Psychological Review (in press)
Revised June 14, 1999.

An Evaluation of a Computational Model of Lexical Access: Comments on Dell et al. (1997)

Wheeler Ruml and Alfonso Caramazza
Harvard University

Abstract

We evaluate the computational model of lexical access proposed by Dell, Schwartz, Martin, Saffran, and Gagnon (1997). They argue that fits of their model to naming data obtained from normals and brain-damaged patients support assumptions regarding interactivity in the lexicon, global damage in aphasia, and continuity between normal and aphasic naming behavior. Our investigation reveals that the model fits the empirical data poorly and that the claims Dell et al. make on the basis of the model's performance would not follow even if the model were accurate. Although we improve the model's fit using a novel automatic regression procedure, it cannot account for five of Dell et al.'s twenty-one patients (24%), and we show that its limitations are inherent in its design. We argue that claims such as those made by Dell et al. can only be addressed by considering evidence from multiple related tasks and by comparing multiple computational models.

(Caution: This is a preprint.
Please check any quotations against the published version.)

When one names an object, the semantic specification of the name of the object is mapped to the sequence of phonemes that is that word's pronunciation. This cognitive process, known as *lexical access*, is quick, automatic, and usually successful, but very occasionally it can result in slips of the tongue: groups of phonemes representing words that are

---

Wheeler Ruml, Division of Engineering and Applied Sciences, Harvard University; Alfonso Caramazza, Cognitive Neuropsychology Laboratory, Harvard University.

We would like to give special thanks to Gary Dell for his helpful comments and suggestions regarding this work, and for providing us with the source code for his implementation of his model. We also thank Deborah Gagnon for sending us detailed patient naming data, and Trevor Harley and an anonymous reviewer for their useful comments. We are indebted to Michele Miozzo, the Harvard Cognitive Neuropsychology Laboratory, Stuart Shieber, and the Harvard AI Research Group for insightful suggestions and stimulating discussions regarding this research. Support was provided in part by National Science Foundation grants IRI-9350192 and IRI-9618848 and National Institutes of Health grant NS-22201.

Please address correspondence concerning this article to Wheeler Ruml, `ruml@eecs.harvard.edu`, Maxwell Dworkin Laboratory, Harvard University, 33 Oxford Street, Cambridge, MA 02138.

semantically or phonologically related to the intended word, or even an unrelated word or gibberish. Studies of naming errors from spontaneous speech and controlled experimental tasks have revealed patterns that should be accounted for by theories of the cognitive mechanisms used by the human brain to perform lexical access (Levelt, 1989).[1] In this paper, we will evaluate the lexical access model proposed by Dell, Schwartz, Martin, Saffran, and Gagnon (1997), which was designed specifically to explain errors during picture naming and word repetition.

Dell et al. present their theory as a computational model, allowing them to compare their model's precise numerical predictions to experimental data. The model is part of a family of related theories proposed by Gary Dell and his collaborators to explain a wide variety of phenomena observed in speech production (Dell, 1986, 1988, 1990; Dell, Juliano, & Govindjee, 1993; Dell & O'Seaghdha, 1991; Martin, Weisberg, & Saffran, 1989). We distinguish here between the abstract theoretical principles which have remained relatively constant over all these models, and the particular implemented computational models themselves, which, to date, have each been evaluated only on specific individual tasks. The model we consider here, for example, only accounts for error frequencies during picture naming, leaving aside such data as substitution errors during spontaneous speech. If this particular model should prove accurate, there is hope that it can be integrated with the others to make progress towards a single unified theory of language production.

Dell et al.'s theory of lexical access makes three particularly interesting and controversial claims. First, it postulates extensive interaction between representations, accomplished by explicit bidirectional communication. As in many other models of lexical access, Dell et al. specify a semantic input representation, a phonological output representation, and a mediating lexical representation. (There is no division into modality-neutral and modality-specific lexical representations, such as the lemmas and lexemes in the theories of Levelt (1992), Roelofs (1992), or Bock and Levelt (1994).) However, unlike models in which each representation only passes information to the next (as in the discrete stage model of Roelofs (1992) and the cascading models of Humphreys, Riddoch, and Quinlan (1988) and Caramazza (1997)), each level in Dell et al.'s model is constantly interacting with its adjacent levels. Since representations in Dell et al.'s model communicate by the spreading of activation, this claim is represented in the model by the spread of activation both forwards (from the semantic level towards the phonological level) and backwards simultaneously. Unlike the models of Harley (1993) or Rapp and Goldrick (in press), which postulate feedback only from the phonological level, Dell et al.'s model has feedback from both the phonological and lexical levels.

The second claim arises because the theory is designed to explain the lexical access process not only of ordinary people, but also of people who have suffered brain damage. To do this, Dell et al.'s model of aphasic naming must make claims not only about the functional mechanism of lexical access, but also about how it is that the mechanism is

[1] Other data that have been used to inform theories of lexical access include investigations of the time course of lexicalization in naming tasks (e.g., Schriefers, Meyer, & Levelt, 1990), the tip-of-the-tongue phenomenon (e.g., Burke, MacKay, Worthley, & Wade, 1991; Miozzo & Caramazza, 1997), the distribution of hesitation patterns in spontaneous speech (e.g., Butterworth, 1979; 1980), and the patterns of naming deficits and other production disorders in aphasic patients (e.g., Butterworth, 1992; Badecker, Miozzo, & Zanuttini, 1995; Garrett, 1992). See Caramazza (1997) for a review.

damaged so as to result in impaired performance. Rather than requiring different levels of damage to different parts of the lexical access system, Dell et al. propose that the different patterns of naming errors of fluent aphasic patients can be explained by the assumption that damage affects all levels equally. They call this claim *the globality assumption*, and it is represented in the model by two parameters which alter the flow and maintenance of activation in all parts of the model. The model therefore incorporates both a theory of the lexical access system and a theory of brain damage.

The third central assumption that Dell et al. propose is what they call *the continuity thesis*. This is the claim that the spectrum of aphasic naming, from mild to severe, can be characterized as a continuum from normal performance at one end and the random error opportunities afforded by the lexicon at the other end. The severity of each case places that individual somewhere along that spectrum. Dell et al. claim that both the patient data and their model's results conform to this characterization.

Dell et al. support these three claims by using a simulation of their model to reproduce the patterns of naming errors observed in individual aphasic patients. Each pattern consists of the frequency of the five types of errors: semantic, phonological, mixed (both semantically and phonologically related to the target), unrelated, and nonword. Dell et al. argue that success in reproducing the error profiles observed in the patients would constitute support for the assumptions of the model:

> The good fit between the patient data and the model suggests three conclusions. First, it extends support for the interactive two-step approach to naming. A model that successfully characterized normal performance could be applied to the range of performance that fluent aphasic individuals exhibit. Although only a restricted set of error patterns is allowed by the model, the patients' patterns appeared to fall within that set. Second, the good fit supports the continuity thesis. A large component of disordered naming can be linked to general severity. More severe aphasic patients have an error pattern that is closer to the error opportunities afforded by the lexicon, whereas less severe aphasic patients have a pattern that is similar to the normal pattern. Finally, the fit supports the hypothesis that variation in patient error patterns can be associated with global lesions in activation transmission, representational integrity, or both. (p. 820)

In short, Dell et al. argue that since the model embodies the three claims, then if the model fits the data, we have support for the claims.

