as is observed).

Zelaznik, Hawkins, Frank, and Quinn (1979) developed a motor analog of Weber's Law, showing that the variable of a forced produced scales linearly with the magnitude of that force; and both simple movements (e.g., pointing—Soechting & Lacquaniti, 1981) and extremely complex movements (e.g., handwriting—Viviani & Terzuolo, 1980) are scale-invariant in both space and time.

We have argued that scale-invariance applies across a wide range of cognitive phenomena. Indeed, we suggest that, unless there is reason to suppose that some aspect of the environment and agent depends on a specific scale (e.g., as in speech processing) scale-invariance may be expected. Thus, a wide variety of psychological regularities, many of which are viewed as requiring special theoretical explanation, may be viewed as coming from a common source, the symmetry induced by the lack of any special phenomena at any particular scale. We suggest, therefore, that it is departure from, rather than adherence to, scale-invariance that requires theoretical explanation.

Thus, for example, Gilden, Thornton, and Mallon (1995) asked people to tap each time they thought that a specific time interval had elapsed (in the range 0.3 sec–10 sec). They then calculated the temporal displacement from a "regular" rhythm from each successive interval. These displacements showed correlations across a wide range of time scales—and indeed, the power spectrum of revealed a power law, which is characteristic of scale-invariance (the exponent of this power law was roughly −1, a noise-structure often observed in physical data; Handel & Chung, 1993). Crucially, however, the power spectrum also shows a well-defined kind, at high temporal frequency, after which the power spectrum has the flat structure characteristic of white noise. Gilden et al. argued that this kink, indicating a violation of scale-invariance, is evidence for two cognitive processes underlying performance: a scale-invariance timing mechanism combined with high frequency "jitter" from the motor system, a viewpoint for which there is independent support (e.g., Wing, 1980).

Violations of scale-invariance may be informative at a qualitative level, too. Whereas the human auditory system generally processes acoustic signals in a similar way across a wide

range of temporal frequencies (i.e., reflecting scale-invariance), there may be characteristic psychophysical transitions around frequencies where the dominant coding mechanism for frequency changes (i.e., around 4–5 kHz, after which at which neural phase locking does not occur; Moore, 2003). Whereas music and ambient sounds typically sound qualitatively similar when shifted in frequency, speed sounds that are shifted in frequency typically sound utterly bizarre. This presumably arises because there is a characteristic band of frequencies in human speech, which is itself a function of the fact that the human vocal chambers have a characteristic size, rather than being scale-invariant (see Lewicki, 2002). The fact that scale-invariance in frequency is violated for human speech processing is, therefore, interesting and informative; the maintenance of scale-invariance for many other types of stimuli is, by contrast, merely the natural default. More broadly, we suggest that *violations* of scale-invariance are likely to be especially revealing of the underlying neural and cognitive mechanisms.

2. Candidate law 2: The simplicity principle

Many problems faced by the cognitive system can be viewed as a type of inductive inference—as finding patterns in data. The perceptual system finds patterns, providing information about the external world, from input to the sensory receptors. Acquiring a language involves finding the wide variety of levels of structure described by linguistic theory. Learning to classify items, on the basis of experience, involves deriving category structures from experience (Feldman, 2000; Tenenbaum, 1999).

Inductive inference problems of this form can be formulated in terms of Bayesian inference (e.g., Chater, Tenenbaum, & Yuille, 2006). Accordingly, the probability of each pattern or structure, after the data is received, is proportional to the product of the conditional probability of the data, given that structure (i.e., how well the structure "fits" the data); and the prior probability of that structure. A central issue in the Bayesian program is how these prior probabilities should be set.

One approach is to set the priors using simplicity—specifically, we assume that the probability of a model, grammar, or pattern is inversely proportion to its complexity. How can this intuition be made precise? The mathematical theory of Kolmogorov complexity (Li & Vitányi, 1997) provides a natural framework. The complexity of a formal object, x, is defined as the length of the shortest code in a universal programming language that generates x. This code length, the Kolmogorov complexity of x, is written $K(x)$. A variety of mathematical considerations[2] suggest that the most natural prior distribution over possible probabilistic models, M, of the data is proportional to $2^{-K(M)}$ (deterministic models are, of course, merely a special case in which all probabilities are 0 or 1).[3] This conclusion can be viewed as a mathematical version of Occam's razor—that, other things being equal, simple theories should be preferred.

Suppose we collect data D. From a probabilistic standpoint, a natural objective is to choose the most probable model; that is, the M that maximizes $\Pr(M|D)$. By Bayes's theorem, we know that this will be the M that maximizes:

$$\Pr(D|M)\Pr(M), \tag{2}$$

which, substituting the universal prior into Equation 2, gives:

$$\Pr(D|M)2^{-K(M)},\tag{3}$$

and, of course, the M that maximizes Equation 3, also maximizes:

$$-\log_2(\Pr(D|M)2^{-K(M)})\tag{4}$$

by the monotonicity of log. This M therefore also *minimizes*:

$$-\log_2(\Pr(D|M)2^{-K(M)}),$$
$$=-\log_2(\Pr(D|M))=K(M)\tag{5}$$

An elegant result from Kolmogorov complexity theory is that, for a computable probability distribution $\Pr$, $-\log_2\Pr(x)=K(x)$ up to a constant term independent of x, for almost all x (strictly, with probability of measure 1, if the x are drawn from $\Pr$—i.e., there are counterexamples, but they are rare).[4] An analogous result holds for conditional probabilities (i.e., with high probability):

$$-\log_2\Pr(D|M)=K(D|M)\tag{6}$$

where the *conditional* Kolmogorov complexity, $K(y|x)$, is the length of the shortest program that yields y, given x as input. Roughly, conditional complexity is small where there is a simple transformation that maps input to output—we shall see that this notion of simple transformation provides the basis for an interesting general notion of similarity, later, in our discussion of the universal law of generalization.

Combining Equations 5 and 6, we conclude that, with high probability, the models M with maximal, or near maximal, posterior probability will also be the M with minimal code length (i.e., in symbols):

$$\underset{M_1}{\arg\max}\,\Pr(M_1|D)=\underset{M_2}{\arg\min}\{K(M_2)+K(D|M_2)\}\tag{7}$$

This derivation, which is derived and presented rigorously by Vitányi and Li (2000), can be informally stated as follows: Models that have high *a priori* probability according to a Bayesian analysis using any computable prior probability will (with high probability) correspond to models that allow short two-part descriptions of the data, and vice versa. This result[5] provides a fundamental normative underpinning for the simplicity principle—that the cognitive system should prefer models that provide the shortest codes for the data, where such codes consist of (a) a specification of the model or hypothesis; and (b) an encoding of the data, given that model.

In the light of the close relation between Bayesian and simplicity-based reasoning, how should we conceive of the relation between them? Some theorists, both in statistics (Rissanen, 1987) and cognitive science (Leeuwenberg & Boselie, 1988) argue that simplicity should be viewed as basic—essentially because the Bayesian approach attempts to assign probabilities of regularities in the "real" world, and it may be doubted that such regularities are, at least outside fundamental physics, likely to be anything more than approximations. Others, in machine learning (e.g., Wallace & Freeman, 1987) and cognitive science (Chater, 1996; Mumford, 1996) view simplicity and Bayesian inference as equivalent—and view the choice of theoretical

framework as a matter of methodological convenience. A simplicity approach is particularly useful in situations where it is easier to form hypotheses about cognitive representations than to form hypotheses about appropriate probability distributions. For example, theories about the representation of formal languages provide a framework for coding grammars; and hence this framework will induce code-lengths for such grammars. But it is arguable less clear how directly to specify a probability distribution over grammars, and hence in this case, a simplicity-based perspective may be more useful. Conversely, specific background knowledge, or statistical information, is easier to incorporate in a probabilistic framework. Hence, a Bayesian framework may be more appropriate for ecological stimuli (e.g., natural textures or images), where underlying physical regularities are fairly well-understood (e.g., concerning the physical regularities, including optics, that generate natural images; Richards, 1988) and where statistical properties of the stimuli can be measured empirically (e.g., Field, 1987).[6]

This result provides reason to believe that the simplicity principle is a reasonable principle of inductive inference. Another normative justification concerns prediction, which arises in the slightly different setting in which a corpus of data accumulates at each time step. A fundamental result, which we call the prediction theorem (Solomonoff, 1978), shows that for any sequence of data produced by a computational process (which can be combined with randomness), prediction of the next item based on simplicity converges with high probability on the "true" probabilities. Specifically, the expected sum-squared error of the infinite number of predictions of the next item, as the sequence unfolds, is bounded by a *finite* sum, proportional to the Kolmogorov complexity of the data-generating process (see Li & Vitányi, 1997) that can be viewed as indicating the possibility of induction, under very general conditions.

