Cognitive Science 34 (2010) 406–433

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01079.x

Through Neural Stimulation to Behavior Manipulation: A Novel Method for Analyzing Dynamical Cognitive Models

Thomas Hope*, Ivilin Stoianov, Marco Zorzi

Department of General Psychology and Center for Cognitive Science, University of Padova, Padova, Italy

Received 10 December 2008; received in revised form 29 June 2009; accepted 8 July 2009

Abstract

The dynamical systems' approach to cognition (Dynamicism) promises computational models that effectively embed cognitive processing within its more natural behavioral context. Dynamical cognitive models also pose difficult, analytical challenges, which motivate the development of new analytical methodology. We start by illustrating the challenge by applying two conventional analytical methods to a well-known Dynamicist model of categorical perception. We then introduce our own analysis, which works by analogy with neural stimulation methods, and which yields some novel insights into the way the model works. We then extend and apply the method to a second Dynamicist model, which captures the key psychophysical trends that emerge when humans and animals compare two numbers. The results of the analysis—which reveals units with tuning functions that are monotonically related to the magnitudes of the numbers that the agents must compare—offer a clear contribution to the contentious debate concerning the way number information is encoded in the brain.

Keywords: Dynamical systems; Genetic algorithms; Evolution; Recurrent neural networks; Analysis

Dynamical systems' (''Dynamicist'') models of cognition are attractive because they embed cognition within its more natural, behavioral context (e.g., Beer, 1995, 1996, 2000; van Gelder, 1995, 1998; Port & Gelder, 1998; Thelen & Smith, 1996). But they can also pose difficult, analytical challenges, which demand new analytical methods in response. The problem is partly a function of the complexity of a dynamical agent's interactions with its world, which makes it difficult to reduce these agents' internal dynamics to atomic, individually interpretable components (e.g., Beer, 2003). The challenge these systems pose is a microcosm of the broader challenge posed by the analysis of biological neural systems,

Correspondence should be sent to Dr Thomas Hope, Room 503, Medical Faculty Building, Imperial College, St. Mary's Campus Norfolk Square, London W2 1PG, UK. E-mail: t.hope@imperial.ac.uk
*Present address: Department of Biosurgery and Surgical Technology, Imperial College London, UK.

and so has attracted increasing attention in recent years (e.g., Beer, 2003; Keinan, Meilijson, Ruppin, Hilgetag, & Meilijson, 2003; Seth, 2008). The current study presents a new response to that challenge, founded on the logic of neural stimulation, which can offer a novel perspective on the way Dynamicist models work.

The bulk of this study is divided into two sections. We start with a focus on Beer's well-known Dynamicist model of categorical perception (Section 1), employing two popular analytical methods to illustrate the interpretative challenge that it poses. We then describe our novel approach, which meets that challenge directly and yields a novel insight into the functional architecture of this model. Armed with the knowledge that our method can be useful, we then turn our attention to a second model that has a rather more direct cognitive relevance than the first. This time, the focus is numerical cognition: a Dynamicist model of an agent's ability to compare two numbers (Section 2). The model captures the key psychophysical trends that emerge when humans (and some animals) are engaged in this task, and our analysis reveals a partial explanation for that behavior—highlighting units with tuning functions that are consistent with the debate concerning the format with which numerical information might be represented in the brain. The result makes a novel contribution to that debate, illustrating that the analytical method can be put to productive use.

1. Introducing the method: Analyzing a dynamical model of categorical perception

The focus of this section is Beer's (1996, 2003) dynamical model of categorical perception. This model should be very familiar to most researchers in the field and seems a good medium for the introduction of a new methodology. For pragmatic reasons, our implementation (described in Sections 1.1 and 1.2) is not an attempt at precise replication; our goal was simply to generate agents that can perform the familiar task. But the differences between our version and its precursors are rather less important than the analyses that follow (Sections 1.3–1.5).

1.1. Description of the system

Our system is a combination of an agent and an environment that it inhabits. The environment is a two-dimensional square plane, with sides measuring 100 units; we use the notation <*X*, *Y*> to denote positions on this plane. The agent is a circle of radius 5, which begins each run at the center of the plane's lower boundary (i.e., with center <50, 0>). Agents are exposed to a series of trials in which shapes (squares or circles) fall from the square's upper boundary toward its lower boundary; the trial ends when a shape touches either the agent or the *x* axis. The agent's goal is to categorize the shapes that fall toward them, catching (i.e., touching) circles, while avoiding squares.

Shapes fall with a speed of 0.5 units per time step and occur with a range of possible radii (3–6). Squares are specified relative to the circumcircle defined by their radii, and they also occur with random rotation. Together, these two sources of variation (size and rotation) complicate the relationship between apparent shape width and actual shape type—a

confound identified by Beer (2003). Each shape starts with a random *X* position, but their centers will always fall within two 10-unit bands—one on each side of the agent's starting position (i.e., at the start of each trial, shape centers have a *Y* value of 100, and *X* values in the ranges 20–30, or 70–80); this latter restriction was intended to eliminate shapes that fall from directly above the agent, as these special cases have previously been shown to raise particular problems in the past (Beer, 2003).

Each agent is a continuous-time dynamic recurrent neural network, updated synchronously in time steps. The activity **u** of unit **i** at time step **t** is calculated using Eq. 1:

$$u_i(t) = u_i(t-1) + (1/\tau_i)\sigma\left(\sum_{j=1}^{N} w_{ji}(u_j(t-1))\right) \quad 1$$

where $\mathbf{w_{ji}}$ is the weight of the connection from unit **j** to unit **i**, $\boldsymbol{\sigma}()$ is the sigmoid function and refers to a unit-specific time constant (higher time constants indicate a lower dependence on incoming activity).

The agents' visual system is analogous to a laser range-finder. Seven rays project upwards from the center of each agent, spanning a 60° angle—adjacent rays subtend an angle of 10°, and each ray has a length of 110 units. If a ray intersects with a shape, activity is passed to its associated sensor unit; the value passed is inversely proportionate to the distance to that intersection. Agents can move in only one dimension, along the *x* axis—at each update, the change in an agent's position is proportionate to the difference between the activity values of two effector units, with a maximum speed of five units per update. To improve the similarity between our agents and those analyzed in previous work, we also restrict the agents' hidden layers to include exactly seven hidden units (as in Beer, 2003). With the exception of the sensor units, which are always fixed by properties of the ''world,'' so receive no incoming connections, the agents' neural networks are universally connected; every unit is directly connected to every other, and to itself. Fig. 1 presents a schematic of the agents' network architecture, together with an illustration of an agent in its environment.

We used a Microbial genetic algorithm (Harvey, 2001) to design the agents in this system, which are initially specified at random. During each iteration, two ''parents'' are randomly selected from the population and compared. The weaker of the two parents is replaced by their ''child,'' which is defined by mixing the parents' weight vectors and time constants (each parent contributes a parameter with 50% probability) and applying a mutation. The mutation operator usually implements a small, random change (±0.01) to a randomly selected weight, but it will sometimes (p = 1%) increment or decrement a unit's time constant instead.

An agent's fitness score is the sum of the absolute distances between that agent and each of a set of 100 shapes at the end of each of 100 trials; each set is composed of 50 pairs of shapes, identical in every respect but for their type. Distances to circles (which should be caught) are counted negatively, and distances to squares (which should be avoided) are counted positively; the fitness **f** of individual **i** is calculated as in Eq. 2:

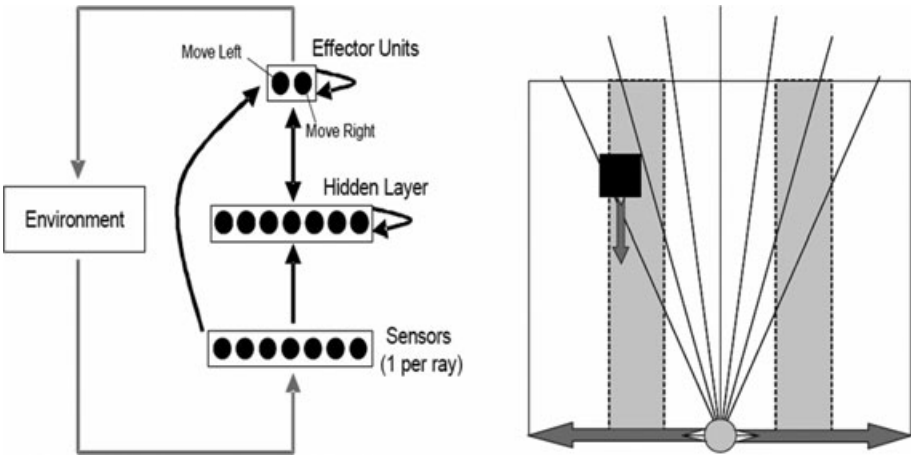

Fig. 1. (Left) Schematic network structure for agents designed to solve a visual object classification problem. The agent has seven sensor units, seven hidden units, and two effector units. The sensor units' activities are always fixed by the agent's environment, but the hidden and effector units all receive direct input from every other unit, and from themselves. (Right) An agent in its environment. Shapes fall from the environment's upper boundary toward its lower boundary—their centers always fall within one of the two shaded areas. The agent's task is to touch falling circles and to avoid falling squares; they can move in only one dimension, along the environment's lower boundary.

$$f_i = \left| \left(\sum_{s=1}^{N} |x_i^s - x_s| \right) - \left(\sum_{c=1}^{N} |x_i^c - x_c| \right) \right| \quad 2$$

where **s** indicates squares, **c** indicates circles, x_t is the *x*-axis position of a particular shape (of type **T**) at the end of a trial, and x_i^T is the *x*-axis position of agent **i** at the end of the same trial. Sets of shapes were generated randomly for each competition, but two prospective parents were always compared with the same set. Evolutionary runs were ended once an agent in the population had achieved 100% accuracy on any shape set.