We will evaluate Dell et al.'s argument in three different ways. First, we can evaluate the empirical evidence for Dell et al.'s model. Although Dell et al. claim at several points in their paper that "the fit between the model and patients was good" (p. 819), our close inspection of their analysis will show that they do not actually present any evidence that their model's fit is in fact good. Since their claims are based on the fit of the model to the data, it is essential to evaluate the model's empirical performance. We will evaluate its fit using three different evaluation metrics, and to improve the model's accuracy, we will use a novel numerical optimization procedure. However, we will see that the model cannot fit five of Dell et al.'s twenty-one patients (24%), and performs worse than a simple mathematical model of naming.

Second, we can also ask whether there are any other known patterns of naming errors

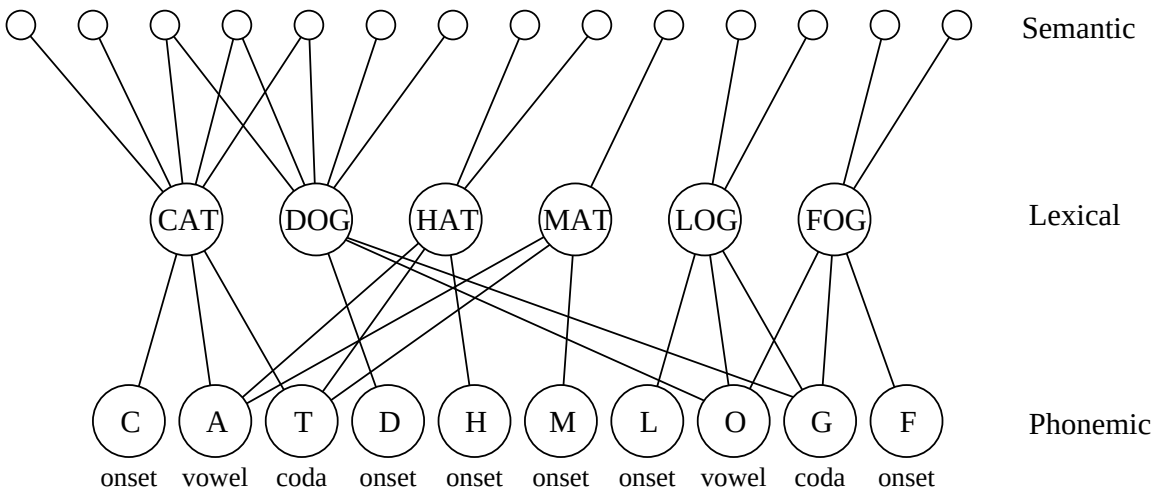

Figure 1. The structure of part of the core model.

that contravene the predictions of the model. Although Dell et al. show fits of their model to data from twenty-one patients, they do not provide a comprehensive or precise analysis of the range of patients that the model is capable of providing predictions for. It is not clear that it can handle error patterns other than those exhibited by the specific patients they tested. By plotting the results of exhaustive simulations, we will gain an understanding of the global behavior and scope of the model, discovering that it cannot model well-known patient patterns.

Finally, we can evaluate the logical force of Dell et al.'s argument. Even if their model provided good fits to patient data and were consistent with other sources of empirical evidence, it is not obvious that one would be compelled to adopt the conclusions they reach. As we will see, Dell et al.'s claims would require evidence from multiple tasks and multiple computational models, and could not follow from the single study they conducted.

But before examining Dell et al.'s claims in detail, we will begin with an overview of the model.

The Model

Dell et al.'s model comes in two parts: a core model of phoneme retrieval, and then supplemental parts that use the core model to predict other behaviors, such as word repetition performance.

The Core Model.

The core model takes as input a semantic specification of a word, and returns an ordered list of phonemes that is intended to represent the word's pronunciation. It is based on a theory of lexical access first proposed by Dell in 1986. The model is structured in a localist connectionist style, as illustrated in Figure 1. The model involves three kinds of structures: nodes, links, and layers. There are three layers in the model, and they are ordered from highest (semantics) to lowest (phonemes). The middle layer is intended to correspond to lexical entries. Each layer contains a set of nodes and a possibly-empty

ordered list of selected nodes. Each node has a type (semantic, lexical, or phonemic), an activation level (a real number), and a set of links to certain other nodes in other layers. All links are bidirectional (if a is linked to b, then b is also linked to a). Each layer contains nodes of only one type, and nodes in a given layer connect only to nodes in adjacent layers. Nodes of type phoneme also have subtypes corresponding to syllabic position (onset, vowel, or coda).

Processing occurs in the model by the updating of each node's activation level according to a weighted sum of the node's current activation level and those of its neighbors. A small amount of noise is also added to each node, some of which is proportional to its current activation level, and the rest of which represents an absolute level of ambient noise. More formally, if $a_t(m)$ represents the activation at time t of a particular node m with neighbors N, and $R(x)$ represents a random sample drawn from the normal distribution with mean zero and standard deviation x, and *decay*, *connection*, *intrinsic*, and *activation* are parameters of the model, then

$$a_{t+1}(m) = \textit{old-activation} + \textit{incoming-activation} + \textit{noise}$$

where

$$\begin{aligned} \textit{old-activation} &= (1 - \textit{decay}) \times a_t(m) \\ \textit{incoming-activation} &= \textstyle\sum_{n \in N}(\textit{connection} \times a_t(n)) \\ \textit{noise} &= R(\textit{intrinsic}) + (R(\textit{activation}) \times a_t(m)). \end{aligned}$$

Note that activation is not necessarily conserved—the number of neighbors a node has does not influence the amount of activation they acquire. This means that the total amount of activation in the network depends on the number of neighbors of the active nodes. Input is given to the model by raising the activation level of the semantic nodes connected to the target word to an arbitrary constant. Activation then spreads according to the equation given above for eight time steps. After this time, the lexical layer chooses its most active node and sets that node's activation level to a constant (ten times the initial semantic activation). This jolt is intended to correspond to lexical selection. After eight more time steps, the phoneme layer then selects the most active phoneme node of each of the three position subtypes (onset, vowel, and coda). This ordered list represents the output of the model, which can then be classified as correct, semantically related to the target, phonologically related, related in both ways (a mixed error), unrelated but representing an actual word, or gibberish (a nonword error). The activation levels of all nodes are then set to zero, preventing any influence of spurious activation from one trial on the next.

There are more semantic nodes in the actual network structure than are depicted in Figure 1. Each of the six lexical nodes in the model connects to exactly ten semantic nodes, although the two nodes representing semantically related words (cat and dog) share three of their semantic nodes. The target word for every trial is cat. Every tenth lexical access simulation is run using a slightly different network structure in which, instead of there being two words phonologically related to the target (hat and mat), there is only one (mat) and the sixth word is now a possible mixed error (rat). This arrangement is motivated by concerns regarding the random error probabilities of the simulation—see Dell et al. (1997) for details. All the parameters of the core theory and the values used by Dell et al. are summarized in Table 1.

Table 1: Parameters in the core theory and their default values.

Parameter	Description	Default value
Nodes	the number of nodes in the semantic, lexical, and phonological layers	57, 6, and 10
Connectivity	the nodes each node connects to	see Figure 1 and text
Connection strength	the coefficient by which a given node's neighbors' activation levels are multiplied during spreading	0.1
Decay rate	the coefficient by which a given node's activation is multiplied during spreading	0.5
Semantic jolt	the activation level to which semantic nodes are set to represent the model's input	10
Lexical jolt	activation level to which the selected lexical node is set	100
Spreading steps	the number of time steps for which activation is spread through the network before lexical selection or phoneme selection takes place	8
Intrinsic noise	standard deviation of the distribution of activation-independent noise	0.01
Activation noise	standard deviation of the distribution of the noise that is proportional to a node's activation level	0.16

An example run of the simulation is shown in Figure 2, using the network structure containing a mixed error (rat). The activation levels of only a few of the nodes are shown. The node corresponding to the word cat is the most active lexical node at the eighth time step, and hence receives the increase in activation corresponding to lexical selection. After eight more time steps, it happens that the most active onset phoneme is K, the most active vowel is AE, and the most active coda is T, so the trial is scored as a correct response. The legend in the figure indicates the rank order of the nodes' activation levels at the end of the run.