How far, however, is the simplicity principle not merely normatively justified but a useful *description* of how the cognitive system finds structure? A direct test is difficult because detailed predictions from the principle in any domain appears to require having independent, and fairly detailed, knowledge of the relevant mental representations—because code lengths are defined in terms of these representations.[7]

Now, one of the attractions of the theoretical notion of Kolmogorov complexity is that it allows us to abstract away from details concerning the precise coding language (as long as the language is sufficiently powerful—i.e., as powerful as a universal programming language, which turns out to be a surprisingly low hurdle). Indeed, the celebrated *invariance* theorem (see Li & Vitányi, 1997) shows that, for any two coding languages, the difference between their code lengths, for all objects that can be coded at all, cannot exceed a constant. This result allows us to abstract away from coding languages, and make mathematical progress (just as computational complexity theory; Garey & Johnson, 1979), allows computer scientists to describe the time-complexity of algorithms, independent of the details of the machine on which the algorithms run). We shall see the usefulness of this level of abstraction, in considering Shepard's Universal Law, and psychological models to which it can be related, below.

From the point of view of providing detailed psychological predictions, however, the need for a theory of mental representation, in terms of which codes can be constructed, remains (just as, if we are interested in building a model of reaction times for a particular task, assumptions concerning the specific computational and neural machinery involved will be of crucial importance; computational complexity theory is not, of course, enough). Fortunately,

Table 2
The simplicity principle

Domain	Principle	References
Perceptual organization	Find grouping that minimizes cost	Attneave and Frost (1969); Koffka (1935); Leeuwenberg (1971)
Early vision	Efficient coding and transmission	Blakemore (1990); H. B. Barlow (2001)
Similarity	Similarity between representations measured by code length between them	Chater and Vitányi (2003); Hahn, Chater, and Richardson (2003)
Categorization	Categorize items to find shortest code (high-level perceptual organization)	Pothos and Chater (2002); Feldman (2000)
Memory storage	Shorter codes easier to store	Chater (1999)
Memory retrieval	Explain interference by cue → trace complexity; a rational foundation for distinctiveness models	Brown, Neath, and Chater (2007)
Language acquisition	Find grammar that best explains child's input	Chomsky (1955); Goldsmith (2001); Chater and Vitányi (2007)

Note. The view that the cognitive system prefers simple representations has been a recurring theme throughout the history of study of mental processes and ranges across cognitive processes, from perception to memory and language acquisition.

however, a great deal of work within cognitive science has been devoted to building theories of mental representation, in particular domains; for example, in the case of language processing, linguistics provides a rich set of potential levels of description, which have been enriched, modified and extended by work in psycholinguistics and computational linguistics. Similarly, psychological, neuroscientific and computational, work has provided a rich range of hypotheses concerning the representation of the perceptual world. Given a representational system, the simplicity principle predicts that the cognitive system will prefer to encode the available data using the structure or model that provides the shortest encoding of the available data. This program has been investigated by a range of formal and computational models, in language (e.g., Brent & Cartwright, 1996; Dowman, 2000; Goldsmith, 2001) and perception (e.g., Attneave & Frost, 1969; Bienenstock, Geman, & Potter, 1998; Hochberg & McAlister, 1953; Leeuwenberg, 1971).

Keeping the argument at a general level, however, still allows us to draw on number of lines of evidence that appear consonant with the simplicity viewpoint (Table 2; these have been reviewed elsewhere—see Chater, 1999, 2005; Chater & Vitányi, 2003). For example, a vast range of phenomena in perceptual organization, including the Gestalt laws of closure, good continuation, common fate, and so on, have been widely interpreted as revealing a preference for simplicity (see Fig. 5). Moreover, grouping effects that may be determined by simplicity also have wide ramifications for other aspects of perception (e.g., Adelson, 2000). Items with simple descriptions are typically easier to detect in noise and easier to detect (Garner, 1974). Finally, note that the physiology of early vision, including receptive field shapes, and phenomena such as lateral inhibition, seems adapted to maximize information compression in vision (Blakemore, 1990).
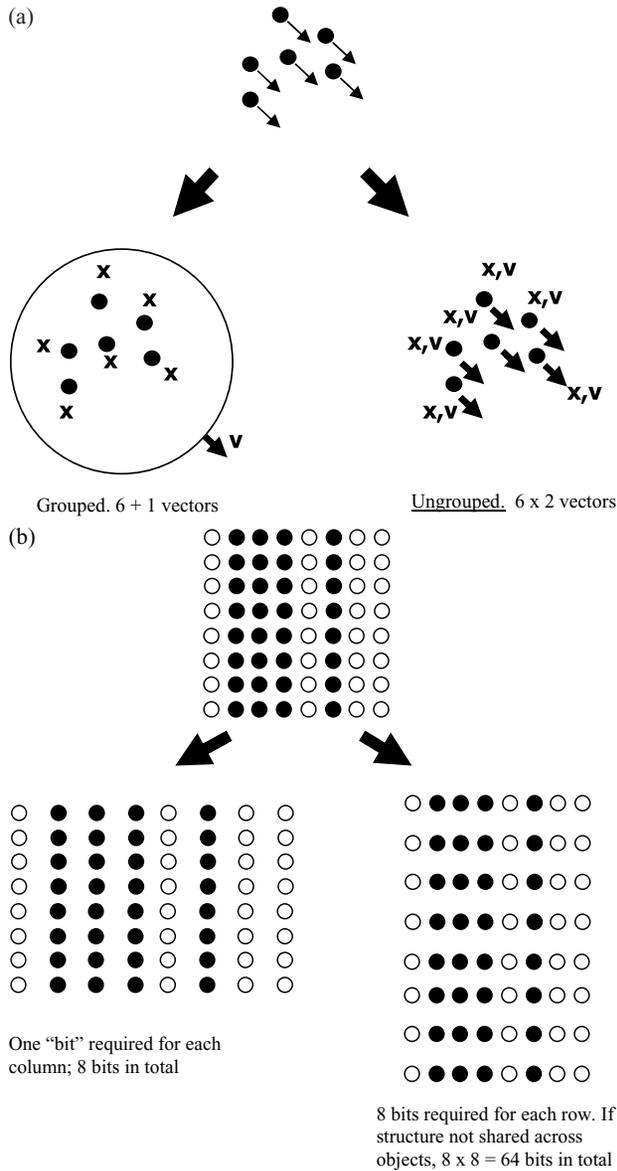
(a)

Grouped. 6 + 1 vectors

Ungrouped. 6 x 2 vectors

(b)

One "bit" required for each
column; 8 bits in total

8 bits required for each row. If
structure not shared across
objects, 8 x 8 = 64 bits in total

Fig. 5. Simplicity and Gestalt principles. *Note*: The Gestalt principles of perceptual organization can be explained in terms of the simplicity principle. (a) Grouping by similarity: We assume that objects are encoded separately; or, more modestly, that sharing information is more difficult between objects than it is within objects. This assumption is consistent with object-based views of attention (e.g., Duncan, 1984). If the top array is organized in terms of columns, color is easy to encode—each column is either black or white. Hence, coding color requires just 1 bit of information for each column. By contrast, if the array is organized by rows, each row has an irregular structure, which must be coded separately and will require a longer code length. Hence, according to the simplicity principle, the column organization is preferred; more generally, groups containing similar items can be encoded more briefly. (b) Common fate: Objects with the same motion, such as flocks of birds, tend to be grouped together. When the objects are grouped together, the stimulus can be coded by a vector indicating the location of each object and a single vector for the movement of each group. The ungrouped code would require a separate vector for the motion of each item.

Moving to higher level cognition, the simplicity of a code for a stimulus is related to the amount of structure uncovered in that stimulus. The more structure people can find in a stimulus, the easier they find it to process and remember and the less random it appears (Falk & Konold, 1997; see also Griffiths & Tenenbaum, 2003). The speed of learning for Boolean concepts (e.g., A or B or C; A and B or C, etc.) is well predicted by the shortest code length for those concepts (Feldman, 2000, 2006). Moreover, as we shall now see in the following, the simplicity principle provides a natural machinery for building a general theory of similarity (Chater & Vitányi, 2003; Hahn, Chater, & Richardson, 2003).

The simplicity principle relates closely to a number of other approaches to inference. As we noted in deriving Equation 7, the simplicity principle typically favors models with high Bayesian *a posteriori* probability. Indeed, the simplicity principle can be viewed as corresponding to Bayesian inference, using a particularly general "ignorance" prior—that is, one for which hypotheses, H, have a prior probability given by $2^{-K(H)}$, where $K(H)$ is the length of the shortest code for H.[8] The simplicity principle, as stated in terms of Kolmogorov complexity, is a highly idealized notion; practical statistical methods based on simplicity "scale-down" the approach to consider the shortest code length given restricted coding schemes, typically assuming, for example, that data is independently drawn of each trial from a single distribution; that it is generated from some particular class of probability distributions, and so on. Statistical perspectives include Rissanen's (1987) Minimum Description Length, Wallace and colleagues' Minimum Message Length (Wallace & Freeman, 1987), and Dawid's (1984) prequential approach to statistical inference.

Given these close relations between concepts based on coding and based on probability, a natural question is: Which should be viewed as basic? We suggest that, from a theoretical point of view, coding and probabilistic concepts may be viewed as equivalent; but that, in practice, the most workable framework should be selected. In many areas of cognition, we suggest, we can make reasonable assumptions about mental representations (especially in language and in early vision), and, indeed, much of cognitive psychology has been oriented toward questions of representation. In such cases, viewing coding and simplicity as basic, and probability as a derived notion, may be most appropriate.