The best agents in this system achieve good performance after about five million iterations of the microbial algorithm; the evolutionary process was repeated five times, and all runs produced agents with similar shape-following/avoidance behavior. The material that follows will focus on an agent from the first run, which achieved 100% accuracy on one shape set and retained good performance (>98%) when tested against 100 other randomly generated sets.

1.2. Behavioral analysis

Fig. 2 graphs the global absolute distance between the agent's center and the shape centers in trials of different type; the data illustrate that the agent does in fact catch circles while successfully avoiding squares. There is also an apparent similarity in the paths during the early stages of both trial types—an active scanning strategy (at least superficially similar to that identified by Beer, 2003 that gives way to genuine divergence only around the 90th

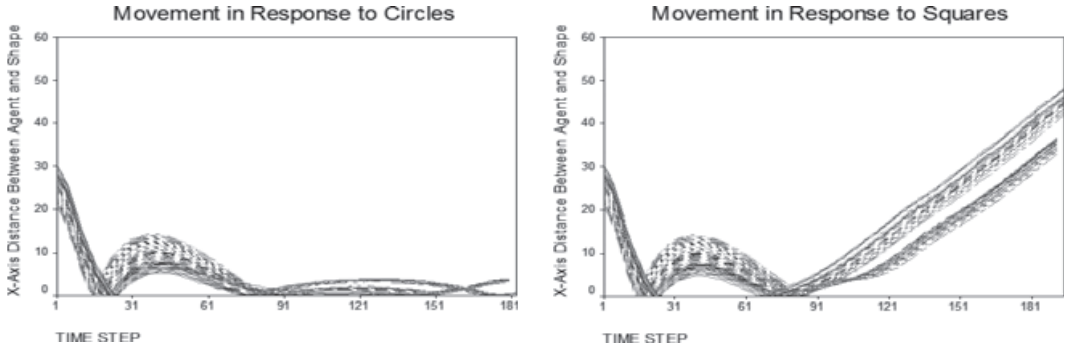

Fig. 2. Absolute lateral distances between the agent and shape-centers in a random selection of 100 shapes (squares and circles). Each series refers to a single trial. (Left) Behavior in response to circles. (Right) Behavior in response to squares.

time step.) Intuitively, this pattern suggests a perceptual process, which drives a categorization ''choice'' that defines subsequent behavior.

In the material that follows, we offer analyses that are designed to explain this shape-following and avoidance behavior. A complete review of the analytical methods that have been proposed is beyond the current scope. Instead, we have chosen to employ two familiar analytical methods that we believe will clarify the contribution that the third—our own proposal—can make to the field.

1.3. Principal components analysis (PCA)

One of the most popular tools for neural network analysis, PCA is a technique for expressing high-dimensional data sets as lower dimensional data sets, while preserving the data's underlying variance. Neural network state-spaces have at least as many dimensions as they have units—usually far too many to comprehend directly. PCA can reduce that apparent complexity, exposing the fundamental dimensions of a network's state trajectory.

The mathematics underlying PCA is well-described elsewhere (e.g., Gonzalez & Richard, 1992; Oja, 1989; Rao, 1964); we will provide only a brief summary. Our version of this method, based on that used by Elman (1991), begins by recording step-by-step hidden unit activities as the agent attempts to categorize falling shapes. The series composes an $[N \times T]$ matrix, where N is the number of hidden units (seven in this case), and T is the total number of time steps required to complete the 100 trials. [1] From this ''activity matrix,'' we can calculate a covariance matrix; the dimensions that PCA identifies are eigenvectors of this matrix, and their eigenvalues correspond to the variance that each accounts for.

Three principal components account for 88% of the variance in hidden unit activities; in Fig. 4, we plot the way these components change throughout each trial. The features of interest here are *differences* between these components in trials of different type, as a sensitivity to shape type is a precondition (i.e., necessary but not sufficient) for the representation of shape type. Moving a bit beyond Elman's method, we can quantify these differences sta-

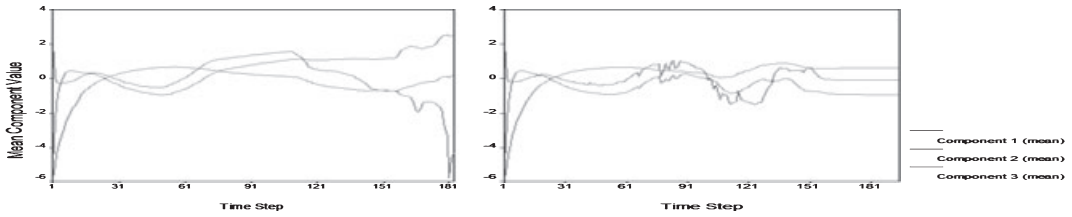

Fig. 3. The agent's hidden unit state trajectory, projected onto three Principal Components, during shape categorization trials involving circles (left) and squares (right).

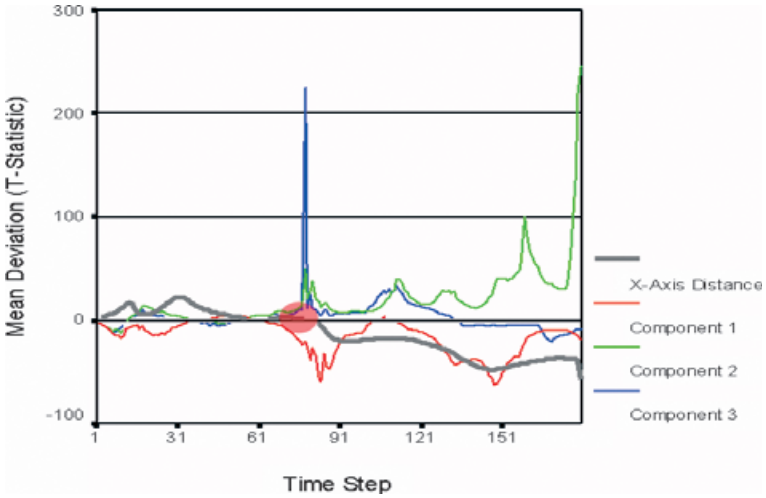

Fig. 4. T-statistics ($p < 0.001$) for tests of shape-sensitive deviation. The red circle marks a possible ''decision point,'' the point beyond which the agent's shape-sensitive behavioral deviation is consistently significant.

tistically—comparing the values of each component (and of the *x*-axis distances from Fig. 3) at each time step in trials of different type. Remember that the shape set includes 50 squares and 50 circles, and every square is paired with an equivalent circle, identical in every respect but for its type. At each time step, there are therefore 50 values for each component in square-trials, paired with 50 values for each circle-trial. None of the samples deviates significantly from a normal distribution, so we use *t*-tests for paired samples to quantify the differences. Fig. 4 displays the *T*-values (where $p < 0.001$) derived from these tests; the series represents the extent to which each principal component, and also the agent's lateral distance from the shapes, is different depending on the categorization decision that the agent is required to make.

Our proposed decision point is visible in Fig. 4—the final period of behavioral similarity (i.e., *t*-values for lateral distance between the agent and shape centers are close to zero) before the shape-dependent deviation in the agent's behavior is *consistent* in its statistical significance. Further, one of the principal components (component 3) displays a very

extreme pattern of shape-sensitive deviation during that period; this pattern reflects a stark reduction in the variance of component 3 at that time, which implies significant uniformity across trials involving the same shape type. The temptation might be to conclude that this agent *does* make a decision, and that component 3 conveys the result.

The problem with this interpretation is that it risks circularity—we have decided that there must be a decision point, then ''found'' it in the model and used it to justify our original intuition. And the conventional logic that researchers use to justify their interpretation of principal components does not provide much independent support. That logic implies using linear regression to associate shape-dependent differences in the agent's behavior with shape-dependent deviations in particular components. With shape-dependent differences in behavior (T-values) as the dependent variable, and the shape-dependent differences in the three principal components (T-values) as separate independent variables, we have three regression analyses with series that each contain 50 values; the results emphasize components 1 ($p < 0.001$, $R^2 = 0.33$) and 2 ($p < 0.001$, $R^2 = 0.32$), whereas marginally dismissing component 3 ($p = 0.054$, $R^2 = 0.02$). However, shape-sensitive behavioral deviation is evident very early in each trial—and certainly before our proposed decision point—so it is far from clear that these associations can justify strong clams about the best interpretation of any of the principal components. We are left with an apparent limit on the scope of the results that PCA can provide. It can focus our attention on the key dimensions of variation in a model's internal dynamics, and we can relate that variation to task-relevant stimuli, but any attempt to attach a firm meaning and/or causal role to the components will be difficult to justify.

One other limitation of PCA emerges when we consider the values in Table 1, which indicate the correlations between hidden unit activities and the extracted components. Six of the seven hidden units are significantly correlated with component 3; the analysis has collapsed their activity nicely, but it does not tell us how to weigh the contribution of one network unit against the others. Must we inspect all six of these units to observe the key dynamics of this system? If so, the analysis has carried out little to reduce the agent's apparent complexity. Should we define some minimum correlation below which we can ignore particular units? Although perhaps appropriate in some circumstances (such as the analysis

Table 1
Correlations between hidden unit series and the three principal components

	Principal Components		
Hidden Unit	1	2	3
Unit 1	.699**	.279**	−.041**
Unit 2	.070**	−.313**	−.876**
Unit 3	.984**	−.120**	−.003**
Unit 4	.410**	.608**	.401**
Unit 5	.108**	.196**	.966**
Unit 6	.128**	.957**	−.119**
Unit 7	−.015**	−.639**	.195**

**$p < 0.001$; *$p < 0.05$.

of fMRi data: for example, Friston, Worsley, Frackowisk, Mazziotta, & Evans, 1994), this approach seems a poor compromise when better options are available.