Because of the influence of the random noise, each simulation of the core model behaves slightly differently, and may result in any of the six possible outcomes (correct, semantic, phonological, mixed, unrelated, nonword). By running the simulation multiple times, one can estimate the probability distribution over the six response types. This pattern can be compared with probabilities estimated from experiments with humans. If the model can replicate the human error probabilities, then insofar as the data summarize language production behavior, the model can be said to exemplify a mechanism sufficient to simulate human behavior.

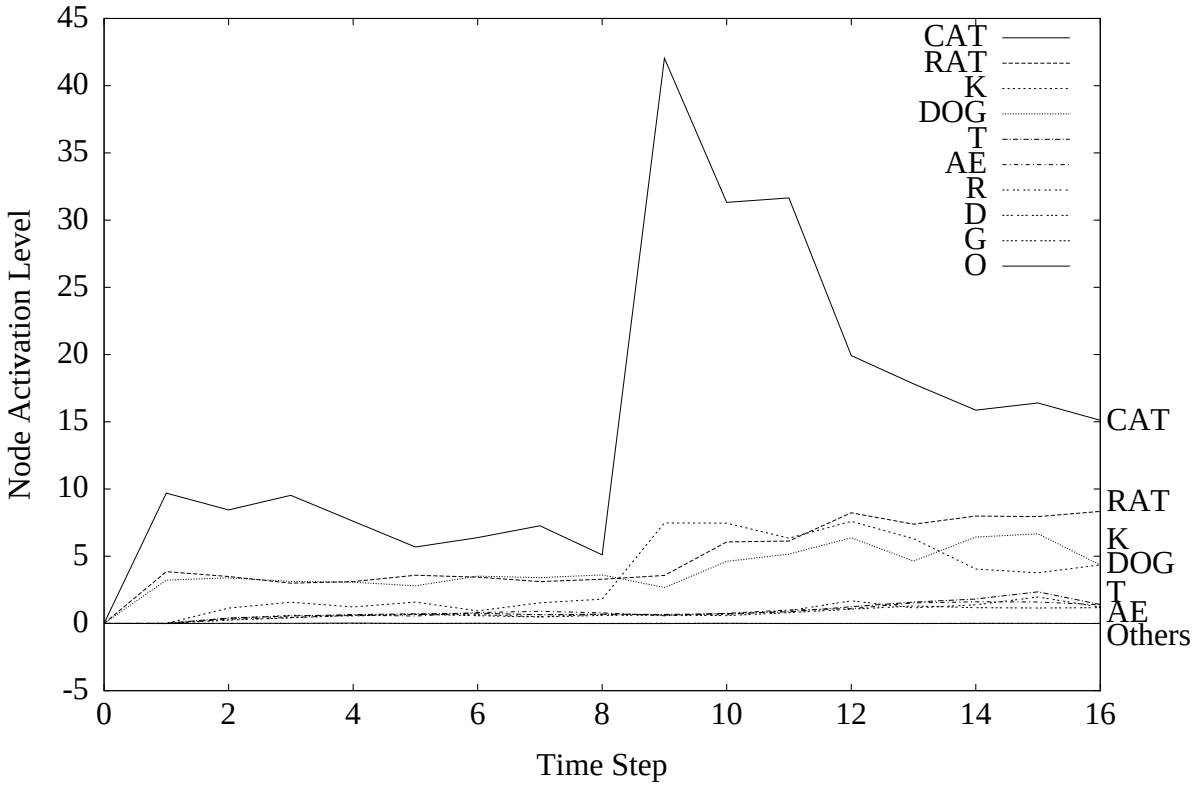

Figure 2. Activation levels of some of the nodes in the model during a simulation run.

Using the Core Model.

By varying the values of *connection* and *decay*, the distribution of errors of the core model can be changed. This allows the replication of a variety of error patterns. The task of matching a particular patient's performance is essentially a non-linear regression problem, using *connection* and *decay* as a two-parameter summary representation for an error pattern. One seeks the values for the parameters such that the resulting distribution of the core model's errors most closely matches the data from the given patient. Speaking more formally, we can try to minimize the X^2 statistic of a χ^2 test to detect a difference between the patient's distribution and the model's. If P represents the distribution of the patient's responses over n possible response categories, and M represents the distribution of the responses generated by the model, then this statistic is calculated as

$$X^2(M,P) = \sum_{j \in \{M,P\}} \sum_{i=1}^{n} \frac{(observed_{ij} - expected_{ij})^2}{expected_{ij}}$$

where

$$expected_{ij} = total_j \times \frac{M_i + P_i}{total}.$$

Varying *connection* and *decay* to minimize X^2 will maximize the probability that the model matches the patient (given the observed samples and assuming an underlying multinomial).

Although Dell et al. optimize the fit by hand, this step is mathematically well-specified, and can be carried out by an appropriate automatic numerical optimization method, as we will demonstrate in the next section.

Note that this use of the core model of picture naming does not itself provide any specific predictions, since the two input parameters do not correspond to any psychologically measurable quantities or stimuli. Although Dell et al. refer to the distributions produced by the model as predictions, their model only makes the general claim that any patient naming pattern can be represented by some setting of *decay* and *connection*. In addition to this general claim, Dell et al. use the fitted core model parameters for each patient to derive specific predictions about additional aspects of each patient's performance. We will defer discussion of those predictions until later in this paper.

In this paper, we are not concerned with evaluating the assumptions of the underlying general theory of lexical access (Dell, 1986). We merely wish to test the specific model of aphasic naming presented by Dell et al. (1997) on the tasks for which it was proposed. We will first consider the model's general ability to simulate patients' naming error patterns.

Replicating Picture Naming Performance

Evaluation Methodology

As we discussed above, the central task of Dell et al.'s core model is to reproduce patient naming behavior. By allowing different values of *connection* and *decay*, the model functions as a two-parameter summarization of a patient's response distribution during picture naming. There are many different ways to assess the model's performance at this task, and we will consider three of them, ranging from formal numerical measures to graphical visualizations.

First, we will formally test the model's ability to replicate observed patient error patterns. Since Dell et al. argue that their theoretical claims rest on the empirical ability of the model to match patient response distributions, it is critically important to assess the fit of the model to the patient data. If even one fluent aphasic patient could not be fit by Dell et al.'s model, this would indicate at least one incorrect assumption regarding a relevant aspect of lexical access, limiting the model's ability to function as a basis for theoretical claims. Without a quantitative assurance of fit, one cannot use the model's performance as a basis for further inferences.

Although necessary for evaluating Dell et al.'s argument, measuring the goodness of fit to patients doesn't provide a way to judge the model's performance relative to the difficulty of the task. Therefore, we will also examine the model's fitting abilities relative to two simple mathematical theories. This will give us a sense of the general difficulty of the modeling task, and whether Dell et al.'s cognitive theory is performing well or poorly in a wider sense.

Finally, we will present plots of the range of patients that the model can account for. This visual evaluation will enable us to spot patterns in the model's behavior and compare those trends with the patient data. While less formal than a statistical evaluation, such a display will allow us to understand the model's performance more intuitively. Although none of these evaluation methods are entirely satisfactory on their own, by considering all

three we should obtain a comprehensive view of the model's ability to account for patient picture naming.