3. The universal law of generalization: A derivation from simplicity

The Universal Law of Generalization (Shepard, 1987) considers the question of generalization from observation of one item, S_i, to a second item, S_j. In abstract terms, this can be viewed as a problem of induction, with just one prior specific instance (Heit, 2000)—that is, suppose that I learn that Daisy the cow has property X; how likely do I believe it is Bill the horse has property X; or that Herbie the car has property X. At first sight, it might seem that type of problem is entirely resistant to quantitative analysis (e.g., Bush & Mosteller, 1951; J. A. Fodor, 1983; Oaksford & Chater, 1998; Pylyshyn, 1987). Such problems appear to depend crucially on how different kinds of object are represented and the nature of the property X under consideration. Furthermore, the process of inductive inference can potentially depend on arbitrary amounts of relevant background knowledge.

The elegance of the Universal Law of Generalization is that it aims not to deny, but to cut through, such complexity. The proposal is that the probability of generalizing from one stimulus item, S_a, to another, S_b, is proportional to the negative exponential of the *psychological distance*, $D(a, b)$, between them. In symbols:

$$G(a, b) = Ae^{-BD(S_a, S_b)} \tag{8}$$

The elegance of the approach is that, although both quantities may be influenced by any number of specific representational or algorithmic constraints, the law states that they will be influenced in corresponding ways. Specifically, if, for whatever complex reasons, two items are nearby in psychological space, then generalization between them will be high; if they are not, generalization will be low.

For the Universal Law of Generalization to be practically useful, it is of course crucial to be able to independently measure the psychological distance between pairs of items. A key breakthrough in achieving this arises from Shepard's (1962a, 1962b) pioneering work on non-metric multidimensional scaling. This method takes as input an ordering of the distances between all pairs of items—that is, all that is required is that we can determine whether the psychological distance between, say, A and B, is greater or less than the distance between C and D; no "metric" assumptions are required concerning the size of these distances or the size of the difference between them. The output is a "map" with each point embedded in a (typically Euclidean) space, where each item is located in the space so that the ordering of the distances between points matches the ordering of distances that were the original data. Thus, this space now provides an estimate of the relative psychological distance between the items; and this measure of psychological distance can be used to make predictions concerning generalization.

The key ideas behind the simplicity principle can also be applied to provide a general picture of similarity and confusion between them. Intuitively, it is natural to suggest that the similarity between two representations is, to some degree, related to the complexity of the transformation required to turn each representation into the other. Thus, two nearby points in a geometric space can be viewed as similar, because a small shift in locations suffices to map one to the other; and a pair of items with highly overlapping features may be viewed as similar, because relatively few "flips" will be required to map one set of features onto the other. This general idea can, of course, apply to representations of any kind: We may view two sentences as similar if they can be related by a simple grammatical transformation; or two pictures of the same object may be viewed as similar if they can be related by a change of viewpoint or lighting.

These intuitions are the starting point for the Representational Distortion theory of similarity (Hahn et al., 2003), according to which the dissimilarity between a pair of mental representations x and y is a function of the complexity of the transformation between them, given a particular representational coding language. The *conditional* Kolmogorov complexity, $K(x|y)$, the length of the shortest code that transforms y into x, is the natural starting point for this kind of account, although a variety of specific formulations are possible (e.g., imposing symmetry or allowing asymmetry, applying various types of normalization, factoring apart structure from noise for each object, etc.). Just as the invariance theorem, described earlier, allows us to abstract away from specific coding languages, and just speak of the complexity of a particular an object, x—that is, $K(x)$—so the analogous result for conditional Kolmogorov

complexity, allows to us abstract away from specific coding languages, and speak of the complexity of a transformation between two objects, $K(x|y)$. Note that we do not have to assume a specific coding languages for transformations—rather we can merely appeal to the representation language in terms of which the objects x, y and so on are coded. As we noted earlier, although this abstraction is mathematically useful (as we shall see shortly), precise behavioral predictions concerning the similarities of individual stimuli will, of course, depend on the coding language used, just as we described earlier, when describing the simplicity principle. Of course, the same goes for feature-based (Tversky, 1977) or geometric theories of similarity (Shepard, 1980)—only when specific features, or dimensions, are specified for the set of objecs under consideration, can we make precise behavioral predictions, although general properties of the accounts can be analyzed in the absence of such detailed information.

Hahn et al. (2003) provided empirical evidence for the representational distortion account, in experiments in which people judge the similarities of three dimensional arrangements of blocks (items that, they argue, cannot readily be captured in either Euclidean or featural terms, and need a more general "structured" representation). The account of similarity is attractive because it provides a general notion of similarity that may be relevant to wide variety of representations used in cognitive science (e.g., J. A. Fodor & Pylyshyn, 1988; Russell & Norvig, 2003; Tenenbaum, Griffiths, & Niyogi, 2007); and also because it collapses into geometric and feature-based models of similarity, in special cases.

Here, however, we focus instead on more formal aspects of this approach. Conditional Kolmogorov complexity can be the basis for a "universal" measure of (dis)similarity, DU, the "information metric" (Bennett, Gács, Li, Vitányi, & Zurek, 1998), which can be defined as follows[9]

$$D_U(x, y) = \tfrac{1}{2}(K(y|x) + K(x|y)). \tag{9}$$

This metric is universal in the sense that, if any cognitively plausible distance measure treats two items as similar, the information metric also treats them as similar. Specifically, if we impose some mild restrictions on the notion of distance—call these admissible distances,[10] the informational distance is, in a sense, a minimal distance between any pair of items (for a discussion, see Chater & Vitányi, 2003; for technical details, see Bennett et al., 1998).

More formally, an admissible distance $D(x, y)$ is *universal* if, for every admissible distance $D'(x, y)$, we have:

$$D(x, y) \leq D'(x, y) + c_D, \tag{10}$$

where c_D is a constant that depends on D, but not on x and y. It turns out that, remarkably, that universal distances exist; and that $D_U(x,y)$ is, indeed, a universal distance. This implies that it assigns nearly as small a distance between two objects as any cognitive distance will do. Thus, for example, although the positive and negative image of the same picture are far away from each other in terms of Euclidean distance (because each pixel value is different), they are at almost zero distance in terms of universal distance because interchanging the black and white pixels transforms one picture into the other. The universal similarity metric is, like other notions rooted in Kolmogorov complexity, an "ideal" notion in the sense that it ignores the limitations on processing capacity, or the goals of the cognitive system.

D may therefore be viewed as a "default" distance metric—that is, if we know nothing about the particular similarity metric that the cognitive system uses, in some context, we know at least that it is not "too far from" *D*. Yet, as we noted earlier, *D* can also be viewed as providing a positive theory of similarity—a theory that states that cognitive distance is determined by the complexity of the transformations required to turn the representation of one object into the representation of the other, and vice versa. Rather than describe empirical evidence of this approach here (Hahn et al., 2003), we focus on how the application of simplicity to similarity can generate non-trivial psychological generalizations. Specifically, and especially relevant to the theme of this special issue, Shepard's contribution to cognitive science, we show that psychological distance between mental representations yields Shepard's Universal Law of Generalization (Chater & Vitányi, 2003).[11]

Although intended to have broader application, the Universal Law of Generalization is primarily associated with a specific experimental paradigm—the identification paradigm (for other derivations of the Universal Law, focusing on generalization rather than confusability, see Shepard, 1987; Tenenbaum & Griffiths, 2001). In this paradigm, humans or animals are repeatedly presented with stimuli concerning a (typically small) number of items. We denote items (e.g., phonemes, colors, or tones) as a, b, and x; the representations of the corresponding perceptual stimuli as S_a, S_b, and S_x; and the representation of the corresponding distinct responses as R_a, R_b, and R_x. In the identification paradigm, experimental participants are required to associate a specific, and distinct, response with each item—a response that can be viewed as 'identifying' the item concerned. The stimulus S_a is associated should evoke response R_a. In practice, this occurs only probabilistically—the more similar a and b are, the greater the chance that S_a might evoke R_b. We leave open, for now, the question of whether these responses arise from confusion of perception, or memory, or through deliberate generalization from one item to another.