As we will see, better options *are* available. In Section 1.4, we consider an alternative that addresses a general concern which lurks behind much of the discussion so far: Correlations and covariance offer at best a limited view of their object's underlying causal structure. To begin to garner evidence of this more causal sort, lesion studies are required.

1.4. Multiperturbation Shapley value analysis (MSA)

Just as neurological disorders can illuminate the functional structure of normal brains (e.g., Shallice, 1988), so lesion analyses can clarify the functional architecture of neural networks. There are almost as many specific methods for lesion analysis as there are researchers to use them. In this section, we will focus on one of the method's more systematic variants, called multiperturbation Shapley value analysis (MSA). MSA was originally inspired by the economics of share-dividend calculation (Keinan, Hilgetag, Meilijson, & Ruppin, 2004), and its results associate each of a network's hidden units with a contribution value (CV) or causal significance, relative to some defined measure of behavioral performance. That CV is essentially a Shapley value.

The Shapley value (Shapley, 1953) is a familiar concept in game theory, and it describes an approach for calculating the fair allocation of gains obtained through the cooperation of groups of actors—allowing for the possibility that some actors may make a greater contribution than others. In formal terms, this situation can be described as a *coalitional game*, defined by a pair ($\boldsymbol{N}$, $\mathbf{v}$), where $\boldsymbol{N} = \{1,...., n\}$ is the set of all *players* and $\mathbf{v}$ is a is a real number associating a worth, or *payoff*, with the game; the goal is to calculate a *payoff profile*, associating each player with a specific proportion of that total payoff. Shapley's approach started by measuring the *marginal importance* of each actor $\mathbf{i}$ relative to each subgroup of actors ($\mathbf{S}$, where $\mathbf{S} \subset \mathbf{N}$)—this is the difference between the payoff for group ($\mathbf{S} \cup \mathbf{i}$) minus the payoff for group $\mathbf{S}$ alone. Actor $\mathbf{i}$'s Shapley value is then simply its average marginal importance for all permutations of $\mathbf{S}$.

As applied to the analysis of neural networks, this formulation requires access to the performance scores associated with every subgroup of the networks' units. That ''full information'' approach may be prohibitive for large networks—and the MSA method's authors do offer a reduced, or ''predicted'' approach to reduce that load (Keinan et al., 2003, 2004)—but the current agent is quite small, so perfectly susceptible to this kind of exhaustive analysis. The agent has seven hidden units, so there are $2^7 = 128$ subgroups in all (one of those groups includes all of the agent's hidden units), and the analysis requires that we conduct 128 performance tests. Each performance test is defined by a ''Lesion Configuration,'' which specifies the units that will be removed for that test.

Following Keinan and colleagues' own preference, the current study also employs ''informational lesions,'' rather than the more traditional ''biological lesions'' to implement each lesion configuration. Biological lesions are so-called because they mimic the probable impact of neural lesions, effectively removing units either by setting the weights of their outgoing connections to zero (e.g., Joanisse & Seidenberg, 1999) or by adding ran-

dom noise to their activity values (e.g., Plaut & Shallice, 1991). As the name suggests, informational lesions are merely intended to remove a unit's information, and they work by fixing its activity to an average value. There are numerous reasons for this choice (see Aharonov, Segev, Meilijson, & Ruppin, 2003; for a discussion), but the central intuition behind it is that functional analyses can be misleading if their objects—the networks under study—are too far removed from their ''natural'' state (e.g., Seth, 2008). The different roles of biological and informational lesions can also be illustrated with the simple example of bias units.

Bias units, a common feature of neural networks, have activity values that are always close to '1' regardless of a network's other dynamics. Often explicitly specified, bias units can also (and often do) emerge through learning, or simulated evolution. When applied to bias units, biological lesions can have a profound effect on a network's state trajectory, as the lesion drastically alters a consistent feature of the network's default state. By contrast, informational lesions will have no effect whatsoever when applied to these units. The preference for informational lesions can be interpreted as expressing a position on what constitutes a ''good'' explanation of network function—bias units are of minimal interest in those explanations.

In practice, the MSA method starts with a baseline performance test (with no lesions), during which the activity values of every hidden unit at every time step are collected; these data define the average values that informational lesions employ, as well as a performance standard (a categorization accuracy rate) for the analysis that follows. We then repeat the same series of trials (with the same shapes in the same order), while applying the informational lesions defined by each of the lesion configurations (one per performance test). Lesion configurations can be thought of as binary lists with one cell for each of the agent's hidden units; if the value in a unit's cell is '1,' a lesion is applied, whereas a '0' indicates that the

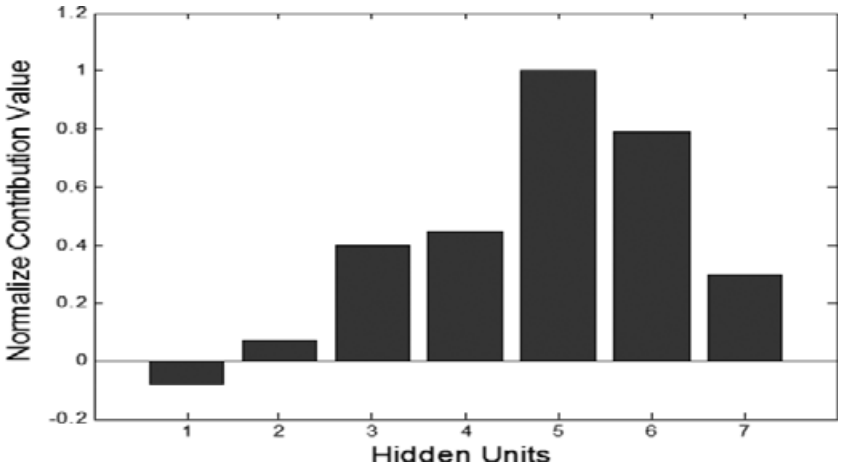

Fig. 5. Normalized contribution values of each of the agent's seven hidden units; higher values indicate that units appear to play a more significant role in the agent's classification performance.

unit is allowed to vary freely. The results associate performance scores with each lesion configuration—and by implication with every subgroup of network units; Fig. 5 displays normalized contribution (Shapley) values for each unit as calculated by Keinan and colleagues' own Matlab implementation of the process.[2]

Of the seven units, two (units 1 and 2) appear relatively insignificant; we can probably ignore those in our search for representations. Note that, although largely causally insignificant, unit 2 did display a strong correlation with the most intuitive source for representations (principal component 3) that we observed with PCA—a good illustration that correlations and covariance really are an imperfect metric of causal significance. The five other units all do seem to play some role, and two of them (units 5 and 6) may justify particular attention. To interpret these results, we need to inspect the unit activity series themselves; Fig. 6 displays the average series of each of those five units during trials involving circles and squares, respectively.

The visual similarity between Figs. 6 and 3 is unsurprising—three of these hidden units are extremely highly correlated with the three factors that we previously extracted. Note too that, given this result, we are still faced with much the same problem of interpretation as we had before; at least five of the agent's seven units seem to demand some scrutiny. The MSA method's original motivation stemmed from the intuition that, often, task-specific functional significance will be localized to small subsets of units (Keinan et al., 2003, 2004). If this prediction is satisfied, MSA may be useful—but there is no guarantee that it will be. Larger networks, with more complex behavior, might yield results that are simply too complex to be useful.

Like PCA, the results of MSA only go so far. The shift to causal evidence is an improvement over PCA, but both methods still depend critically on a subjective eye-balling process, which limits the scope of the conclusions they can support. These problems highlight the need for a method that addresses interpretation directly.

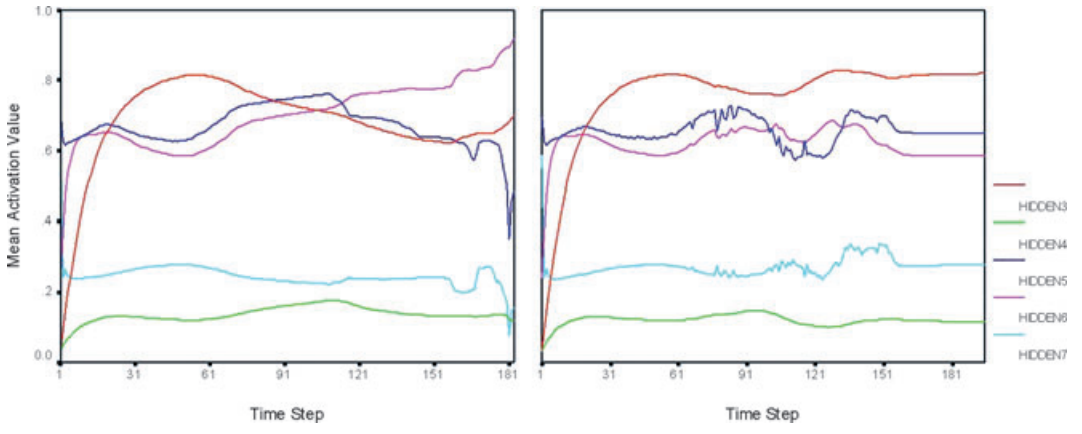

Fig. 6. Average activities of five units that are causally significant to the agent's categorization performance. (Left) Trials involving circles. (Right) Trials involving squares.