Evaluating Dell et al.'s Fits

We begin by evaluating the fit of Dell et al.'s model to their patient data. Although, as we discussed above, Dell et al. fitted the model by minimizing the X^2 statistic, they do not report the resulting X^2 values or analyze the model's goodness of fit using a standard χ^2 test. Instead, they present their results using root mean squared deviation (RMSD), which is a measure of similarity between distributions. Given the two probability distributions M and P over a set of n response types,

$$\mathrm{RMSD}(M,P) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(M_i - P_i)^2} \quad .$$

RMSD values can range from zero through one, where a value of zero implies identical distributions. Note that, while X^2 is calculated on the basis of the number of samples in each category, RMSD is based on the probability of a response in each category. This makes it insensitive to the number of samples taken. While RMSD seems to capture the average magnitude of discrepancies between the model and patient probabilities in an intuitive way, it is not clear how to use it as a test of the goodness of fit. Such a test would require a way of calculating the probability of obtaining the given RMSD value under the assumption that the two different sample distributions were both drawn from the same fixed underlying distribution. Presumably, one would need to know the number of samples taken in each of the two groups, information which the RMSD statistic does not directly capture.

Since it is difficult to assess goodness of fit using RMSD directly, Dell et al. try to interpret the quality of their fits by comparing against fits to random data (p. 819). They constructed ten sets of six random numbers (normalized to sum to one), fit the core model to each set, and then calculated the RMSD between the model and the data for each of those ten fits. Then they compared the median RMSD of their fits of the random data (0.220) to the median RMSD of their fits to the patient data (0.026). Although this comparison suggests that the model fits human data better than it fits random data, it does not tell us how well the model is fitting either set. It could well be the case that both fits are poor.

Dell et al. also present an analysis of the deviation of the model's response patterns from random error opportunities (p. 819–820, Figures 9–11). Although they repeat the analysis for the patient data, it is not clear whether one can compare the two deviations in a quantitative way. The analyses don't address the quality of the model's fits, since they do not quantify the significance of discrepancies between the deviation of the model and the deviation of the patients. Some of the average discrepancies between model and data in the figures appear to be the size of an entire unit on their logarithmic scale, which corresponds to a factor of 2.7 times. It appears that nowhere in their paper do Dell et al. provide evidence that their model has a good fit to patient data, even though their argument depends critically on the match.

The most straightforward way to evaluate the model's fit is to use a standard χ^2 test of independence between each patient distribution and the distribution of the fitted model. This is easily done by computing the significance of the X^2 value of each fit, which yields

the probability that we would be incorrect in claiming that available data shows that the model's distribution differs from the patient's. One common objection to such a test is that, with enough samples, one is almost certain to detect subtle and theoretically unimportant differences between two distributions. This problem cannot arise in our situation, since we have only 175 samples for each patient (one response for each stimulus in the Philadelphia Naming Test (Roach, Schwartz, Martin, Grewal, & Brecher, 1996)). Although we can shrink the region of uncertainty surrounding the model's distribution by simulating more trials, the uncertainty regarding the patient's distribution is fixed, and will always allow many similar distributions to appear as likely matches.

A naive application of the χ^2 test would quickly show that nineteen of Dell et al.'s twenty-one patients (90%) could not be fit by the model. This is because all but two of them (L.B. and H.B.) gave at least one picture-naming response that did not fall into the well-defined response categories (correct, semantic, phonological, mixed, unrelated, or nonword). Responses such as descriptions or refusals to name would fall into this seventh 'miscellaneous' category. Since the core model cannot generate such responses, it cannot match patients who make them. Any distribution the model can produce will, after enough samples have been taken, be judged highly unlikely to be identical to any patient distribution that contains such a response. To alleviate this problem, we must ignore uncodable responses, and treat them as outside of the domain of the model. This means that we should calculate X^2 values based only on the six response categories that can be generated by the model. (It seems that Dell et al. followed a similar approach by ignoring miscellaneous responses in their RMSD calculations, although they continued to include them when converting counts to percentages.)

We implemented Dell et al.'s model and tested its fit to their patient data. The patient distributions were recovered from the probabilities reported by Dell et al. by multiplying by 175, the number of stimuli in their naming task, and then rounding. For each patient, we simulated 10,000 trials of the model using the values of *connection* and *decay* they recommended for that patient. Results of a χ^2 analysis are shown in Table 2, calculated assuming five degrees of freedom.[2] For each patient, the table compares the response probabilities of the patient and the model, and lists the RMSD and X^2 value of the fit. The significance value (p) represents the probability that we would have obtained a match of this quality if the model and patient had had the same response distribution. A low value therefore indicates a significant mismatch between the model and a patient. Confidence intervals on the X^2 values and their associated significance levels were derived using a Monte Carlo procedure described in Appendix A.

Twelve of the twenty-one fits (57%) do not match ($p < 0.05$) the corresponding patients. Nine of the fits are very poor ($p < 0.01$). This is surprising, given the emphasis that Dell et al. place on the accuracy of their model's fit to the data. An examination using the statistic they used to present their data, RMSD, shows that it does not provide a reliable indication of the goodness of fit. The fit for patient J.Fr., for instance, has a lower RMSD than the fit for patient V.C. (.013 as opposed to .020), even though J.Fr.'s fit is much worse ($X^2 = 18.4$, $p = .002$ as opposed to $X^2 = 5.7$, $p = .334$). A test using

[2]One could argue that, since we generate the model's distribution by estimating the two parameters *connection* and *decay* from the patient data, there are actually only three degrees of freedom instead of five. We will err on the lenient side, perhaps rejecting fewer non-matches then is warranted.

Table 2: Fits of Dell et al.'s model to their patients, using the parameters they suggest.

Patient and	Naming response						Fit				
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	X^2 conf.	p	p conf.
W.B.	.94	.02	.01	.01	.00	.01					
conn .0200, *dec* .56	.93	.04	.01	.00	.00	.02	.008	4.2	2.9–6.5	.523	.258–.711
T.T.	.95	.01	.01	.02	.00	.00					
conn .0200, *dec* .56	.93	.04	.01	.00	.00	.02	.017	21.3	16.4–28.9	**.001**	.000–.006
J.Fr.	.93	.01	.01	.02	.00	.02					
conn .0200, *dec* .56	.93	.04	.01	.00	.00	.02	.013	18.4	13.4–27.2	**.002**	.000–.020
V.C.	.92	.02	.01	.01	.00	.03					
conn .0200, *dec* .57	.88	.05	.02	.01	.01	.03	.020	5.7	4.8–6.9	.334	.229–.438
L.B.	.82	.04	.02	.01	.01	.09					
conn .0070, *dec* .50	.82	.05	.03	.01	.02	.07	.011	3.3	2.7–4.1	.653	.531–.748
J.B.	.83	.06	.01	.03	.01	.06					
conn .0065, *dec* .50	.77	.06	.04	.01	.02	.10	.033	14.8	12.9–17.0	**.011**	.005–.024
J.L.	.85	.03	.01	.03	.01	.06					
conn .0250, *dec* .60	.82	.06	.03	.01	.01	.06	.021	15.5	13.4–18.6	**.008**	.002–.020
G.S.	.73	.02	.06	.01	.02	.15					
conn .0057, *dec* .50	.70	.07	.06	.01	.03	.13	.024	6.5	6.0–7.1	.263	.216–.305
L.H.	.71	.03	.07	.01	.02	.15					
conn .0057, *dec* .50	.70	.07	.06	.01	.03	.13	.020	5.5	5.0–6.1	.358	.292–.417
J.G.	.59	.06	.09	.04	.03	.20					
conn .0450, *dec* .70	.56	.10	.11	.02	.05	.16	.029	10.9	10.0–12.1	.053	.034–.074
E.G.	.94	.03	.00	.02	.00	.01					
conn .1000, *dec* .60	.95	.03	.00	.01	.00	.00	.008	11.1	6.8–21.3	**.049**	.001–.233
B.Me.	.89	.03	.01	.05	.01	.00					
conn .1000, *dec* .82	.85	.09	.01	.03	.00	.02	.034	50.9	37.0–91.1	**.000**	.000–.000
B.Mi	.88	.05	.01	.02	.01	.01					
conn .0550, *dec* .70	.84	.08	.02	.02	.00	.03	.023	7.1	6.2–8.5	.211	.131–.284
J.A.	.88	.05	.00	.03	.01	.03					
conn .0580, *dec* .70	.88	.07	.01	.02	.00	.02	.016	19.1	15.0–26.9	**.002**	.000–.010
A.F.	.78	.02	.03	.06	.04	.07					
conn .1000, *dec* .85	.78	.12	.02	.04	.00	.05	.044	88.7	73.6–106.4	**.000**	.000–.000
N.C.	.80	.03	.07	.01	.00	.09					
conn .1000, *dec* .85	.78	.12	.02	.04	.00	.05	.047	42.3	39.5–46.1	**.000**	.000–.000
I.G.	.77	.10	.06	.03	.01	.03					
conn .1000, *dec* .86	.73	.13	.03	.04	.01	.06	.026	9.2	8.3–10.3	.101	.067–.142
H.B.	.61	.06	.13	.02	.01	.18					
conn .0500, *dec* .71	.60	.11	.10	.02	.04	.13	.035	13.8	12.9–14.7	**.017**	.012–.024
J.F.	.66	.16	.01	.13	.01	.03					
conn .1000, *dec* .86	.73	.13	.03	.04	.01	.06	.051	38.1	35.1–42.1	**.000**	.000–.000
G.L.	.29	.04	.22	.03	.10	.32					
conn .0790, *dec* .85	.28	.11	.18	.03	.10	.30	.036	9.7	9.1–10.4	.085	.065–.106
W.R.	.08	.06	.16	.05	.35	.30					
conn .1000, *dec* .94	.18	.10	.18	.03	.12	.39	.113	98.1	93.6–102.5	**.000**	.000–.000