Shepard (1987) uses a specific, symmetric measure, $G(a, b)$, to capture what he termed the *generalization* between items a and b:

$$G(a, b) = \left(\frac{\Pr(R_a|S_b) \Pr(R_b|S_a)}{\Pr(R_a|S_a) \Pr(R_b|S_b)} \right)^{1/2}. \tag{11}$$

Now we are in a position to directly relate $G(a, b)$ to information distance. Recall that, for a computable probability distribution Pr, with high probability, $-\log_2 \Pr(x) = K(x)$ up to a constant term independent of x:

$$\log_2 G(a, b) = 1/2 \{ \log_2 \Pr(R_a|S_b) + \log_2 \Pr(R_b|S_a) \\ - \log_2 \Pr(R_a|S_a) - \log_2 \Pr(R_b|S_b) \} \tag{12}$$

We assume mapping an item onto itself can be achieved by a fixed finite program, independent of the particular items a and b (this is essentially an identity mapping),[12] and hence the terms $-\log_2 \Pr(R_a|S_a)$ and $-\log_2 \Pr(R_b|S_b)$ can be collapsed into the $o(1)$ term. Replacing the $\log_2 \Pr(R_a|S_b)$ with $K(R_a|S_b)$ and $\log_2 \Pr(R_b|S_a)$ with $K(R_b|S_a)$, we obtain:

$$\log_2 G(a, b) = -1/2 \{ K(R_a|S_b) + K(R_b|S_a) \} \pm o(1), \tag{13}$$

where the $o(1)$ indicates that Equation 12 holds up to a constant. Assuming that there is a fixed program that maps the one-to-one correspondence between the S_x to R_x, this means that complexities will be invariant (up to a constant, and depending on the complexity of this mapping) if responses are replaced with stimuli, throughout (or, indeed, vice versa). This is typically a reasonable assumption—for example, in phoneme identification, the correspondence between the sound /b/ and the production of a /b/ sound by the learner can be taken as built in, before the experiment begins. In an identification paradigm, it is particularly natural because we assume that the participant has learned the labels of the stimulus items during a training phase. In cases where the mapping is complex and must be learned, collapsing the S_x to R_x would be inappropriate.[13] Making this substitution and applying Equation 9, we obtain:

$$\log_2 G(a, b) = -1/2\{K(S_a|S_b) + K(S_b|S_a)\} \pm o(1) \tag{14}$$
$$= -D_U(S_a, S_b) \pm o(1).$$

Rearrangement leads to:

$$G(a, b) = Ae^{-BD_U(S_a, S_b)}, \tag{15}$$

where A and B are arbitrary constants.[14] This is the Universal Law of Generalization, defined over representations of arbitrary form.

Two questions arise regarding the scope of this derivation. The first question is whether the derivation is too general. Does the Universal Law hold for data that are not best modeled by data derived using a Euclidean metric? Evidence on this question appears to be sparse, perhaps because scaling techniques that embed items in Euclidean spaces are particularly well-developed and widely used. One piece of evidence that the law may extend to other metrics is given in Cunningham and Shepard (1974).

Confusability data for Morse code signals collected by Rothkopf (1957) wereanalyzed by a very general scaling method, which makes only the metric assumptions. These data showed qualitatively the same pattern as in conventional non-metric multidimensional scaling analysis, consistent with the Universal Law. Moreover, Tenenbaum and colleagues (e.g., Tenenbaum et al., 2007) has used generalizations of multidimensional scaling, for example, modeling taxonomic hierarchies using tree structures, or causal dependencies in terms of directed causal graphs. These methods provide a rich set of tools for studying the scope of the Universal Law for broader classes of representation. The second question concerns whether this analysis can be adapted to deal with the case of deliberate generalization, rather than mere stimulus confusion, as considered here—and, indeed, an interesting question concerns the strength of the empirical basis for the universal law in this case (for a discussion, see Chater & Vitányi, 2003). In any case, however, the analysis for confusability will turn out to supply a robust empirical generalization; and, as we now see, one that can be combined with assumptions about scale-invariance to provide an account in a different area of psychological theory: memory for serial order.

4. Constructing models from principles: The SIMPLE model of memory for serial order

We noted at the outset that one of the attractions of general cognitive principles is that they provide a set of building blocks out of which models of specific cognitive phenomena may be constructed. Here, we provide a concrete example—indicating how a theory of memory for serial order, Scale Invariant Memory, Perception, and Learning (SIMPLE) can be constructed by assuming both scale-invariance and the simplicity-based generalization of Shepard's Universal Law (Brown, Neath, & Chater, 2007; Surprenant, Neath, & Brown, 2006). SIMPLE was originally developed from ideas of scale-invariance, and strongly influenced by Shepard's Universal Law; the precise derivation presented here, relating simplicity and scale-invariance together is, however, new. An interesting future direction will be to see if it is possible to successfully constrain new theories using general cognitive principles; here, however, we consider the derivation of existing theories.

Let us begin, by simplicity, considering to-be-recalled items as indexed along a single psychological dimension: the length of time that has elapsed since they were encoded. Consider an experiment in which participants are presented with a sequence of, say, five items. They are then probed with a particular location in the sequence (i.e., asked, "What was the fourth item in the sequence?"). Participants will give a range of responses, typically peaked, of course, around the correct answer (see Fig. 6). Let us assume that the items themselves are easily distinguishable, and hence that errors arise in confusion between the times at which those items are encoded. We assume that times are encoded independently for each item and that errors occur because of uncertainty about these times.

This setup can be viewed as an identification paradigm—in which the objects to be identified are locations in time, and their labels are the memory items, $I_1, I_2 \ldots I_5$ (as in Fig. 6). Thus, we can immediately apply a slightly variant of our generalization of Shepard's Universal Law, to describe the probability of confusion between items. Using the law in a simplified form (which follows directly from the derivation outlined earlier)

$$\Pr("I_i"|I_j) \propto e^{-aD(I_i|I_j)}. \tag{16}$$

Where $\Pr("I_i"|I_j)$ is the probability of mistakenly responding with item I_i, when the correct item in that position is item I_j. Now, by the assumption that the confusion between item arises because of confusion between their times (rather than the items themselves), the key question is how to measure the code length $D(t_i|t_j)$ where t_i and t_j represent how long ago items I_i, I_j occurred at the point of testing.[15] Notice that his assumption crucially depends on viewing time as an explicitly encoded psychological dimension—otherwise the question of *encoding* time would not arise. Thus, we face the question: how complex is it to code time t_i, given the "hint" of time t_j? By standard information theory, we can view this problem of coding in terms of probability—that is, how likely is an item at time t_i to occur, given that we know that an item at time t_j has occurred. Specifically, we can use the relation:

$$D(t_i|t_j) = -\log_2 \Pr(t_i|t_j)\Delta t_i, \tag{17}$$

where Δt_i is the precision of the encoding of t_j. Weber's Law (and scale-invariance) imply that $\Delta t_i \propto t_i$.
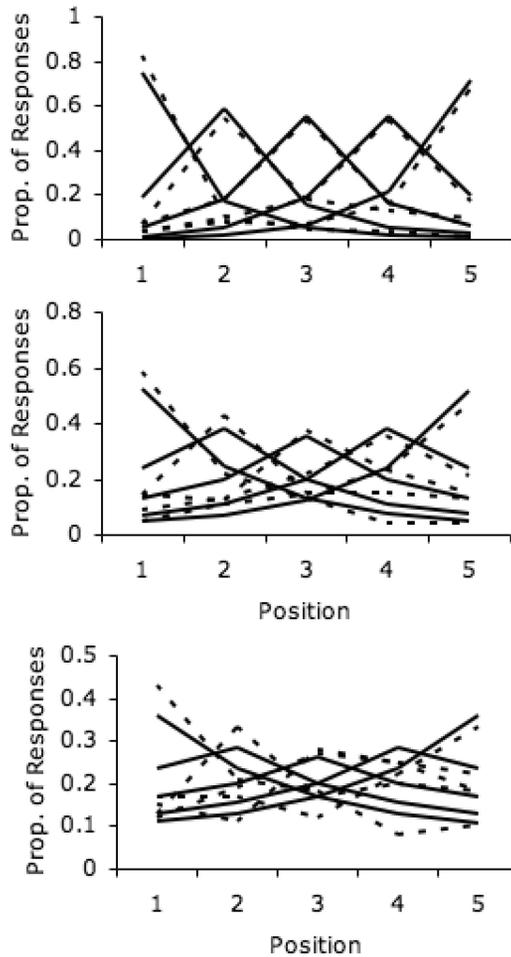
Fig. 6. Memory for serial order over three time scales. *Note*: Proportion of responses in a serial order reconstruction task. Each graph indicates the serial position of the items that participants produce, typically peaking on the correct serial position. The three panels indicate performance for delays before testing of 30 sec, 4 hr, and 24 hr. Notice that, although performance is clearly degraded after a longer retention interval, the pattern of results is qualitatively the same. The solid lines give fits from SIMPLE are shown alongside each set of data. Data are re-plotted from Nairne (1992).

Scale-invariance implies that the probability of encountering an item at t_i (within a precision Δt_i), given a prior input t_j, must be the same as the probability of encountering αt_i (within a precision $\alpha \Delta t_i$), given a prior input αt_j; that is, $\Pr(t_i|t_j)\Delta t_i$ is purely a function of the ratio between t_i and t_j. We assume, further, that this probability is symmetrical, so that only the absolute value of the ratio is important. Let us define the function $r(x,y)$ to be $\min(x/y,\ y/x)$:

$$\Pr(t_i|t_j)\Delta t_i = f(r(t_i, t_j)). \tag{18}$$

A further application of scale-invariance is to assume the function f itself is scale-invariant, and hence a power function:

$$f(r(t_i, t_j)) \propto (r(t_i, t_j))^b. \tag{19}$$

Connecting Equations 16 through 19 together, and simplifying, we can reason that:

$$\Pr("I_i"|I_j) \propto e^{-aD(I_i|I_j)} = e^{a \log_2 \Pr(t_i|t_j)\Delta t_i)},$$
$$\propto \left\{\Pr(t_i|t_j)\Delta t_i\right\}^c = \left\{f(r(t_i, t_j))\right\}^c = (r(t_i, t_j))^d \tag{20}$$

where a, b, c, d are arbitrary real constants. Hence, we can conclude that:

$$\Pr("I_i"|I_j) \propto (r(t_i, t_j))^d. \tag{21}$$

Given an input I_j, we will produce *some* output, so that $\sum_k \Pr("I_k"|I_j) = 1$. Hence, we can normalize Equation 21 to write the probability of responding I_i, when probed with the location of I_j, as:

$$\Pr("I_i"|I_j) = \frac{(r(t_i, t_j))^d}{\sum_k (r(t_k, t_j))^d}. \tag{22}$$

This equation is the core of a model of memory for serial order, SIMPLE (Brown et al., 2007). The scope of SIMPLE is broad—indeed, much broader than the derivation here would suggest. Indeed, SIMPLE provides a fairly comprehensive model of data on serial and free recall, and provides a single mechanism that explains many data that are typically viewed as arising from a variety of distinct memory stores.