1.5. The behavior manipulation method (BMM)

Consider a hypothetical agent, designed to solve our categorical perception problem, which uses a simple strategy in which perception is used to establish internal states that stand for the salient features of the task (i.e., shape types) in a clear and transparent manner. As the agent is simulated, we have great freedom to constrain its state trajectory; if we know what those representations are, we should be able to force the agent to ''perceive'' squares as circles, and vice versa. When subject to that forced perception, the agent should behave as if squares really are circles, and circles really are squares.

Our BMM is a practical extension of this example's logic—an analysis based on targeted lesions, which share some features in common with MSA. Like MSA, the BMM employs integer vectors to define the lesions, but the meaning that those vectors convey is rather different. In deference to that difference, we will refer to these lists as ''Candidates,'' rather than lesion configurations, in the material that follows. In the language of our hypothetical example, Candidates can be construed as hypotheses concerning the best way to control an agent's perception of its environment; better Candidates permit ever-more effective and predictable mediation of the agent's behavior.

Our particular implementation of this method borrows from the concept of the informational lesion, described previously, which involves fixing a unit to its average activity. We extend the concept to define a ''partial informational lesion,'' which involves fixing a unit to its average activity *in specific circumstances*. Where the informational lesion is designed to remove a unit's information, the partial informational lesion offers a positive hypothesis concerning the meaning of that unit's activity—that averages convey information about the circumstances that define them. This choice is a useful starting point because it reduces the complexity of unit activity series, and because it accords with the way in which, in practice, researchers manage the apparently random variation in neural spike train data (e.g., Tomko & Crapper, 1974).

1.5.1. The BMM in practice

Like MSA, the BMM begins with a series of ''natural'' trials, providing both a baseline for the agent's categorization performance (the number of correct categorizations: 100% in this case) and a record of its hidden units' activity values throughout each trial. From these latter data, we can calculate two average activities for each unit—one for trials involving circles and one for trials involving squares (these averages group every time step in each trial type together). As with MSA, we then repeat the same series of categorization trials (employing the same shapes in the same order) while lesioning the agent's hidden units; each new experiment is defined by a Candidate, which specifies the lesions that should be performed.

Like lesion configurations, our Candidates associate each of the agent's hidden units with either a '0' or a '1'. In the former case, the unit is allowed to vary freely, whereas in the latter, a PIL is applied. To assess the quality of each Candidate, we attempt to use them to reverse the way the agent responds to shape types; in trials involving squares, lesioned units are fixed to their average activities for trials involving circles, whereas in trials involving

circles, lesioned units are fixed to their average for trials involving squares. ''Good'' Candidates will encourage the agent to catch squares and avoid circles. The best Candidates should encourage incorrect categorizations of most, or all, of the shapes. As with MSA, the current version of the BMM implements an exhaustive search of the agent's Candidate-space, testing each of the 2^7 Candidates.

1.5.2. Results

Startlingly, the results of this analysis yield several Candidates who permit very accurate manipulation of the agent's categorization choices. One Candidate yields a perfect result—a 100% categorization error rate (i.e., the Candidate defines a set of PILs that let us force it to avoid all circles and catch all squares). This Candidate defines a concrete role for particular hidden units, which appear to convey information about shape type through their average activity values. Fig. 7 displays the best Candidate who we found: a distributed solution that implicates units 1 and 3 through 7. Each of these units displays deviations in their average activities that appear sensitive to shape type; these deviations can be used to ''fool'' the agent into confusing square for circles and vice versa.

1.6. Comparing the BMM with PCA and MSA

Like PCA and MSA, the BMM can be interpreted as a kind of filter, directing our attention to those features of a network's dynamics that drive the behavior of interest. During our discussions of both PCA and MSA, we noted that there is no guarantee that these ''analytical filters'' will provide results that are both sufficiently well justified to be useful and transparent enough to be interpretable. Although the BMM also implicates many hidden units, its results are much more interpretable; each unit is associated only

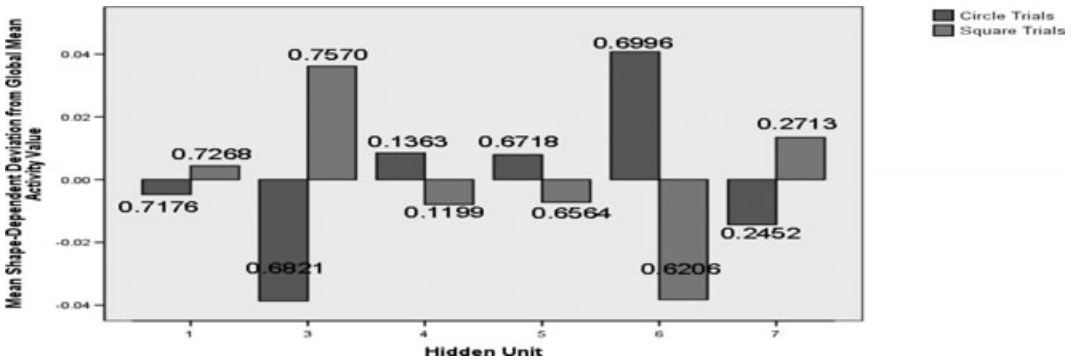

Fig. 7. Average shape-type-dependent deviation in hidden unit activity values. The bars indicate the positive and negative differences between each unit's shape-dependent average activities (one for circles, and one for squares) and their shape-*independent* average activity (i.e., calculated across all trials regardless of shape type). The absolute values of shape-dependent average activities for each unit are appended to each bar. This deviation can be recruited to reverse the way the agent categorizes shapes. The implication is that this deviation stands for (or represents) the agent's knowledge of shape type.

with a pair of values (averages)—and we know that these values have played a definite, causal role.

Even on its own, that knowledge can be useful. Figs. 3, 4, and 6 all graph mean values of the series under study, but the choice is really a pragmatic attempt to clarify the presentation. Nothing in the logic of either PCA or MSA can demonstrate that these mean values are causally significant in themselves. This conflation is all the more tempting because it is largely accepted in the analysis of, for example, neural spike train data (e.g., Roitman, Brannon, & Platt, 2007). Just as minimum correlation thresholds are acceptable in the analysis of fMRi data (mentioned in Section 1.3), so this conflation is acceptable when no better methodological options are available. But as the BMM illustrates, computational models permit far more invasive analyses than their biological referents. As we *can* verify the causal significance of unit averages directly, it seems reasonable to require that we should.

Another encouraging feature of our results is that they appear to be largely consistent with those derived from MSA. The best Candidate that the BMM identifies includes units 3–7—the same group to which our MSA assigned high CVs. If PILs are applied only to these units, a 96% categorization error rate can be achieved; the natural influence of the dynamics of these five units appears to be well captured by their average activities in trials of different type. This consistency also clarifies the different contributions that each method makes. MSA allows us to rank hidden units by their causal significance, and that information is absent in the results of the BMM (at least as currently defined). On the contrary, the BMM supplies a justifiable interpretation of the meaning that those units convey—in this case by implementing the agent's sensitivity to shape type—whereas MSA leaves this to the observer. However, despite this overall consistency, the BMM does seem to disagree with MSA in the way it characterizes hidden units 1 and 2.

Unit 1 has a negative CV, implying that informational lesions actually improved the agent's performance when applied to this unit, but Unit 1 is also part of the best Candidate that we found (displayed in Fig. 7). The implication is that the effect of our PIL is very close to that of a normal informational lesion. The shape-dependent averages for all units are numerically quite close, but they are closest for unit 1; in this case, the partial informational lesions' probable primary role is to remove the unit's variance, helping the agent to act on its ''knowledge'' of shape type (encoded by units 3–7).

In the case of Unit 2, a positive CV does not yield a positive role in our best Candidate. The implication here is that Unit 2 helps the agent not by encoding its knowledge of shape type, but by helping to guide the shape-following and avoidance behavior that this knowledge informs. Note that if the agent's control of movement depended mostly on its hidden units, performance would fall apart when a PIL is applied to them—but the Candidate that includes all hidden units displayed fairly accurate behavior (88% reversed accuracy). The implication is that—once the agent ''knows'' which type of shape it is dealing with, so has ''decided'' whether to try to catch or avoid that shape—the control of the movement behavior itself is mainly carried out by the agent's direct sensor-to-effector connectivity. This is as it should be, as each of the two behaviors (catching and avoiding) can be expressed by a linear mapping from the sensor units. Nevertheless, a freely varying Unit 2 is clearly critical to the perfect performance that this agent achieves.

1.7. Interim discussion

The goal of this section was to describe and demonstrate the BMM—a novel method for analyzing Dynamicist cognitive models. To clarify that demonstration, we began by describing two more familiar methods—PCA[3] and MSA. Each of these methods can act as a filter, guiding the focus of our attention, but there is no guarantee that either will filter the data into a neatly interpretable form. By contrast, the BMM offers a formal, statistically justifiable route toward the identification of causally significant, transparently interpretable states. By establishing an ''anatomical'' distinction between the components which implement shape-type-sensitivity, and the components which use that sensitivity to guide behavior, we also confirmed that the BMM can provide some insight into a model's functional architecture.

However, the details of this model's architecture are rather less important than the fact that we can analyze it effectively; this agent is rather less convincing as a model of cognition than it is as a spur for the ''mental gymnastics'' (Beer, 1996) required to develop good analyses. However, armed with the BMM, we can turn our attention to models with much more direct relevance to cognition.