Note: RMSD stands for root mean squared deviation. Boldface indicates significant mismatches.

the standard X^2 measure, however, reveals that Dell et al.'s reported results provide a very poor match to their patient data. Almost half of the naming patterns were not replicated.

Improving the Model's Fit

One possible explanation for the core model's poor fit could be the manual regression procedure used by Dell et al. They report that "the fitting process was informal" (p. 818) and "if a patient could be reasonably well fit with a [parameter] combination that we had already used, we just used that combination rather than try to fine tune the fit" (p. 818–19). So it seems likely that the model's fit to the patient naming data could be improved by the use of a more formal regression process.

To this end, we developed an automated fitting procedure based on numerical optimization. An overview of its operation can be found in Appendix A. We used the procedure to refit the core model to Dell et al.'s patients, deriving new *connection* and *decay* values for each patient response distribution. To allow finer adjustments at low values (and following the representation of the parameter space in Dell et al.'s Figure 5), the algorithm manipulated *connection* on a logarithmic scale. The algorithm stopped fitting each patient when it was confident that any better solution had a *decay* value within 0.01 of the current solution, and a *connection* value whose logarithm (base 10) was within 0.025.

The fits found by the algorithm are shown in Table 3, with 95% confidence intervals on the X^2 values and their associated significance levels. A comparison of the X^2 confidence intervals with those resulting from the parameter settings reported by Dell et al. (our Table 2) shows that, for thirteen of the twenty-one patients (62%), our algorithm found fits with significantly lower X^2 values. For the remaining eight cases, the fits reported by our algorithm were not significantly better or worse, given about 10,000 samples from each distribution. Overall, the algorithm's fits had a mean X^2 of 13.2 (median 8.0) compared with a mean X^2 of 23.5 (median 13.8) for the fits reported by Dell et al. Measured by RMSD, the algorithm's fits had a mean RMSD of 0.027 (median 0.024) compared with a mean RMSD of 0.030 (median 0.024) for Dell et al.'s reported fits. These results give us confidence that our fitting algorithm is, in general, at least as good as the manual fitting process followed by Dell et al., and confirms our previous suspicion that a more formal fitting process would improve the model's fit.

Even with this improved fit, however, five of the twenty-one patients (24%) cannot be fit by the model (patients J.L., B.Me., A.F., J.F., and W.R.). Four of the fits are very poor, with even the lower bound of a 95% confidence interval on X^2 indicating a failure to match with 99.8% significance. The rigorous nature of the fitting algorithm suggests that these failures are due to intrinsic inadequacies in the model, rather than experimental happenstance. Our results lead us to conclude that, contrary to the claims of Dell et al., the fit of their model to patient naming data is mediocre and that, in general, the model cannot account for many patterns of aphasic naming.

Although aphasic patients are the main focus of our discussion, we also attempted to fit the model to the error pattern of normal participants (as reported by Dell et al., p. 810). The results are shown in Table 4. The closest fit found by our optimization algorithm results in a distribution that is significantly different ($p < 0.0005$) from normal performance, even though it is closer than the distribution achieved by the parameters suggested by Dell et al.. When adjusted for a high probability of a correct response, the model cannot make

Table 3: Fits of Dell et al.'s model to their patients, using the parameters found by our optimization algorithm.

Patient and	Naming response						Fit				
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	X^2 conf.	p	p conf.
W.B.	.94	.02	.01	.01	.00	.01					
conn .0375, *dec* .62	.92	.04	.01	.01	.00	.01	.012	2.5	1.9–3.2	.778	.665–.861
T.T.	.95	.01	.01	.02	.00	.00					
conn .0422, *dec* .63	.94	.04	.01	.01	.00	.01	.014	8.0	6.6–10.1	.156	.072–.251
J.Fr.	.93	.01	.01	.02	.00	.02					
conn .0563, *dec* .69	.90	.06	.01	.02	.00	.02	.024	7.9	6.9–9.1	.160	.107–.227
V.C.	.92	.02	.01	.01	.00	.03					
conn .0313, *dec* .61	.91	.05	.01	.01	.00	.02	.011	3.1	2.5–4.1	.684	.542–.779
L.B.	.82	.04	.02	.01	.01	.09					
conn .0104, *dec* .53	.81	.06	.03	.01	.02	.08	.013	2.8	2.1–3.9	.734	.560–.835
J.B.	.83	.06	.01	.03	.01	.06					
conn .0453, *dec* .67	.82	.08	.03	.02	.01	.04	.014	4.5	3.6–5.8	.481	.330–.602
J.L.	.85	.03	.01	.03	.01	.06					
conn .0453, *dec* .67	.82	.08	.03	.02	.01	.04	.026	11.2	9.6–13.3	**.048**	.021–.087
G.S.	.73	.02	.06	.01	.02	.15					
conn .0057, *dec* .50	.69	.07	.06	.01	.03	.14	.025	6.4	5.5–7.4	.274	.193–.362
L.H.	.71	.03	.07	.01	.02	.15					
conn .0055, *dec* .50	.67	.07	.07	.01	.03	.15	.023	5.2	4.4–6.2	.394	.285–.487
J.G.	.59	.06	.09	.04	.03	.20					
conn .0470, *dec* .70	.59	.11	.10	.02	.04	.14	.031	10.4	8.8–12.4	.065	.029–.117
E.G.	.94	.03	.00	.02	.00	.01					
conn .0703, *dec* .71	.93	.05	.00	.02	.00	.01	.009	2.9	2.1–4.1	.715	.534–.836
B.Me.	.89	.03	.01	.05	.01	.00					
conn .0672, *dec* .73	.85	.08	.01	.02	.00	.02	.031	22.2	18.9–27.3	**.000**	.000–.002
B.Mi	.88	.05	.01	.02	.01	.01					
conn .0484, *dec* .67	.87	.07	.02	.01	.00	.02	.010	5.0	3.5–7.2	.420	.205–.616
J.A.	.88	.05	.00	.03	.01	.03					
conn .0550, *dec* .70	.84	.08	.02	.02	.01	.03	.025	9.0	7.8–10.8	.109	.055–.166
A.F.	.78	.02	.03	.06	.04	.07					
conn .0531, *dec* .71	.72	.10	.06	.02	.02	.07	.044	26.5	23.7–29.7	**.000**	.000–.000
N.C.	.80	.03	.07	.01	.00	.09					
conn .0065, *dec* .50	.77	.06	.05	.01	.02	.09	.021	8.8	7.6–10.3	.117	.067–.181
I.G.	.77	.10	.06	.03	.01	.03					
conn .0641, *dec* .74	.77	.11	.04	.02	.01	.05	.011	2.6	2.0–3.7	.754	.601–.852
H.B.	.61	.06	.13	.02	.01	.18					
conn .0375, *dec* .67	.57	.10	.11	.02	.05	.16	.030	9.7	8.5–11.0	.085	.051–.130
J.F.	.66	.16	.01	.13	.01	.03					
conn .0984, *dec* .86	.69	.14	.05	.04	.01	.07	.046	36.5	31.9–42.4	**.000**	.000–.000
G.L.	.29	.04	.22	.03	.10	.32					
conn .0734, *dec* .83	.28	.11	.19	.03	.10	.30	.032	8.3	7.4–9.6	.138	.088–.193
W.R.	.08	.06	.16	.05	.35	.30					
conn .0969, *dec* .94	.15	.08	.17	.02	.13	.45	.114	82.8	76.0–90.8	**.000**	.000–.000