Fig. 6 illustrates how SIMPLE provides excellent fits to data from a serial order reconstruction experiment (Nairne, 1992). Participants viewed lists of five items and rated them for pleasantness; at test participants were provided with the items required to arrange them in the order of presentation. The positional uncertainty gradients in Fig. 6 show the different output positions into which an item was placed after 30 sec (panel a), 4 hr (panel b), and 24 hr (panel c).

The model fit illustrated is with the model described in Equation 22, with the single parameter d allowed to vary with retention interval (for details, see Brown et al., 2007). We note that such scaling will be needed if performance is to be invariant with respect to the psychological spacing of items in temporal memory, as is (within limits) observed for the analogous case of absolute identification. In practice, perhaps due to the infeasibility of capturing all possible sources of memory interference, perfect scale-invariance is seldom observed in estimated model parameters. This is not surprising because changes of scale tend to lead to other changes (e.g., changes in the number of intervening items that may cause proactive or retroactive interference); and aspects of memory performance do have a characteristic scale (e.g., the rate at which items are rehearsed, or the rate at which they are reported at recall). Nonetheless, we suggest that, other things being equal, the more it is possible to minimize the effect of such factors (e.g., by using high distinctive items, with which intervening material does not interfere), the more closely observed behavior should fit with the predictions of scale-invariance.

We have illustrated how general principles can be used as a basis for constructing cognitive models in specific domains. We do not, of course, take this process to be merely algorithmic—

that is, there will be many assumptions and idealizations that may be adopted, that will utilize basic principles in different ways (just as in the physical sciences). Nonetheless, working with a set of basic principles substantially constraints the process of developing specific theories; we may hope that these constraints will assist the development of cognitive models, rather than unhelpfully restricting it. This derivation is an indication that, in some contexts at least, there is cause for optimism.

5. Extensions to identification and categorization

The derivation that we have outlined can be applied more broadly. We noted earlier that SIMPLE depends on the confusability of pairs of items, and we focussed purely on a single dimension, time. However, if time is not different (at least, in relevant respects) from any other psychological dimension the same derivation can be applied to confusability between items on any dimension (e.g., weight, brightness, or loudness). Neath, Brown, McCormack, Chater, and Freeman (2006) noted, indeed, that the application of SIMPLE successfully captures a wide range of phenomena in the task of absolute magnitude identification, where participants assign numerical labels to a fixed set of stimuli, with feedback (although, for a more precise model, see Stewart, Brown, & Chater, 2005).

If, by contrast, we consider more complex stimuli, distance is often not measured in physical terms (i.e., in terms of underlying physical parameters of sound pressure level, Newtons, and so on), but is left as a psychological primitive (and may, e.g., be measured by the application of multidimensional scaling; Shepard, 1962a, 1962b). If we rerun the derivation above, but leave the distance D unanalyzed, we obtain:

$$\Pr("I_i"|I_j) = \frac{e^{-D(t_i,t_j)}}{\sum_k e^{-D(t_k,t_j)}}, \tag{23}$$

which is a special case of the Nosofsky's (1986) model of identification (see also Luce, 1963; Shepard, 1957). Specifically, this special case lacks multiplicative "bias" weights, which can be used to capture response bias (i.e., an intrinsic tendency to choose some responses over others, independent of the stimulus). Note, too, that this formulation assumes, in line with Shepard's universal law, that generalization is an exponentially decreasing function of psychological distance. In some contexts, however, particularly when stimuli are readily confusable, a Gaussian, rather than exponential, function appears to provide a better fit with the empirical data (e.g., Nosofsky, 1986):

$$\Pr("I_i"|I_j) = \frac{e^{-D(t_i,t_j)^2}}{\sum_k e^{-D(t_k,t_j)^2}}. \tag{24}$$

When the Gaussian model is appropriate, scale-invariance is violated—and hence, according to our earlier arguments, we may expect that this reveals some interesting cognitive discontinuity. Shepard (1986) and Ennis (1988) provided arguments supporting this viewpoint—arguing that generalization follows the exponential universal law (following scale-invariance), except where the stimuli are difficult to discriminate, and hence where the accuracy

of the representation of the stimuli is the limiting factor on performance. Thus, perceptual noise introduces a scale when working at the limits of stimulus discriminability, just as motor noise introduces a scale, at very high frequencies in timing behavior (Gilden et al., 1995), as we discussed earlier. In both cases, scale-invariant behavior informatively breaks down at the performance limits of the system.

Nosofsky (1986) noted a natural extension of this model to categorization based on labelled examples. Specifically, if the identifying labels for the examples are not unique, but can be shared between exemplars (i.e., examples with the same label are viewed as belonging to the same category), then the probability of producing a particular category label, C, is the sum of the probabilities of the exemplars, i, with that label (the i such that $i \in C$. Specifically, this yields:

$$\Pr("I_C"|I_j) = \frac{\sum_{i \in C} e^{-D(t_i, t_j)}}{\sum_k e^{-D(t_k, t_j)}}. \tag{25}$$

This is Nosofsky's influential model of categorization, but, again, without free parameters to deal with the possibility of response bias.

6. General discussion

6.1. The scope of universal laws

We noted at the outset that cognitive science looks to computer science, and allied disciplines, as its source of theoretical hypotheses—and that computer science, unlike physics, is not replete with universal laws. Indeed, it might be tempting to conclude that understanding the detailed computational principles of the brain might be no more amenable to general laws than the computational principles underpinning a complex piece of software such as a word processor; statistical software package; a data-base; or, indeed, a model in traditional symbolic artificial intelligence. Such a piece of software is, of course, highly structured; the representations and methodologies it uses are carefully constrained; and there are many common features across different aspects of its behavior. But its behavior does not seem to exhibit quantitative universal laws.

It seems entirely possible, and indeed seems highly likely, that there are many aspects of cognition that must be understood in terms of specific representations and algorithms, which will not be neatly described by universal principles. But each individual case should, we suggest, be considered on its merits—and the possibility that general principles may combined to explain apparently complex phenomena should not be discounted. After all, the physical world, too, is full of extremely diverse and idiosyncratic objects and processes (e.g., stars, comets, tides, earthquakes)—yet many aspects of these can usefully be understood in terms of basic physical theory.

Language provides a particularly interesting example. Chomsky (e.g., 1965, 1980) influentially argued for a genetically encoded "universal grammar" that specifies information that is specific to language and does not derive from functional considerations. If this picture is correct, then universal grammar is the antithesis of universal principles as stated here: It focuses,

not on general cognitive principles, but on information specific to a particular (linguistic) domain. For example, binding constraints (Chomsky, 1981) provide elaborate, subtle, and apparently arbitrary restrictions on co-reference. For example, consider examples 26a through 26d), where the subscripts indicate co-reference, and asterisks indicate ungrammaticality:

John$_i$ likes himself$_i$ (26a)

*John$_i$ likes him$_j$ (26b)

John$_i$ said he$_{i/j}$ is happy (26c)

*He$_i$ said John$_i$ is happy (26d)

In 26a, the pronoun *himself* must refer to John; in 26b, it cannot. In 26c, the pronoun *he* may refer to John or to another person; in 26d, it cannot refer to John. Although apparently arbitrary, these constraints can, however, be explained in terms of pragmatic factors (Levinson, 1987)—for example, in 26b, the availability of the more specific *himself* to co-refer with John has the implicature that if the less specific *him* is used, then co-reference is not intended; in 26d, the use of *John* is unnecessarily specific, as a second pronoun would successfully co-refer to the first *he*; and alternative explanations have been govern in terms of processing preferences (O'Grady, 2005; for a more general account of grammatical patterns as arising from processing constraints, see Hawkins, 1990); and both pragmatic and processing biases may themselves become part of the grammar over generations of language change (e.g., through grammaticalization; Hopper & Traugott, 2003).

According to explanations of this type, it is possible that the linguistic patterns may arise from the confluence of a range of more basic cognitive principles, rather than being determined by genetically encoded language-specific principles. This raises the possibility that the interaction of relatively simple cognitive principles, of learning, processing, and pragmatics, might explain language structure, just as the interaction of basic physical principles is assumed to underpin the enormous variety and complexity of macroscopic physical phenomena. This style of explanation is consonant with much recent linguistic work, including construction grammar (e.g., Goldberg, 2006), and usage-based models of language more generally (e.g., M. Barlow & Kemmer, 2000; Tomasello, 2003). It is interesting to note that it may also be consistent with Chomsky's recent thinking. Hauser, Chomsky, and Fitch (2002) suggested that much of language may be explicable by general cognitive mechanisms; and even the exception they suggest, recursion, appears to be widespread across cognitive domains (e.g., in planning and motor control) as well as arising across species (Conway & Christiansen, 2001).