2. Analyzing a dynamical model of quantity comparison

Knowing that one quantity is greater or smaller than another is the basis of cognitive number knowledge. Quantity comparison is therefore fundamental to researchers with an interest in the cognitive neuroscience of numeracy (e.g., Butterworth, 1999; Dehaene, 2000). Like categorical perception, quantity comparison has attracted accounts that emphasize the neural representations which drive the behavior of interest.

2.1. The representation debate

Contemporary models of quantity comparison all take very specific stances on the way in which quantity information is represented; at least four formats (Dehaene & Changeux, 1993; Gallistel & Gelman, 2000; Verguts & Fias, 2004; Zorzi & Butterworth, 1999) have been proposed. Arguments in favor of one or another theory invoke (at least) two empirical phenomena—the Size and Distance effects (Moyer & Landauer, 1967)—which reliably emerge when subjects compare two quantities. The Size effect refers to a characteristic trend of increasing errors and reaction times as the numerical magnitudes of manipulated numbers increase; for example, the comparison ''2 versus 3'' appears easier to solve than the comparison ''8 versus 9''. The Distance effect refers to the pattern of increasing errors and reaction times as the numerical distance between compared numbers decreases; selecting the larger among the pair ''4 versus 5'' appears a more difficult judgment than among the pair ''4 versus 6'' (Moyer & Landauer, 1967).

The most popular accounts of these effects refer directly to the logical structure of subjects' semantic number knowledge. Consider, for example, the ''compressed number line'' theory advanced by Dehaene & Changeux (1993) (see Fig. 8). Different positions on this

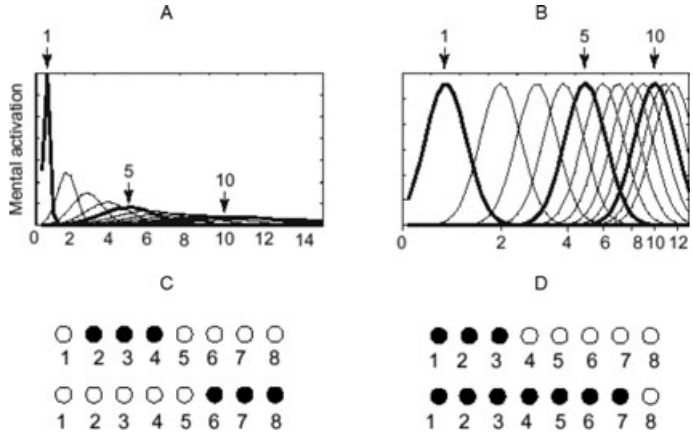

Fig. 8. Four popular proposals for the format of semantic number representations. (a) The linear number line with scalar variability (Gallistel & Gelman, 1992, 2000); the *noisy MNL*. (b) The compressed number line with fixed variability (Dehaene & Changeux, 1993); the *compressed MNL*. (c) The ''barcode'' (or shifting bar) representation (Verguts & Fias, 2004); the *barcode*. (d) The cardinal accumulator representation (Zorzi & Butterworth, 1999); the *numerosity code*.

line are defined by the magnitude-specific tuning of neural receptive fields, with each field approximating a Gaussian shape. In other words, particular neurons may ''prefer'' magnitude X, but they will also be partially activated by magnitudes $X + 1$ and $X - 1$, and less so by $X + 2$ and X-2, and so on.

The key to this theory is the logarithmic compression of the receptive field centers associated with magnitudes; as the magnitudes increase, the distance between the centers associated with numerically adjacent values decreases. The effect is to ensure greater overlap between the representations of larger numbers, which are therefore harder to compare (the source of the Size effect). The Distance effect arises because numerically more distant numbers imply representations with lesser overlap, which therefore exhibit less mutual interference and are easier to compare. Although different in detail, two other theories of semantic number knowledge—the linear number line with scalar variability (Gallistel & Gelman, 1992, 2000) and the numerosity code (Zorzi & Butterworth, 1999)—offer analogous accounts of the Size and Distance effects.

Conventional models of this process are pitched at the level of the functional module. Number representations—of the designer's preferred form—are used to represent numerical problems, which are typically associated with ''answers'' through learning (Zorzi, Stoianov, & Umiltà, 2005, for a review). In the case of quantity comparison, the problem might be a pair of operands (e.g., ''7 versus 4''), and the answer a decision between them (e.g., ''7 is larger''). The quality of a model is then judged by comparing its behavior with empirical data.

Although effective as far as it goes, this approach inserts an important level of indirection between theories of representation and the computational evidence that supports them. Interface specifications can have a profound impact on model behavior (e.g., the structure of

input and output representations in models of reading aloud; see Perry, Ziegler, & Zorzi, 2007; for discussion), but these choices are still a few among many. Other details, such as the particular choice of learning algorithm and the number of hidden units employed, can also be important, although these choices are usually rather difficult to justify (e.g., McCloskey, 1991). Even if conclusive judgments could be made about the relative qualities of competing models, it would still be difficult to connect that quality directly to a particular choice of representation. In this context, Dynamicism is attractive because its results can be ''minimal'' in the sense described by, among others, Nowak (2004)—because so much of their eventual structure can be problem-driven rather than designer-driven. If successful, they should provide an attractive alternative source of evidence for the number representation debate.

2.2. An evolutionary start-up kit?

The particular context of numerical cognition permits a further justification of the Dynamicist method, at least as currently employed. Recent evidence suggests that both prelinguistic infants (Feigenson, Carey, & Hauser, 2002) and animals—from apes (Biro & Matsuzawa, 2001) and monkeys (Nieder, Freedman, & Miller, 2002) to pigeons (Brannon, Wusthoff, Gallistel, & Gibbon, 2001) and salamanders (Uller, Jaeger, Guidry, & Martin, 2003)—are sensitive to numerical quantity. Although learning is clearly involved in the development of an adult human's ''number sense'' (Dehaene, 2000), this evidence supports the theory that some rudiments of that sense—an ''evolutionary start-up kit'' (Butterworth, 1999)—are genetically determined. Instances of apparent Size and Distance effects have also been observed in animals (Jordan & Brannon, 2006a,b); as these are traditionally explained by reference to the format with which quantities are represented, the implication is that this genetic contribution might include a preparedness to employ that format.

In other words, in this case at least, genetic algorithms can take on a theory-driven dimension—we can interpret the behavior-based selection as a deliberate attempt to capture the impact of evolution. Our model is founded on the intuition that quantity comparison emerges from a selective pressure to forage effectively (e.g., Gallistel & Gelman, 2000); effective foragers will tend to ''go for more'' (Uller et al., 2003) food, implying an ability to judge relative quantity. We implement this logic by ''evolving'' quantity-sensitive foragers.

2.3. The artificial ecosystem

The environment is a simplified ''berry world''; a 2D toroidal grid, composed of 100×100 square cells, where each cell can contain up to nine berries. Food is initially randomly distributed throughout the environment, with a uniform probability that a given cell will take any of the possible food values (0–9). As food is ''eaten,'' it can be replaced by random ''growth'' in other cells. Growth rates are adjusted to maintain the total quantity of available food at no less than 80% of its original value.

The ecosystem includes a fixed population of 200 agents, which traverse their environment by moving between adjacent cells. The agents are recurrent, asymmetrically connected, rate-coded neural networks; the activation value **u** of the unit **i** at time **t** is calculated using Eq. 3:

$$u_i(t) = \sigma\left(\sum_{j=1}^{N} w_{ji}(u_j(t-1))(1-m)\right) + u_i(t-1)m \qquad 3$$

where $\mathbf{w_{ji}}$ is the weight of the connection from unit **j** to unit **i**, $\boldsymbol{\sigma}()$ is the sigmoid function, and m is a fixed momentum term with a value of 0.5. This momentum term replaces the unit-specific time constants used in Section 1, and it is equivalent to fixing all of those constants to '2'. We also impose the further restriction that sensors and effectors can only mediate each other through hidden units (Fig. 9).

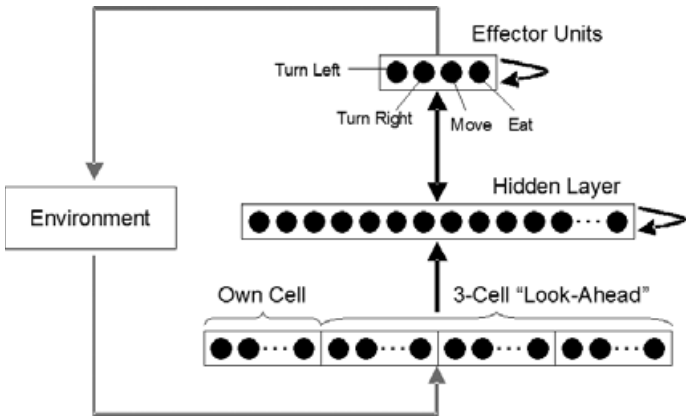

Fig. 9. Schematic structure of the quantity comparison agents' neural network architectures. The sensor layer is composed of four cells, each with nine units. The hidden layer is initialized at 10 units, but agents in the final population invariably have between 23 and 26 hidden units.

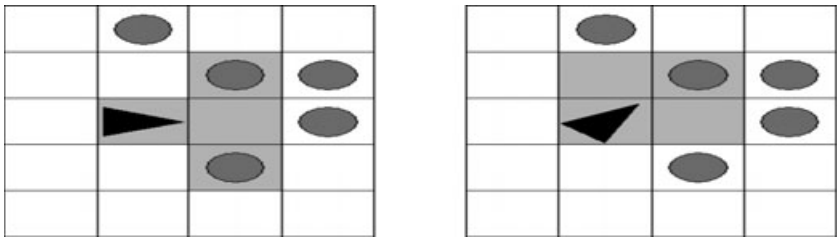

Fig. 10. An agent in its environment. (a) The agent—a black triangle—is facing right and can sense food (gray circles) in its right- and left-most sensor fields. (b) The same agent, after making a single turn to the left. It can now sense only one cell containing food.