Note: RMSD stands for root mean squared deviation. Boldface indicates significant mismatches.

Table 4: Fits of Dell et al.'s model to control participants, first using the parameters found by our optimization algorithm, and also using the parameters they suggest.

	Naming response						Fit		
Data source	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	p
Controls	10,094	120	6	90	28	5			
conn .0806, *dec* .55	9,748	164	0	85	0	3	.002	41.7	.000
vs. .1 and .5	9,668	219	0	104	0	9	.006	68.5	.000

Note: RMSD stands for root mean squared deviation.

any phonological or unrelated errors. It appears that, in addition to its failures on aphasic patients, the model is unable to match the performance of normal participants.

Evaluating in Relative Terms

We have seen that Dell et al.'s model cannot replicate the behavior of a quarter of their patients (even when ignoring descriptions and non-responses), and therefore that its performance cannot be used as evidence in support of theoretical claims such as interactivity, globality, or continuity. But perhaps patient naming patterns are inherently very difficult to model successfully. Coming as close as Dell et al.'s model has might then be an impressive achievement. We could then interpret the model's promising performance as strongly suggestive of directions for further research. Although we still could not use it as a basis for theoretical inferences, the model would have important heuristic value.

Evaluating the model in relative terms will give us a quantitative basis for such judgments. The model's performance would be impressive only if it were difficult to match sixteen of twenty-one patients (or score a mean RMSD of 0.028) using a model that is as simple as Dell et al.'s. (Ockham's razor tells us that we are only interested in theories that are as simple or simpler.) We can determine if this is the case by comparing the model's performance to that of some simple alternatives.

Perhaps the simplest theory of aphasic naming is that all patients exhibit the same performance. If we choose that constant pattern as the mean of the patient data, then comparing against this simple theory is known as computing the variance accounted for (VAF). VAF, also called the proportion of variance explained, is widely used to evaluate models produced by techniques such as linear regression (Mosteller, Fienberg, & Rourke, 1983), classical multidimensional scaling (Cox & Cox, 1994), or clustering (Shepard & Arabie, 1979). To calculate VAF, one compares the error between the model and the data to the error between the mean of the data and the data. If the model is explaining at least a part of the measured phenomenon, it should perform better than a constant model that always guesses the mean. Although the constant model is clearly incorrect, its performance provides a quantitative way to assess how far Dell et al.'s model has come toward explaining the data, even if we know that it can't provide a match.

We performed a VAF analysis of Dell et al.'s model, using the fits found by our regression algorithm. We will express the VAF of the model and the patient data as the percentage of the variance in the data that is accounted for by the model, computing the

mean separately for each response category and summing over all categories:

$$\mathrm{VAF}(M,P) = 1 - \frac{\sum(data - model)^2}{\sum(data - mean_{cat})^2}$$

If the model were perfect (and tolerant of sampling error), it would account for 100% of the residual error left from guessing the mean. And, if it were worse than guessing the mean, the model's VAF would be negative. The model accounted for 87% of the variance, which is much better than guessing the mean. However, when we calculate VAF by summing over categories, we are implicitly weighting response categories with higher variance (such as correct and nonword) more than those with lower variance (such as phonological or unrelated). This is inappropriate here, since patient response rates in low-variance categories such as phonological or mixed errors are just as important for evaluating theories of lexical access as response rates in the high-variance categories. (In fact, we will see later than phonological and mixed errors play a large role in the predictions that Dell et al. derive from the model.) Calculating the VAF separately in each category, we find that the model accounted for the following percentages of the variance in each category:

correct	98%
semantic	-40%
phonological	90%
mixed	11%
unrelated	55%
nonword	83%.

The mean of these category VAFs is 49%. This means that, on average, the model performed better than guessing the mean in each category, reducing the error by half. It performed best in the correct category, which is not surprising, since we optimized the model's fit using the X^2 statistic, which is sensitive to the number of samples, and the correct category is the one in which most patients made most of their responses. It is surprising, however, that the model fares so poorly in the semantic category. It performs worse here than guessing the mean of the data.

Of course, a constant model is an overly simplistic theory, even as a baseline. It is obvious that different aphasic patients exhibit different patterns of naming errors. It would be more interesting to compare against a baseline model that had adjustable parameters to separately fit the particular aphasia of each patient. Like Dell et al.'s model, such a model would be able to represent a range of different patterns, depending on the adjustments to the parameters. The simplest such model would be a linear theory of naming. In other words, the baseline hypothesis would be that there is a single spectrum of naming performance, with each patient falling somewhere exactly between two extremes. This means that we are assuming that, if we plotted all the patients' error mixes as points in six-dimensional space, they would all happen to line up along a single straight line shooting through the space in some orientation. Since the theory restricts us to a line, we could then represent each error pattern by a single number representing its position along that line. Comparing against such a simple linear model would still be misleading, however, since it has only one adjustable parameter we can tune to fit each patient, whereas Dell et al.'s model has two. We can extend the linear model with a second parameter, expanding its range of representable

patterns from a line to a two-dimensional plane in the space of patterns. As we explain in Appendix B, such a model embodies the theoretical claim that patient performance is composed only of two components that interact with patient response probabilities simply by summing. The strengths of the two components then locate each patient on that plane. Like the constant model, this two-parameter linear model does not identify specific mental representations and the theory it embodies is clearly false. It is useful only as a benchmark two-parameter theory against which we can compare Dell et al.'s model. Its simplicity implies that it could well be missing important features of aphasic naming behavior that are explicitly articulated in Dell et al.'s model of aphasic lexical access. (Indeed, we will see in the next section that Dell et al.'s model is more complicated in the specific sense that it allows the patients to lie on a complex curved two-dimensional surface in the response space.) If those additional assumptions are useful, we would expect Dell et al.'s model to outperform the linear one.