How far it proves to be possible to provide useful explanatory purchase on complex areas of cognition, including language, using an inventory of general cognitive principles is currently uncertain. Yet, we suggest that the apparent variety and complexity of cognitive domains and phenomena should not be taken as evidence against general principles, any more than this is appropriate in the physical sciences. Indeed, this very variety should, we suggest, be a stimulus to the search for underlying generality. If general principles cannot be found, the cognitive sciences are likely to be highly fragmented, piecemeal, and intractable.

6.2. Universal laws as reflections of the world

If we suppose that there are broad regularities in mental phenomena, what is their origin? One possible source of universality is common processing mechanisms, whether understood and an algorithmic or neural level. For example, common principles in memory might arise from a common neural substrate, in terms of many or all memories might be encoded. Alternatively, however, it might be that some universal aspects of cognition have a functional, rather a mechanistic, basis—they may arise as adaptations to the structure of the environment, and the agent's role within it. In an important sequence of articles, Anderson and colleagues have argued that many features of memory can be understood as adaptive in this way (Anderson & Milson, 1989; Anderson & Schooler, 2000; Schooler & Anderson, 1997). The core idea is that, in the environment, items occurs in a highly clustered way. For example, if a particular word or object has occurred at one time, the probability that it will occur at later times is not constant, but decreases systematically with time. Specifically, this decrease with time is modelled be a power law—this is, of course, a further example of scale-invariance. Anderson and colleagues (Anderson & Milson, 1989; Anderson & Schooler, 2000; Schooler & Anderson, 1997) made the reasonable assumption that an adaptive memory system should make items available in proportion to their likely occurrence. This leads to the prediction that forgetting should also have a power law, and thus scale-invariant, structure.

How can the adaptive and mechanistic viewpoints be distinguished? Perhaps the most straightforward line of argument concerns whether the regularity of interest arises across a range of different mechanistic systems—for example, across a wide range of stimulus materials (presumably engaging different neural machinery) or across species. To the extent that very different underlying machinery leads to the same regularity, it is reasonable to suspect that the regularity has some adaptive basis. Thus, for example, temporal and spatial scale-invariance in the motor system seems to apply across a wide range of different effectors and tasks, ranging from handwriting, to gross body movements (e.g., Viviani & Terzuolo, 1980). A second line of argument concerns whether the cognitive regularity follows "bumps" in environmental structure. For example, human perception is roughly invariant to pitch (e.g., the same tune can readily be identified, even if shifted through several octaves); but speech perception is not invariant in this way. Thus, we have the familiar observation that music played at the wrong speed sounds only mildly odd until singing begins. From an adaptive point of view, this makes good sense—because the human vocal tract, and hence the sounds that it produces, have a particular scale, to which the auditory and speech processing system is attuned. Anderson and Schooler (1991) used this type of argument in supporting their rational analysis of memory retrieval based on the repetition structure in the environmental. Suppose that a particular item is presented ten times over a period of time; or in rapid succession. On the latter account, a natural assumption of an adaptive memory system adjusts its assessment of how likely an item is to recur, based on the spacing of occurrences in learning, and the spacing between the last such item, and the test item. Thus, if we fix the gap between the last study item and test to be, say, 20 trials, then memory should be better if previous items were spaced at roughly 20 trial intervals, rather than spaced much more densely. Thus, we should expect that forgetting should be slower, if the prior items are further back in time—a phenomenon that appears counterintuitive, in the light of the natural assumption that memory traces are

active in proportion to their recency. Nonetheless, such "spacing" effects are observed (e.g., Bahrick, 1979; Glenberg, 1976), providing evidence for an adaptive interpretation of the scaling properties of memory.

An important open question concerns which aspects of the scale-invariance in cognition should be viewed as reflections of scale-invariant environmental structure and which should be viewed as arising from cognitive mechanisms. We suggest that this question should be addressed piecemeal for each scale-invariant aspect of cognition—that is, we do not propose that there is a single "unified" origin for scale-invariance in cognition. Scale-invariance is, after all, a null hypothesis in the sense that it proposes the *absence* of effects that are specific to particular scales. Indeed, rather than attempting to provide a "deep" explanation of this null case, it may, in general, be more appropriate to focus attention on explaining departures from scale-invariance—what might term "cognitive bumps"—and whether these appear to be best explained in mechanistic or functional terms.

Notes

1. For example, Barenblatt (1996) described the example of how Taylor (1950a, 1950b) was able to derive that the radius of the fireball of an atomic explosion is proportional to $E^{1/5}t^{2/5}\rho^{-1/5}$, where E is the energy released, t the time elapsed since detonation, and ρ the initial air pressure, purely from considerations of scale-invariance. This analysis agreed astonishingly well with experimental data.

2. This is an example of the so-called universal prior, that the prior probability of any object x is $2^{-K(x)}$. One justification of this prior is that it is, in a specific sense, as neutral as possible—that is, if any computable prior gives some x a certain probability, the universal prior gives x "nearly as much" probability (i.e., for any computable prior Pr, there is a constant m such that for all x, $m2^{-K(x)} = \text{Pr}(x)$). For a discussion, see Li and Vitányi (1997).

3. The class of possible models or hypotheses is restricted only by the requirement that the probability of the data must be computable (strictly, semicomputable from below, see Li & Vitányi, 1997). This is a very mild restriction—and it seems reasonable to assume models with uncomputable predictions would be of limited practical utility to the cognitive system.

4. Intuitively, this result holds because $-\log_2 \text{Pr}(x)$ is the code length for x, given Pr, according to standard information theory; and the Kolmogorov complexity, as the shortest code, must be at least this short. It is possible, but unlikely, that the converse is violated—that is, that Pr can generate objects, x, which have much shorter description than the $-\log_2 \text{Pr}(x)$ from standard information theory. For example, suppose we consider a uniform distribution on strings of length n, so that for each x, $\text{Pr}(x) = 2^{-n}$. Standard information theory will assign each x a code length of $-\log_2 \text{Pr}(x) = \log_2(2^{-n}) = n$ bits of information. Some strings (e.g., $000\ldots000$) will have shorter codes (e.g., using a simple looping program); but there will be few of these, by a counting argument—that is, although there are 2^n strings of length n, there can only be at most 2^k codes of length k, where $k < n$, and hence at most a fraction $1/2^{n-k}$ strings of

length n that have codes as short as k. More generally, for any computable Pr, a similar argument shows that the probability of generating an item with a shorter code is small. Thus, $K(x)$ approximates to Pr(x); and the argument extends to the conditional case, so that K($y|x$) approximates Pr($y|x$), up to a constant term, with high probability.

5. This result suggests that induction is possible in principle, from the available data, given the very modest restriction that the data is generated by computable or probabilistic processes. A restriction, however, is that the simplicity principle cannot be implemented accurately in practice because calculating Kolmogorov complexity is itself, in general, uncomputable. An interesting research question is far how the theoretical results we have outlined apply to practical approximations to the simplicity principle. The successful application of minimum description length, and related, methods in statistics and machine learning indicate that such approximations are often reliable in practice (Grünwald, Myung, & Pitt, 2005; Rissanen, 1987).

6. The divide between cases in which simplicity or probability is most usefully viewed as basic may, however, be less than straightforward. For example, Chandler and Field (2007) argued that, despite the intense study of the underlying physics and statistical structure of natural images, the best estimates of the information content of natural images comes from directly applying a coding method.

7. Kolmogorov complex is, notably, invariant up to a constant, between any two universal programming languages; and for these reason it is possible to develop an abstract theory of Kolmogorov complexity in general; and to prove asymptotic results such as the prediction theorem. Nonetheless, in deriving predictions concerning specific, finite, sets of sensory, linguistic, or other data, the "constant" that can be ignored in theoretical analysis, may turn out to be important. Thus, it may be important to have some prior knowledge of the representations in terms of which the cognitive system operates. Theories of representation are, however, highly developed in some areas of cognitive science; for example, linguistic theory provides a potential starting point for a theory of representations of linguistic knowledge.

8. It is interesting to note that scale-invariance can also be viewed as providing an ignorance prior (albeit an improper prior—i.e., a prior probability that sums to greater than 1), over continuous quantities. Kolmogorov complexity, by contrast, is most readily defined over discrete objects. Whether scale-invariance may usefully be viewed as derivable from simplicity, or whether should be viewed as separate, is an interesting question for future research.

9. There are various variants of information distance of which the formally most elegant is $D_{max}(x,y) = \max \{K(y \mid x), K(x|y)\}$.

10. We say that a distance is normalized if, for every binary string x, $\sum_{y:y\neq x} 2^{-D(x,y)} < 1$, where y ranges over all other binary strings. If the constraint is violated by some cognitive metric, the constraint requires simply that all distances are scaled-up (e.g., by a multiplicative factor) until the constraint does hold. An admissible distance $D(x, y)$ is a total nonnegative function on the pairs x; y of binary strings that is 0 if and only if $x = y$; is symmetric, satisfies the triangle inequality, is upper-semicomputable (roughly, can be approximated from above arbitrarily closely, by a computational process) is normalized, and is a metric (in the standard mathematical sense).