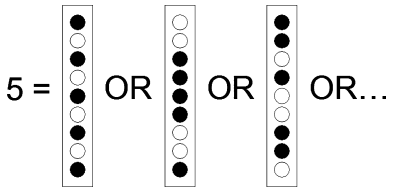

Fig. 11. The random position code, similar to that employed by Verguts, Fias, and Stevens (2004). To represent the quantity *N*, the code requires that exactly *N* (randomly selected) sensor units be active. The code is illustrated for $N = 5$.

The agents' sensors are always clamped according to the food values of the cells within their ''field of view'' (see Fig. 10)—agents are sensitive to the cell that they currently occupy and to the three cells directly ahead. Each sensor field represents a corresponding food value with a ''random position code''; this was used by Verguts and Fias (2004), among others, to capture quantity information without employing any of the popular representational strategies (see Fig. 11). By using this code, we are also restricting the problem that agents must solve, assuming away the perceptual cues, such as element size (Miller & Baker, 1968) and density (Durgin, 1995), that mediate numerosity judgements in humans. These simplifications are important but permissible given the current, primarily methodological focus.

Effector units are interpreted to define the agent's behavior during each update; agents can turn left or right, move forward, or eat. Each action is associated with a unit and is executed if its unit's activity is supra-threshold (here, above 0.5). When two inconsistent actions—turning left and turning right, or eating and moving—are attempted at the same time, neither occurs.

The ecosystem proceeds by iterative update—each update allows every agent the opportunity to sense its environment and act. Agents are updated in a random order, which is recalculated at the beginning of each time step. As in Section 1, we use a genetic algorithm to implement the evolutionary process. The crossover and mutation operators are also identical to those used in Section 1, with the exception that we include a dynamic hidden layer that can grow and shrink in size; additions to and subtractions from the hidden layer (with a p of 0.01) replace the mutation of time constants in that model. In the current case, an agent's fitness is simply the rate at which it collects food, defined as the amount of food collected since its ''birth,'' calculated by Eq. 4:

$$Fitness_i = \frac{Food_i}{Age_i} \tag{4}$$

where the age of individual **i** is the number of time steps since its creation. The goal is to promote the emergence of agents that forage for food in a quantity-sensitive manner—choosing to move into cells that contain the most food by comparing the quantities of food that they can see. The best signal that this behavior has begun to emerge is high food collection efficiency (food collected per moves made in the environment), which rises

toward five after about 10 million iterations (∼50,000 generations). The evolution was repeated three times, and all three yielded populations that achieved similar distributions of food collection efficiency after a similar number of iterations; the results that this chapter reports are based on the first of those populations.

2.4. Behavioral analysis

To capture the agents' quantity comparison performance, we remove them from their ''natural'' environment and placed them into a 3 × 3 ''mini-world'' (Fig. 13). Two cells, the top left and top right of the world, contain food of varying quantity. In its initial position at the center of the world, the agent can ''see'' both of these food quantities, although it can also turn without constraint once each trial begins. Food ''selection'' occurs when the agent moves onto one or other of the filled cells—the only cells onto which it is allowed to move. A correct choice is defined as the selection of the larger of the two food values; this is analogous to the method used by Uller et al. (2003) to capture quantity comparison performance in salamanders.[4] Every agent in the population was tested using this methodology, with 50 repetitions of every combination of food quantities (1–9, 72 combinations in all), for a total of 3,600 trials per agent. The results are displayed in Fig. 12.

A few of the agents perform extremely poorly, indicating that the evolved foraging solutions are brittle in the face of ''evolutionary'' change. This brittleness may also reflect a more general mutation bias against specialized structures (Watson & Pollack, 2001). The main bulk of the population distribution is also apparently bimodal; agents in the left-most cluster perform at roughly chance levels, whereas agents in the right-most cluster perform significantly above chance—only this latter group appears to discriminate quantity. The persistence of nondiscriminating agents reflects the fact that high rates of food collection can be achieved by sacrificing decision quality in favor of decision speed. A visual inspection of the performance scores for these agents indicates strong asymmetry in their behavior; many simply ''choose'' the right-hand square regardless of the food quantities presented.[5] Using the results displayed in Fig. 12, we selected the most accurate agent and recorded its empirical performance in more detail. The results are displayed in Fig. 13.

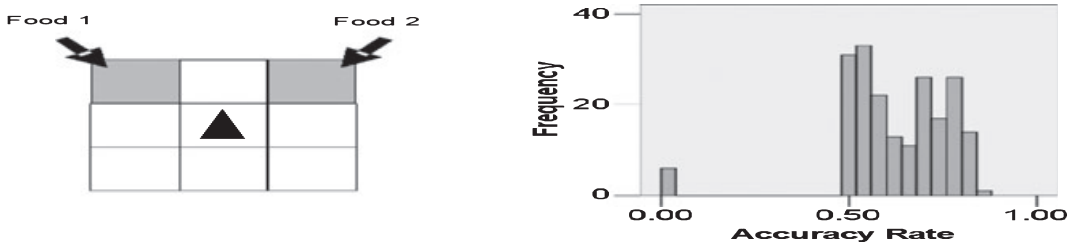

Fig. 12. (Left) The schematic structure of the comparison experiment. The agent (represented by a black triangle) is placed in the center of the mini-world, facing ''up.'' (Right) A histogram of the population performance in the quantity comparison experiment.

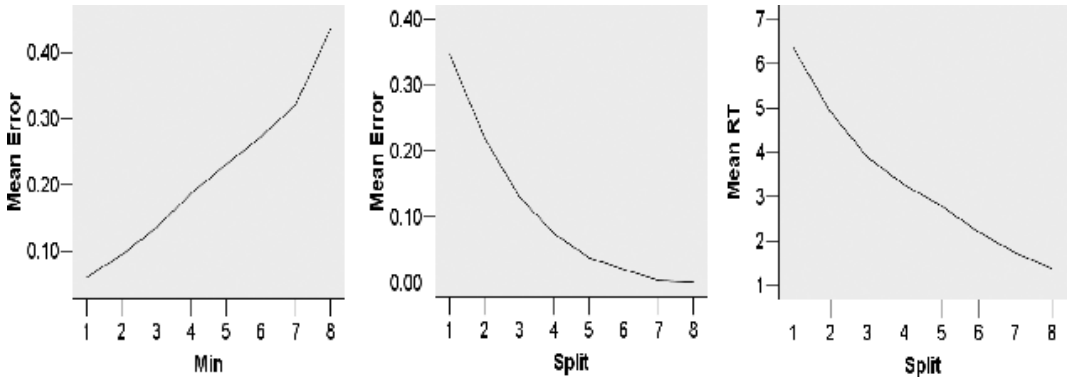

Fig. 13. Accuracy scores are rates of correct choices. (a) Mean accuracy versus minimum quantity of food (Min) in a given trial. (b) Mean accuracy versus numerical distance (Split) between food quantities. (c) Mean ''reaction time'' versus numerical distance between quantities; this latter value is the average number of processing iterations required before the agent makes a defined ''choice.''

As the minimum of the two quantities to be compared increases, there is an increase in discrimination error ($p < 0.001$, $R^2 = 0.34$, $\beta = 0.59$); this is an instance of the Size effect (Fig. 13). As the numerical distance between the quantities increases, there is a corresponding decrease in discrimination error ($p < 0.001$, $R^2 = 0.55$, $\beta = -0.75$); this is an example of the Distance effect (see Fig. 13). Strikingly, this agent also displays a Distance effect for *reaction times* ($p < 0.001$, $R^2 = 0.30$, $\beta = -0.56$; see Fig. 13), just as humans do in analogous tasks. Reaction times are defined as the number of time steps from the start of each comparison trial until the agent ''chooses'' one of the two food values. Although non-discriminating foragers can persist by sacrificing decision accuracy for decision speed, this agent is capable of reversing that trade-off, sacrificing decision speed to more reliably ''go for more.'' As Size and Distance effects drive the classical debate on the structure of quantity representation, a representational account of this agent's behavior—which seems to display those effects—should be relevant to that debate.

2.5. Searching the candidate-space

Although the original logic of the BMM is equally applicable here, the current agent raises some practical issues that demand some extensions. Where our agents in Section 1 had just seven hidden units, the best quantity-discriminator in our evolved population has 25. In the previous case, we derived our results from an exhaustive search of the agent's Candidate-space, with 2^7 lesion experiments in all; for much larger spaces of the sort we now face, this approach is impractical. The space is further enlarged because our quantity comparison problem is rather richer than was the categorical perception problem. This is, in the current case, particular units could represent either of the two numbers that the agent must compare, or (some function of) the difference between them. Each ''meaning'' that a

Table 2
Receptive fields considered in the BMM analysis of the quantity-comparison agent

Lesion Identifier	
0	No lesion
1	Unit average codes for right-hand food value
2	Unit average codes for left-hand food value
3	Unit average codes for relative difference
4	Unit average codes for absolute difference

unit's activity might have corresponds to a different tuning function that must be included in our Candidate-space. Table 2 displays the list of lesion types—or proposed unit tuning functions—that we consider. There are five values in all, so the corresponding Candidate-space contains 5^{25} items.

To search this space, we use precisely the same approach as that recruited to design the agents themselves—a genetic algorithm. When designing the agents, our search optimized a population of neural networks, whereas in this case, we optimize a population of Candidates. These Candidates are structurally similar to those employed in the last section, but different in that their cells can contain integers in the range 0–4 (rather than 0–1). The other important difference is that, where our agent was evolved to be an effective forager, our Candidates must be evolved to manipulate that agent effectively.