An evaluation of the linear model is complicated somewhat by the fact that its fixed parameters must be estimated from the patient data (just as we computed the mean for the VAF measure). This must be done carefully, so as not to allow the model to implicitly memorize the data it will be tested on. Details of our testing methodology and experiments are given in Appendix B—we will only repeat the main results here. The simple two-parameter linear model succeeds in matching ($p > 0.05$) seventeen of the twenty-one patients (81%), with a mean RMSD of 0.022 (median 0.015), a weighted VAF of 89%, and an average category VAF of 65%. The VAF in the individual categories is

correct	99%
semantic	28%
phonological	80%
mixed	57%
unrelated	33%
nonword	92%.

These figures indicate that the variance unaccounted for by Dell et al.'s model in the semantic and mixed categories is not entirely due to noise in the data. While the linear model does not match significantly more patients than Dell et al.'s model (which matched sixteen patients with a mean RMSD of 0.027, a weighted VAF of 87%, and an average category VAF of 49%), it does indicate that Dell et al.'s model is not doing a particularly good job of capturing the patient data. The data seem to be equally well matched with a simple linear model representing a superficial theory that is clearly false. The mediocre performance of Dell et al.'s model therefore remains unimpressive, even when evaluated in relative terms. Although we have not yet considered the accuracy of the supplementary predictions that Dell et al. derive from their model, both formal tests of its ability to replicate patient patterns and relative comparisons against simple models indicate that its fit to picture naming data is poor.

Possible Error Patterns

To fully understand Dell et al.'s model of aphasic naming, we need to know more than how well it can fit particular patients—we need some general intuition about what kinds of error patterns it can represent and what kinds of patterns it can't. Describing the

model's representational range would provide a first step toward understanding why it fits some patients but not others. Just as we saw how a linear model of naming assumed that all patients lay along a line or plane in six-dimensional space, we would like to have a similar graphical understanding of Dell et al.'s model. This would allow us to visually determine which basic patterns in the data the model is able to capture. Since it can be difficult to analyze such computational models abstractly, we will adopt the brute-force approach of empirical testing.

We sampled the output distributions of Dell et al.'s model using a wide range of possible settings of the two parameters, *connection* and *decay*. We varied the logarithm (base 10) of *connection* from -4 through -1 in increments of $\frac{1}{32}$ (resulting in values from 0.0001 through 0.1) and we varied *decay* from 0.5 through 1 in increments of $\frac{1}{128}$. In addition, we also varied *connection* non-logarithmically from 0 through 0.1 in increments of $\frac{1}{64}$, with *decay* varying the same way. This resulted in over 8,000 distinct parameter settings. At each of these parameter settings, we ran the core model simulation until the width of the widest 95% confidence interval around any of the six response probabilities was 0.02 (this took anywhere from 200 to 9,600 trials, depending on the results at that setting). This gave us, for each setting of the parameters, an estimate of the error mix produced by the simulation at that setting. By examining the range of mixes across all of these parameter settings, we can see the structure of those patterns that are describable by Dell et al.'s representation.

Although an error mix is a six-dimensional object, we can visualize the error mixes attainable in the representation by considering only two or three response categories at a time. In Figure 3, we consider only the proportions of semantic and phonological errors. Each small dot in the figure corresponds to a particular setting of *connection* and *decay*, and all settings we sampled are plotted in the figure. Circles represent patients reported by Dell et al. The performance of normal speakers is marked with a large X (partially obscured by points from the model), as are the random error opportunities in English. Dell et al. estimated the random point using the method of Dell and Reich (1981), a combination of random initial phoneme substitutions and target/error re-pairings. Clusters of dots in the figure imply that many different parameter settings for the simulation give rise to error mixes that have similar percentages of semantic and phonological errors, although these mixes may vary widely in the other response types (such as nonword errors) which are not shown. Looking at the boundary of the dotted area shows us the limits of the representation. Note that the clusters and patterns of variation in the density of dots in the figure arise from the systematic way in which we varied the input parameters, and do not represent important features of the representation—only the boundary of the area containing dots is meaningful for our purposes. We assume that, although we didn't happen upon it in our sampling, there exists a parameter setting that can give rise to a distribution that would fall between any two adjacent dots.

From the figure, we see that several patients lie outside the region that can be captured by the model. Note that the distance between the model and each patient is minimized in this two-dimensional projection, since a patient who appears to lie within the model's range may differ along dimensions not represented in the figure. The figure serves as a more intuitive, but less accurate, version of the χ^2 tests we discussed earlier.[3]

[3]Although our main evaluation criterion has been the ability of Dell et al.'s model to replicate patient

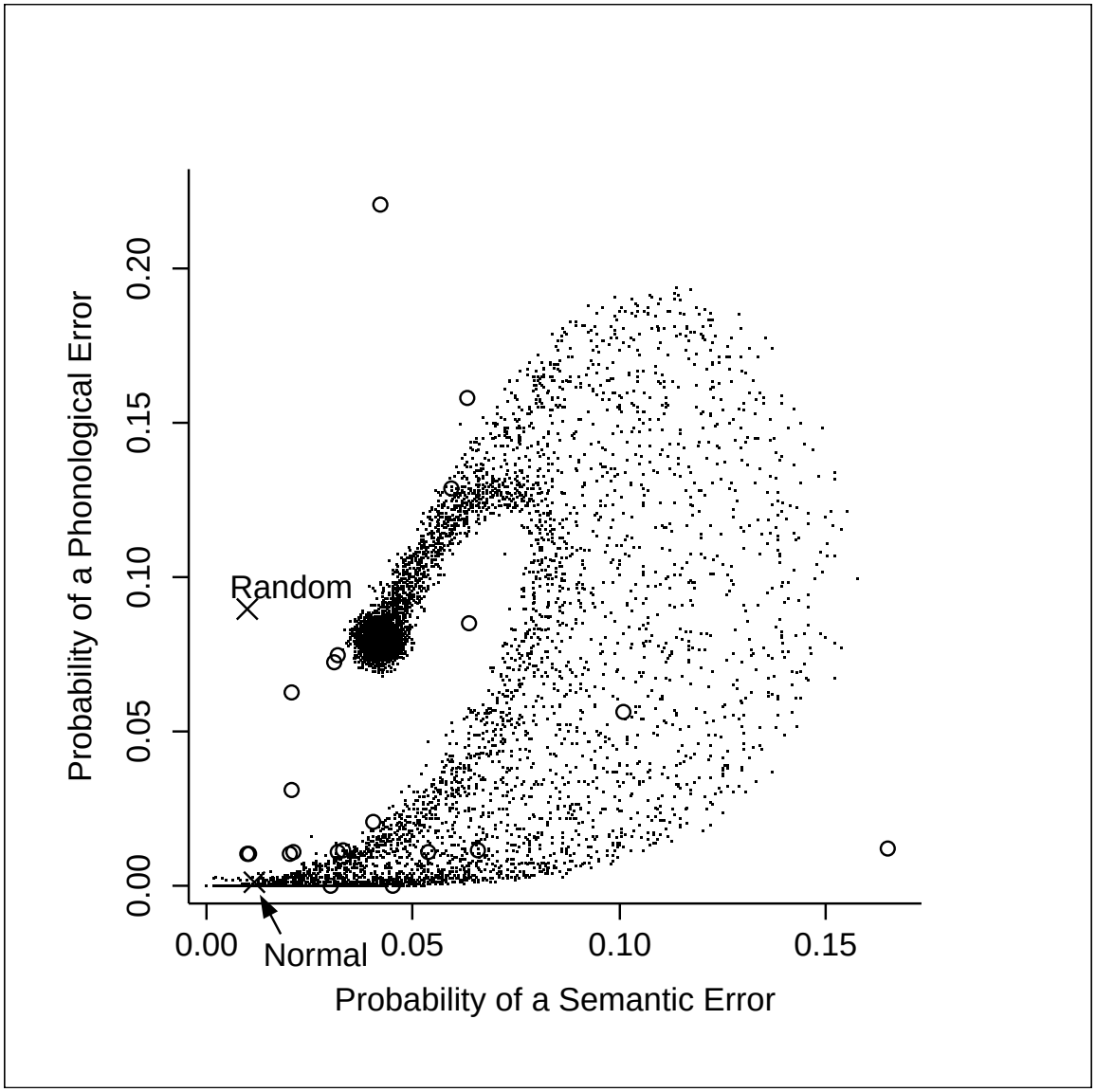

Figure 3. The relations between semantic and phonological errors that are attainable using Dell et al.'s two-parameter representation. Patients are plotted as circles.