11. Note that this derivation assumes that cognitive distance is well-approximated by the average conditional Kolmogorov complexity required to transform each item into the other (in line with the Representational Distortion theory of similarity; Hahn, Chater & Richardson, 2003). Thus, if, in some context, cognitive distance (e.g., of two locations on a computer screen) were closely related to physical distance, then the derivation would not hold—and, indeed, we should not predict Shepard's (1987) Universal Law to hold good.

12. Note that this is a fairly innocuous assumption: We simply assume that the participant knows the relation, in principle, between the inputs and the outputs; that is, if the confusion were between phonemes, we begin by assuming that participants already know the relevant inventory of phonemes, and hence can provide the appropriate response in the task under ideal conditions. Errors in the task are presumed to arise when performance is difficult (e.g., under noise).

13. This case does not arise in the identification paradigm. It does, however, arise in paired associate learning. Here, confusion errors are a function *both* of the similarity of the "stimulus" items and the "response" items; and also, potentially, similarity of the relation between them. For example, if many of the pairs were synonyms and antonyms, this might substantially increase the probability of saying *white* in response to *black*. This type of case can, nonetheless, be analyzed given the machinery described here.

14. It might seem that the constant *B* is not arbitrary, but has a fixed value; but in reality it is a free parameter because Kolmogorov complexities themselves can be rescaled by a linear multiplicative factor. This corresponds to changing the number of primitive symbols in the assumed coding language.

15. We assume that there is no absolute "clock," but that time in the past is estimated by reference to the present.

Acknowledgments

References

Adelson, E. H. (2000). Lightness perception and lightness illusions. In M. S. Gazzaniga (Ed.), *The new cognitive neurosciences* (2nd ed., pp. 339–351). Cambridge, MA: MIT Press.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review, 96*, 703–719.

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

Anderson, J. R., & Schooler, L. J. (2000). The adaptive nature of memory. In E. Tulving & F. I. M. Craik (Eds.), *Handbook of memory* (pp. 557–570). New York: Oxford University Press.

Attneave, F., & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception and Psychophysics, 6*, 391–396.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308.

Balsam, P. D., Fairhurst, S., & Gallistel, C. R. (2006). Pavlovian contingencies and temporal information. *Journal of Experimental Psychology: Animal Behavior Processes, 32,* 284–294.

Barenblatt, G. I. (1996). *Scaling, self-similarity, and intermediate asymptotics*. Cambridge, England: Cambridge University Press.

Barlow, H. B. (2001). The exploitation of regularities in the environment by the brain. *Behavioral and Brain Sciences, 24*, 602–607.

Barlow, M., & Kemmer, S. (Eds.). (2000). *Usage-based models of language*. Stanford, CA: CSLI Press.

Bennett, C. H., Gács, P., Li, M., Vitányi, P., & Zurek, W. (1998). Information distance. *IEEE Transactions on Information Theory, 44*, 1407–1423.

Bienenstock, E., Geman, S., & Potter, D. (1998). Compositionality, MDL priors, and object recognition. In M. C. Mozer, M. I. Jordan, & T. Petsche (Eds.), *Advances in neural information processing systems 9* (pp. 838–844). Cambridge, MA: MIT Press.

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*, 173–189.

Blakemore, C. (Ed.). (1990). *Vision: Coding and efficiency*. Cambridge, England: Cambridge University Press.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61,* 93–126.

Brown, G. D. A., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review, 114,* 539–576.

Bush, R. R., & Mosteller, F. (1951). A model for stimulus generalization and. discrimination. *Psychological Review, 58*, 413–423.

Carlton, E. H., & Shepard, R. N. (1990a). Psychologically simple motions as geodesic paths: I. Asymmetric objects. *Journal of Mathematical Psychology, 34*, 127–188.

Carlton, E. H., & Shepard, R. N. (1990b). Psychologically simple motions as geodesic paths. II. Symmetric objects. *Journal of Mathematical Psychology, 34*, 189–228.

Chandler, D. M., & Field, D. J. (2007). Estimates of the information content and dimensionality of natural scenes from proximity distributions. *Journal of the Optical Society of America A, 24*, 922–941.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103*, 566–581.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology, 52A*, 273–302.

Chater, N. (2005). A minimum description length principle for perception. In P. Grünwald, I. Myung, & M. Pitt (Eds.) *Advances in minimum description length: Theory and applications* (pp. 385–410). Cambridge, MA: MIT Press.

Chater, N., & Brown, G. D. A. (1999). Scale invariance as a unifying psychological principle. *Cognition, 69,* B17–B24.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10,* 292–293.

Chater, N., & Vitányi, P. M. B. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology, 47*, 346–369.

Chater, N., & Vitányi, P. M. B. (2007). "Ideal learning" of natural language: Positive results about learning from positive evidence. *Journal of Mathematical Psychology, 51,* 135–163.

Chechile, R. A. (2006). Memory hazard functions: A vehicle for theory development and test. *Psychological Review, 113,* 31–56.

Chomsky, N. (1955). *The logical structure of linguistic theory*. New York: Plenum.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Chomsky, N. (1980). Rules and representations. *Behavioral and Brain Sciences, 3*, 1–61.

Chomsky, N. (1981). *Lectures on government and binding*. Dordrecht, The Netherlands: Foris Publications.

Conway, C., & Christiansen, M. (2001). Sequential learning in non-human primates. *Trends in Cognitive Science, 5*, 539–546.

Cunningham, J. N., & Shepard, R. N. (1974). Monotone mapping of similarities into a general metric space. *Journal of Mathematical Psychology, 11*, 335–363.

Davelaar, E. J., Goshen-Gottstein, Y., Ashkenazi, A., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: empirical and computational investigations of recency effects. *Psychological Review, 112,* 3–42.

Dawid, A. P. (1984). Present position and potential developments: Some personal views. Statistical theory: The prequential approach (with Discussion). *Journal of the Royal Statistical Society A, 147*, 278–292.

Dowman, M. (2000). Addressing the learnability of verb subcategorizations with Bayesian inference. In L. R. Gleitman & A. K. Joshi (Eds.), *Proceedings of the 22nd annual conference of the Cognitive Science Society* (pp. 107–112). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General, 114*, 501–517.

Ennis, D. M. (1988). Confusable and discriminable stimuli: Comment on Nosofsky (1986) and Shepard (1986). *Journal of Experimental Psychology: General, 117*, 408–411.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a bias for judgment. *Psychological Review, 104*, 301–318.

Farrell, J. E., & Shepard R. N. (1981). Shape, orientation, and apparent rotational motion. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 477–486.

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Feldman, J. (2006). An algebra of human concept learning. *Journal of Mathematical Psychology, 50*, 339–368.

Field D. J. (1987). Relations between the statistics of natural images and the response profiles of cortical cells. *Journal of the Optical Society of America A, 4*, 2379–2394.

Fitts, P. M. (1954). The information capacity of the human motor system in controlling the amplitude of movement. *Journal of Experimental Psychology, 47*, 381–391.

Fodor, J. A. (1983). *Modularity of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture. *Cognition, 28*, 3–72.

Gabaix, X. (1999). Zipf's Law and the growth of cities. *American Economic Review, 89*, 129–132.

Gallistel, C. R., & Gibbon, J. (2000). Time, rate and conditioning. *Psychological Review, 107*, 289–344.

Garey, M., & Johnson, D. (1979). *Computers and intractability; A guide to the theory of NP-completeness*. New York: Freeman.

Garner, W. (1974). *The processing of information and structure*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Gilden, D. L., Thornton, T., & Mallon, M. W. (1995). 1/f noise in human cognition. *Science, 267,* 1837–1839.

Gisiger, T. (2001). Scale invariance in biology: Coincidence or footprint of a universal mechanism? *Biological Reviews of the Cambridge Philosophical Society, 76*, 161–209.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15,* 1–16.

Glenberg, A. M., Bradley, M. M., Kraus, T. A., & Renzaglia, G. J. (1983). Studies of the long-term recency effect: Support for a contextually guided retrieval hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 231–255.

Goldberg, A. (2006). *Constructions at work*. Oxford, England: Oxford University Press.

Goldberger, A. L., Amaral, A. N., Hausdorff, J. M., Ivanov, P., Peng, C.-K., & Stanley, H. E. (2002). Fractal dynamics in physiology: Alterations with disease and aging. *Proceedings of the National Academy of Science, 99*(Suppl. 1), 2466–2472.

Goldsmith, J. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27,* 153–198.

Griffiths, T. L., & Tenenbaum, J. B. (2003). Probability, algorithmic complexity, and subjective randomness. *Proceedings of the 25th annual conference of the Cognitive Science Society* (pp. 480–485).

Grünwald, P., Myung, I. J., & Pitt, M. (Eds.). (2005). *Advances in minimum description length: Theory and applications*. Cambridge, MA: MIT Press.

Hahn, U., Chater, N., & Richardson, L. B. C. (2003). Similarity as transformation. *Cognition, 87*, 1–32.