To achieve this goal, we first recorded the agent's comparison performance scores for every individual combination of food values (72 in all); our test included 10 repetitions of each combination and recorded the number of times that correct choices were made in each case. The result is a vector of performance scores (length = 72), associating each food combination with a score in the range 0–10. A similar list was also generated while the testing of each Candidate; in these tests, agents were always placed in an empty mini-world, and the goal was to discover Candidates that encouraged the agent to behave as if it could ''see'' particular food value combinations.

Following the logic of Section 1, we measured the correspondence between this invoked perception and natural perception by comparing the agent's behavior in each case; good Candidates should encourage choices that correspond to those made under natural conditions. We defined the fitness of each Candidate as the sum of the absolute item-by-item differences between the baseline scores and lesioned scores; the goal of the search was to find a Lesion List that minimized this ''fitness''. The calculation of fitness is described below in Eq. 5:

$$F_i = \left(\sum_{j=1}^{N} |(P_j^u - P_j^l)| \right) \quad 5$$

where P_j^u is the performance score (the number of times a correct choices was made) achieved by the agent for food value combination **j**, P_j^l is the performance score achieved

when partial informational lesions are used to simulate the presence of food value combination **j**, but no food is actually present, and **N** is the number of performance scores in each list (72).

After running the lesion-search (which appears to converge after ∼4 million iterations) and identifying the best discovered solution, one further step was required. As mentioned in Section 1, simulated evolution can yield bias units whose activity remains very close to '1' regardless of any environmental input. On closer inspection, five of the agent's units appeared to behave in this way, and two were part of the best Candidate that we could discover. As bias units do not vary, neither informational nor partial informational lesions should have any impact at all on the agent's behavior. We verified that this was the case by pruning and then retesting the solution; as no fitness costs were incurred, we omit these units in the results that follow.

Unlike in the previous section, our best discovered Candidate does not permit perfect manipulation of that agent's choice behavior. Nevertheless, the results are encouraging. To assess their quality, we employed the standard method of linear regression. The mark of a good Candidate is its ability to reproduce ''natural'' comparison choices when no food is actually present—the dependent variable for the regression is therefore the series of ''natural'' performance scores that we derived earlier. This series is an integer vector, with 72 cells (one for each food value combination), which contains integers in the range 0–10; a score of '10' indicates that the agent always chooses the larger of the two values when faced with that particular combination. The independent variable is the list of performance scores achieved when our best discovered Candidate is used to lesion the agent, and no food is actually present. In this case, a correct choice is made when the agent moves onto the square that *would have been* correct if the food that we have tried to simulate were actually present in the world. By measuring the correspondence between these two series, we are measuring the extent to which the best Candidate that we discovered has allowed us to manipulate the agent's categorization choices.

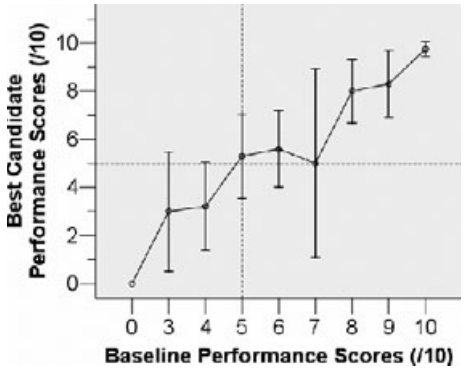

Fig. 14. Agreement between performance scores in unlesioned versus lesioned conditions for the best, identified theory of the agent's representations; error bars are *SEs* of the values (mean averages) at each point. Perfect agreement would yield a perfectly straight diagonal line, of the form '$y = x$'.

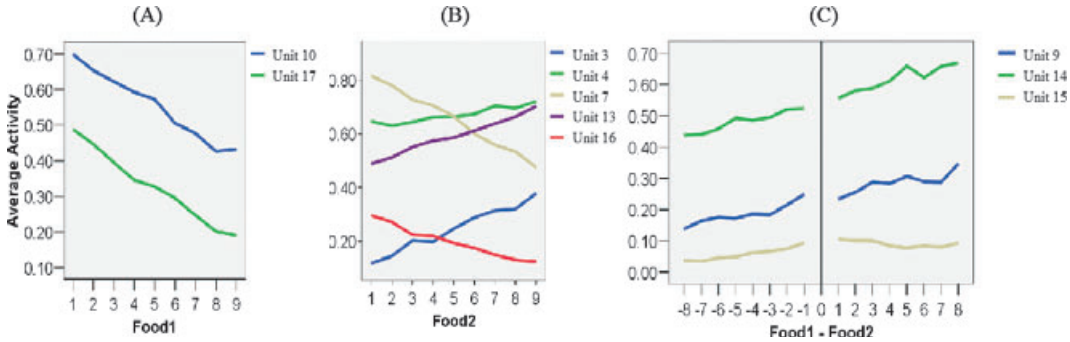

Fig. 15. Classically recognizable representations, emerging from a Dynamicist model of quantity comparison. (a) Representation of food on the agent's right. (b) Representation of food on the agent's left. (c) Representation of the difference between presented food values. Each point in each of these series corresponds to the average activity value of the specified unit in the specified circumstances.

2.6. Results

By linear regression, the relationship between the agent's baseline performance and that obtained using our best Candidate is very strong*:* $p < 0.001$, $R^2 = 0.59$. In other words, our Candidate yields performance scores that are significantly related to the baseline scores, and which account for 59% of the variation in those scores (Fig. 14). The theory itself—the best account that we have discovered of the agent's representational strategy—is graphed in Fig. 15.

Is 59% enough? Questions like this are difficult to answer conclusively. One argument in the result's favor stems from the logic of Spieler and Balota (1997), who argued that a model's item-level predictive power should be judged relative to that of the environmental features that drive the behavior of interest. In this case, the relevant features are the food values themselves, their mean and numerical distance. When these features are regressed (as independent variables) against the agent's performance scores (the dependent variable), an R^2 value of 0.74 is achieved ($p < 0.001$). That figure of '0.74' is the real target for our Candidates, which are designed specifically to capture the way those quantities constrain the agent's internal state. Our best discovered Candidate captures approximately 80% of the influence of the food values themselves on the agent's choice behavior, so cannot be lightly dismissed. The results are also clearly comparable with the proposed formats that drive the number representation debate (Fig. 8). Given that context, our results might best be described as a mental number line (MNL) based on linear single-unit accumulators (i.e., single-units accumulate activity linearly with changes in the magnitude of the number stimuli that they are claimed to encode).

2.7. Interim discussion

The goal of this section was to illustrate that, armed with the BMM, we could use a Dynamicist cognitive model to make a definite, novel contribution to the study of cognition.

Our results satisfy that goal because they speak directly to the number representation debate (Zorzi et al., 2005). Our agent is the first example of a quantity-comparison model that encodes numbers with linear single-unit accumulators, but the format is broadly consistent with the accumulator system proposed by Meck and Church (1983), as well as with the coding of Dehaene and Changeux's (1993) ''summation clusters'' (which precede their numerosity detectors) and with the ''numerosity coding'' scheme proposed by Zorzi and Butterworth (1999). Moreover, there are at least three reasons to prefer the Dynamicist approach in this case—first, because it integrates the skill under study with a more natural (if very simplified) ''behavioral context''; second, because it encourages the problem-driven emergence of both empirical phenomena (Size and Distance effects) and representational strategies at the same time; and third, because it offers the opportunity to explore the role of phylogeny in cognitive development.

Finally, our results appear to enjoy some reasonably direct empirical support. Neurophysiological studies have just begun to tackle the issue of the neuronal correlates of number representations using single-cell recording in behaving monkeys. Nieder et al. (2002) have described ''number neurons'' in the monkey brain with tuning functions that fit the logarithmic coding of Dehaene and Changeux's (1993) ''numerosity detectors.'' This finding would seem to be at odds with the type of coding employed by our agent. However, a different type of ''number neuron'' with tuning properties that are startlingly similar to those employed by our agent (single-cell linear accumulators) has been recent discovered by Roitman et al. (2007) in the lateral intraparietal cortex of monkeys engaged in a numerosity comparison task. After averaging the spike rates recorded over a few hundred milliseconds from single, number-sensitive neurons—a process analogous to our own use of circumstance-dependent unit averages—the authors showed that these neurons encode the total number of elements within their receptive fields in a graded fashion. We were not aware of this work while implementing the model that this section reports—nevertheless, these data provide a huge boost to the confidence that we can attach to it. Moreover, the same neural coding strategy—a monotonic sensitivity to an increase of a particular feature dimension—has been shown to apply to other sensory domains (e.g., the frequency of vibrotactile stimulation; Romo & Salinas, 2003).

3. General discussion

We began by discussing the analytical challenge that Dynamicist cognitive models pose. Section 1 demonstrated that challenge by applying two well-known analytical methods to a familiar Dynamicist model, then illustrated a novel response to that challenge in the form of the BMM. Section 2 extended that response and applied it to a novel dynamical model of number comparison—yielding results that make a direct contribution to the debate concerning the format of neural number representations. Our BMM seems at least useful enough to warrant further study.