Handel, P. H., & Chung, A. L. (Eds.). (1993). *Noise in physical systems and 1/f fluctuations*. New York: American Institute of Physics.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: What is it, who has it, and how did it evolve? *Science, 298*, 1569–1579.

Hawkins, J. A. (1990). A parsing theory of word order universals. *Linguistic Inquiry, 21*, 223–261.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin & Review, 7*, 185–207.

Heit, E. (2000). Properties of inductive reasoning. *Psychonomic Bulletin & Review, 7*, 569–592.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figure "goodness." *Journal of Experimental Psychology, 46*, 361–364.

Hopper, P. J., & Traugott, E. C. (2003). *Grammaticalization* (2nd ed.). Cambridge, England: Cambridge University Press.

Ijiri, Y., & Simon, H. A. (1977). *Skew distributions and the size of business firms*. Amsterdam: North-Holland.

Kelso, J. S. (2000). *Dynamic patterns: The self-organization of brain and behavior*. Cambridge, MA: MIT Press.

Koffka, K. (1935). *Principles of Gestalt psychology*. London: Routledge & Kegan Paul.

Leeuwenberg, E. L. J. (1971). A perceptual coding language for perceptual and auditory patterns. *American Journal of Psychology, 84*, 307–349.

Leeuwenberg, E. L. J., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review, 95*, 485–491.

Levinson, S. (1987). Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics, 23*, 379–434.

Lewicki, M. S. (2002). Efficient coding of natural sounds. *Nature Neuroscience, 5*, 356–363.

Li, M., & Vitányi, P. (1997). *An introduction to Kolmogorov complexity theory and its applications* (2nd ed.). Berlin: Springer.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (pp. 103–189). New York: Wiley.

Mantegna, R. N., & Stanley, H. E. (1995). Scaling behaviour in the dynamics of an economic index, *Nature, 376*, 46–49.

Marr, D. C., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science, 194,* 283–287.

Maylor, E. A., Chater, N., & Brown, G. D. A. (2001). Scale invariance in the retrieval of retrospective and prospective memories. *Psychonomic Bulletin & Review, 8*, 162–167.

Moore, B. C. J. (2003). *An introduction to the psychology of hearing* (5th ed.). San Diego, CA: Academic.

Mumford, D. (1996). *Pattern theory: A unifying perspective*. In D. C. Knill & W. Richards (Eds.). *Perception as Bayesian inference* (pp. 25–62), Cambridge, England: Cambridge University Press.

Nairne, J. S. (1992). The loss of positional certainty in long-term memory. *Psychological Science, 3,* 199–202.

Neath, I., & Brown, G. D. A. (2006). Further applications of a local distinctiveness model of memory. *Psychology of Learning and Motivation, 46,* 201–243.

Neath, I., Brown, G. D. A., McCormack, T., Chater, N., & Freeman, R. (2006). Distinctiveness models of memory and absolute identification: Evidence for local, not global, effects. *Quarterly Journal of Experimental Psychology, 59,* 121–135.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition.* (pp. 1–51). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Norris, D., McQueen, J. M., & Cutler, A. (2000). Merging information in speech recognition: Feedback is never necessary. *Behavioral and Brain Sciences, 23*, 299–370.

Nosofsky, R. M. (1986), Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.

Oaksford, M., & Chater, N. (1998). *Rationality in an uncertain world*. Hove, England: Psychology Press.

O'Grady, W. (2005). *Syntactic carpentry: An emergentist approach to syntax*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Pothos, E., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303–343.

Pylyshyn, Z. W. (Ed.). (1987). *The robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Richards, W. (Ed.). (1988). *Natural computation*. Cambridge, MA: MIT Press.

Rissanen, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society, Series B, 49*, 223–239.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology, 53*, 94–101.

Rumelhart, D. E., McClelland, J. L., & the PDP Research Group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition, Volumes 1 and 2*. Cambridge, MA: MIT Press.

Russell, S. J., & Norvig, P. (2003). *Artificial intelligence: A modern approach* (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

Schmidt, R.A., Zelaznik, H., Hawkins, B., Frank, J. S., & Quinn, J. T. J. (1979). Motor-output variability: A theory for the accuracy of rapid motor acts. *Psychological Review, 47*, 415–451.

Schooler, L. J., & Anderson, J. R. (1997). The role of process in the rational analysis of memory. *Cognitive Psychology, 32,* 219–250.

Shanks, D. R. (1995). *The psychology of associative learning*. Cambridge, England: Cambridge University Press.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22*, 325–345.

Shepard, R. N. (1958a). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review, 65*, 242–256.

Shepard, R. N. (1958b). Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology, 55*, 509–523.

Shepard, R. N. (1962a). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika, 27*, 125–140.

Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika, 27*, 219–246.

Shepard, R. N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology, 1*, 54–87.

Shepard, R. N. (1965). Approximation to uniform gradients of generalization by monotone transformations of scale. In D. Mostofsky (Ed.), *Stimulus generalization* (pp. 94–110). Palo Alto, CA.

Shepard, R. N. (1966). Metric structures in ordinal data. *Journal of Mathematical Psychology, 3*, 287–315.

Shepard, R. N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science, 210*, 390–398.

Shepard, R. N. (1982). Geometrical approximations to the structure of musical pitch. *Psychological Review, 89*, 305–333.

Shepard, R. N. (1986). Discrimination and generalization in identification and classification: Comment on Nosofsky. *Journal of Experimental Psychology: General, 115*, 58–61.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science, 237*, 1317–1323.

Shepard, R. N. (1994). Perceptual-cognitive universals as reflections of the world. *Psychonomic Bulletin & Review, 1*, 2–28.

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations*. Cambridge, MA: MIT Press.

Shepard, R. N., & Zare, S. L. (1983). Path-guided apparent motion. *Science, 220*, 632–634.

Smolensky, P., & Legendre, G. (2006). *The harmonic mind: From neural computation to optimality-theoretic grammar Vol. 1: Cognitive architecture; vol. 2: Linguistic and philosophical implications*. Cambridge, MA: MIT Press.

Soechting, J. F., & Lacquaniti, F. (1981). Invariant characteristics of a pointing in man. *Journal of Neuroscience, 1,* 710–720.

Solomonoff, R. J. (1978). Complexity-based induction systems: Comparisons and convergence theorems. *IEEE Transactions on Information Theory, IT-24,* 422–432.

Stevens, S. S. (1957). On the psychophysical law. *Psychological Review, 64,* 153–181.

Stewart, N., Brown, G. D. A., & Chater, N. (2005). Absolute identification by relative judgment. *Psychological Review, 112,* 881–911.

Surprenant, A. M., Neath, I., & Brown, G. D. A. (2006). Modeling age-related differences in immediate memory using SIMPLE. *Journal of Memory and Language, 55,* 572–586.

Taylor, G. I. (1950a). The formation of a blast wave by a very intense explosion. I. Theoretical discussion. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 201,* 159–174.

Taylor, G. I. (1950b). The formation of a blast wave by a very intense explosion. II. The atomic explosion of 1945. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences, 201,* 175–186.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. Kearns, S. Solla, & D. Cohn (Eds.), *Advances in neural information processing systems 11* (pp. 59–65). Cambridge, MA: MIT Press.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24,* 629–641.

Tenenbaum, J. B., Griffiths, T. L., & Niyogi, S. (2007). Intuitive theories as grammars for causal inference. In A. Gopnik & L. Schulz, L. (Eds.), *Causal learning: Psychology, philosophy, and computation.* New York: Oxford University Press.

Tenenbaum, J. B., Kemp, C., & Shafto, P. (2007). Theory-based Bayesian models of inductive reasoning. In A. Feeney & E. Heit (Eds.), *Inductive reasoning* (pp. 167–204). Cambridge, UK: Cambridge University Press.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84,* 327–352.

Vitányi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism and Kolmogorov complexity. *IEEE Transactions on Information Theory, 46,* 446–464.

Viviani, P., & Terzuolo, C. (1980). Space–time invariance in motor skills. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 525–533). Amsterdam: North-Holland.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding. *Proceedings of the Royal Statistical Society B, 49,* 240–265.

Weber, E. H. (1996). De tactu. Annotationes anatomicae et physiologicae [The sense of touch: Anatomical and physiological aspects]. In H. E. Ross & D. J. Murray (Eds.), *E. H. Weber on the tactile senses* (2nd ed., pp. 21–136). Hove, England: Taylor & Francis. (Original work published 1834).

West, G. B., & Brown, J. H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: Towards a quantitative unifying theory of biological structure and organization. *Journal of Experimental Biology, 208,* 1575–1592.

Wing, A. M. (1980). The long and short of timing in response sequences. In G. E. Stelmach & J. Requin (Eds.), *Tutorials in motor behavior* (pp. 469–486). Amsterdam: North-Holland.

Wing, A. M. (2000). Motor control: Mechanisms of motor equivalence in handwriting. *Current Biology, 10,* R245–R248.

Wixted, J. T., & Ebbesen, E. B. (1997). Genuine power curves in forgetting: A quantitative analysis of individual subject forgetting functions. *Memory and Cognition, 25,* 731–739.

Wolpert, D. M., & Ghahramani, Z. (2004). Computational motor control. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences III* (pp. 485–494), Cambridge, MA: MIT Press.