As a starting point, we chose to focus our analyses on average unit activities, trading explanatory power for interpretative clarity; this choice can be criticized because it implies

a very restricted role for network units. One defense for the choice is that the restriction does not prevent us from discovering encouraging results; we achieved perfect manipulation of an agent's categorical perception performance, and reasonable manipulation of an agent's quantity comparison performance. Even accepting the logic of the BMM, the former result is surprising—cognitive theories rarely aim to capture every detail of the performance under study. Moreover, nothing in the BMM's logic precludes the use of different primitives, such as time-period dependent means, average rates of change, or even centroid time series. Future extensions of this work could consider some or all of these alternatives without departing from the BMM's fundamental logic.

Another key direction for future research concerns the method's applicability to larger-scale models, with richer internal dynamics. Although the current results do at least appear to say something useful about biological information processing, analyses of more complex models may be able to say still more. Work of this sort could also address the concern that the precise form of our results in Section 2 owes more to the details of the artificial ecosystem than it does to a more general connection with the pressure to forage effectively. This kind of connection (between the definition of the sensors and ''evolved'' representations) is probably unavoidable—indeed, its significance is also assumed in the way that researchers employ the statistics of real sensory stimuli to decode the tuning functions of biological neurons (e.g., Atick, 1992)—but the concern could certainly be allayed by results garnered from a broader range of models.

Yet despite its limits, the current work does at least illustrate that the BMM can be useful—that we can use it to generate novel insights into dynamical cognitive models, and results that contribute to our evolving understanding of cognition. By changing the way we interpret the internal dynamics of models engaged in cognitive tasks, work like this can also begin to change the way we think about the connections between cognitive behavior and biological neural dynamics. That goal may still be some way away, but the current work takes what we hope is a useful step toward it.

Acknowledgments

This research was supported by grants from the European Commission (MRTN-CT-2003-504927 NUMBRA) and the European Research Council (210922-GENMOD) to MZ.

Notes

1. This number can vary from trial to trial, because trials can end before a shape touches the *x* axis—when they are ''caught'' by an agent—and because shape size can vary.
2. Available on request; see http://www.cns.tau.ac.il/msa/
3. Note that PCA is formally very similar to multidimensional scaling methods that have gained currency in recent years (e.g., Botvinik & Plaut, 2004); our criticisms of the former should therefore apply equally well to the latter.

4. One important difference is that Uller et al. (2003) exclude trials in which their salamanders fail to choose one option after a maximum length of time—in our method, these ''misses'' (failure to choose after 100 iterations) are treated as incorrect choices.
5. Although lateral asymmetry is a consistent feature of the behavior of agents evolved in this system, its direction is not consistent—some runs yield agents that prefer left-sided food.

References

Aharonov, R., Segev, L., Meilijson, I., & Ruppin, E. (2003). Localization of function via lesion analysis. *Neural Computation*, *15* (4), 885–913.

Atick, J. (1992). Could information theory provide an ecological theory of sensory processing? *Network*, *3*, 213–251.

Beer, R. (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence*, *72* (1–2), 173–215.

Beer, R. D. (1996). Toward the evolution of dynamical neural networks for minimally cognitive behavior. In P. Maes, et al. (Eds.), *Proceedings of the 4th international conference on the simulation of adaptive behaviour* (pp. 421–429). Cambridge, MA: MIT Press.

Beer, R. D. (2000). Dynamical approaches to cognitive science. *Trends in Cognitive Sciences*, *4* (3), 91–99.

Beer, R. D. (2003). The dynamics of active categorical perception in an evolved model agent. *Adaptive Behavior*, *11* (4), 209–243.

Biro, D., & Matsuzawa, T. (2001). Use of numerical symbols by the chimpanzee (Pan troglodytes): Cardinals, ordinals, and the introduction of zero. *Animal Cognition*, *4* (3), 193–199.

Brannon, E. M., Wusthoff, C. J., Gallistel, C. R., & Gibbon, J. (2001). Numerical subtraction in the pigeon: Evidence for a linear subjective number scale. *Psychological Science: A Journal of the American Psychological Society/APS*, *12* (3), 238–243.

Butterworth, B. (1999). *The mathematical brain*. London: Macmillan.

Dehaene, S. (2000). *The number sense: How the mind creates mathematics?* New York: Oxford University Press.

Dehaene, S., & Changeux, J. (1993). Development of elementary numerical abilities: A neuronal model. *Journal of Cognitive Neuroscience*, *5* (4), 390–407.

Durgin, F. H. (1995). Texture density adaptation and the perceived numerosity and distribution of texture. *Journal of Experimental Psychology: Human Perception and Performance*, *21* (1), 149–169.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7* (2–3), 195–225.

Feigenson, L., Carey, S., & Hauser, M. (2002). The representations underlying infants' choice of more: Object files versus analog magnitudes. *Psychological Science: A Journal of the American Psychological Society/APS*, *13* (2), 150–156.

Friston, K., Worsley, K., Frackowisk, R., Mazziotta, J., & Evans, A. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, *1*, 214–220.

Gallistel, C., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44* (1–2), 43–74.

Gallistel, C. R., & Gelman, R. (2000). Non-verbal numerical cognition: From reals to integers. *Trends in Cognitive Sciences*, *4*(2), 59–65.

van Gelder, T. (1995). What might cognition be, if not computation? *The Journal of Philosophy*, *92* (7), 345–381.

van Gelder, T. (1998). The dynamical hypothesis in cognitive science. *Behavioral and Brain Sciences*, *21* (05), 615–628.

Gonzalez, R. C., & Richard, E. (1992). *Woods, digital image processing*. Reading, MA: Addison Wesley.

Harvey, I. (2001). Artificial evolution: A continuing SAGA. In T. Gomi (Ed.), *Evolutionary robotics. From intelligent robotics to artificial life* (pp. 94–109). Heidelberg, Germany: Springer.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America*, *96* (13), 7592–7597.

Jordan, K., & Brannon, E. (2006a). Weber's Law influences numerical representations in rhesus macaques (*Macaca mulatta*). *Animal Cognition*, *9* (3), 159–172.

Jordan, K. E., & Brannon, E. M. (2006b). A common representational system governed by Weber's law: Nonverbal numerical similarity judgments in 6-year-olds and rhesus macaques. *Journal of Experimental Child Psychology*, *95* (3), 215–229.

Keinan, A., Hilgetag, C. C., Meilijson, I., & Ruppin, E. (2004). Causal localization of neural function: The Shapley value method. *Neurocomputing*, *58–60*, 215–222.

Keinan, A., Meilijson, C. C. H. I., Ruppin, E., Hilgetag, C. C., & Meilijson, I. (2003). Fair attribution of functional contribution in artificial and biological networks. *Neural Computation*, *16*, 1887–1915.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, *2*, 387–395.

Meck, W. H., & Church, R. M. (1983). A mode control model of counting and timing processes. *Journal of Experimental Psychology. Animal Behavior Processes*, *9* (3), 320–334.

Miller, A. L., & Baker, R. A. (1968). The effects of shape, size, heterogeneity, and instructional set on the judgment of visual number. *The American Journal of Psychology*, *81* (1), 83–91.

Moyer, R. S., & Landauer, T. K. (1967). Time required for Judgements of numerical inequality. *Nature*, *215* (5109), 1519–1520.

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science*, *297* (5587), 1708–1711.

Nowak, A. (2004). Dynamical minimalism: Why less is more in psychology. *Personality and Social Psychology Review*, *8* (2), 183–192.

Oja, E. (1989). Neural networks, principal components, and subspaces. *International Journal of Neural Systems*, *1* (1), 61–68.

Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP + model of reading aloud. *Psychological Review*, *114* (2), 273–315.

Plaut, D. C., & Shallice, T. (1991). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*, 377–500.

Port, R. F., & Gelder, T. V. (1998). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA: MIT Press.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhyā: The Indian Journal of Statistics, Series A*, *26* (4), 329–358.

Roitman, J. D., Brannon, E. M., & Platt, M. L. (2007). Monotonic coding of numerosity in macaque lateral intraparietal area. *PLoS Biology*, *5* (8), e208.

Romo, R., & Salinas, E. (2003). Flutter discrimination: Neural codes, perception, memory and decision making. *Nature Reviews: Neuroscience*, *4* (3), 203–214.

Seth, A. (2008). Causal networks in simulated neural systems. *Cognitive Neurodynamics*, *2* (1), 49–64.

Shallice, T. (1988). *From neuropsychology to mental structure*. Cambridge, UK: Cambridge University Press.

Shapley, L. (1953). A value for n-person games. In H. W. Kuhn & A. W. Tucker (Eds.), *Contributions to the theory of games* (pp. 307–317). Princeton, NJ: Princeton University Press.

Spieler, D., & Balota, D. (1997). Bringing computational models of word naming down to the item level. *Psychological Science*, *8*, 411–416.

Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action* (p. 408). Cambridge, MA: MIT Press.

Tomko, G. J., & Crapper, D. R. (1974). Neuronal variability: Non-stationary responses to identical visual stimuli. *Brain Research*, *79* (3), 405–418.

Uller, C., Jaeger, R., Guidry, G., & Martin, C. (2003). Salamanders (Plethodon cinereus) go for more: Rudiments of number in an amphibian. *Animal Cognition*, *6*, 105–112.

Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: A neural model. *Journal of Cognitive Neuroscience*, *16* (9), 1493–1504.

Watson, R. A., & Pollack, J. B. (2001). *Coevolutionary dynamics in a minimal substrate. I, GECCO-01* (pp. 702–709). San Francisco, CA: Morgan Kaufmann.

Zorzi, M., & Butterworth, B. (1999). A Computational model of number comparison. In *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 1–4). Vancouver, British Columbia: Lawrence Erlbaum Associates.

Zorzi, M., Stoianov, I., & Umiltà, C. (2005). Computational modeling of numerical cognition. *Handbook of Mathematical Cognition*, *5*, 67–83.