A detailed examination of the figure shows that the model cannot represent any error mix with more than 17% semantic errors, and when representing any mix with more than 13% semantic errors, a significant number of phonological errors must also be included (note that mixed errors are not shown). Similarly, when representing any significant number of phonological errors, a large number of semantic errors must be present. These results weaken Dell et al.'s assertion (p. 802) that their representation accounts for a dissociation between predominately semantic or predominately phonological errors. It is patients exhibiting just such patterns from whom the model is especially distant.

Considering only semantic and nonword errors gives further insight. In Figure 4, we see that the representation also forces a large proportion of non-word errors whenever it represents a significant number of semantic errors. Any error pattern with more than 13% semantic errors must have at least 5% nonword responses. Conversely, the model also forces a significant number of semantic errors whenever it represents patients with a significant number of nonword errors. This forces a poor fit to several patients who have low rates of semantic errors.

The dimensions in which the limited range of the model is most obvious are those for mixed and unrelated errors. It is clear from Figure 5 that the range of variation of the model is inadequate to capture the variety among the patients. It can generate up to 13% unrelated errors while patients can make over 30%, and it can generate up to 5% mixed errors while patients can make more than twice that many.

The most intuitive constraint on the possible error mixes is the relationship between correct and gibberish responses. As shown in Figure 6, there seems to be a steady inverse relationship between the two. The model implies that any patient who gives a substantial percentage of incorrect answers will exhibit a substantial percentage of nonword errors. As the positions of two of the patients in the figure illustrates, this assumption is unwarranted. It is well-known that patients can exhibit a high rate of semantic or phonological errors without a high rate of neologisms (see, for example, Caramazza & Hillis, 1990).

The Continuity Thesis

These figures also allow us to informally assess the continuity thesis, in which Dell et al. claim that the error patterns of both the model and the patients fall along a continuum between normal performance and random error opportunities. Dell et al. define random error opportunities in a theory-neutral way as "...the distribution of error types that would occur if output is 'random' (i.e., if output is not affected at all by lexical retrieval)" (p. 808). However, they do not calculate the opportunities of English by scoring random phonologically legal sequences of phonemes. As we mentioned earlier, they instead use the estimation method of Dell and Reich (1981), which involves phoneme substitutions in the first position of target words. Presumably, this is motivated by a desire to estimate the distribution of error types produced by a process akin to the one used by patients. However, any artificial process of generating patient-like errors necessarily depends on an implicit theory of lexical

---

data, another desirable characteristic of any naming model is the inability to represent error patterns that do not arise in human performance. If enough patient data could be gathered to estimate the portion of error-pattern space occupied by possible human patterns, one could develop a quantitative measure of this inability, perhaps based on volume intersection.

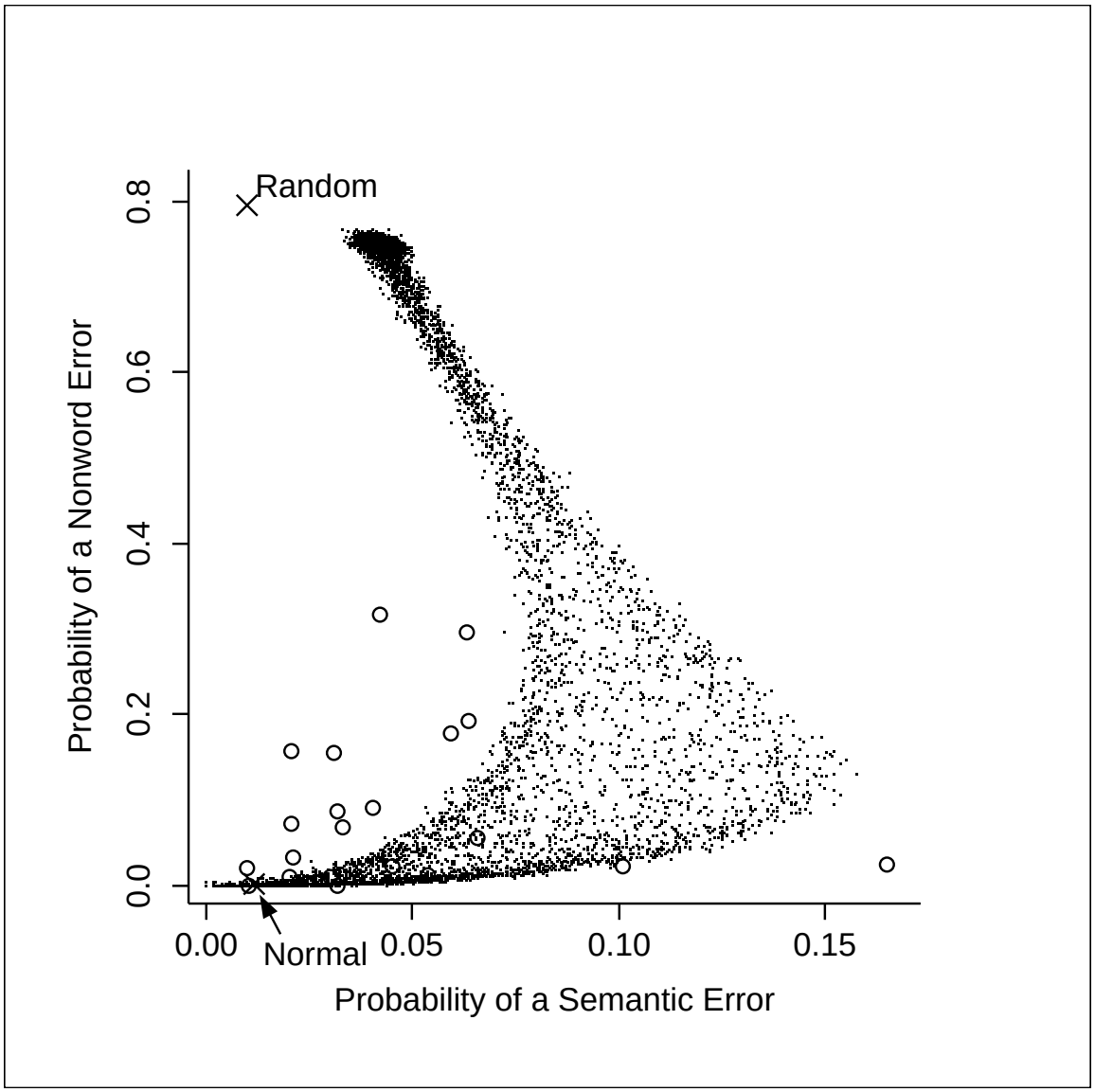

Figure 4. The relations between semantic and non-word errors that are attainable using Dell et al.'s two-parameter representation. Patients are plotted as circles. Note the difference in scale between the two axes.

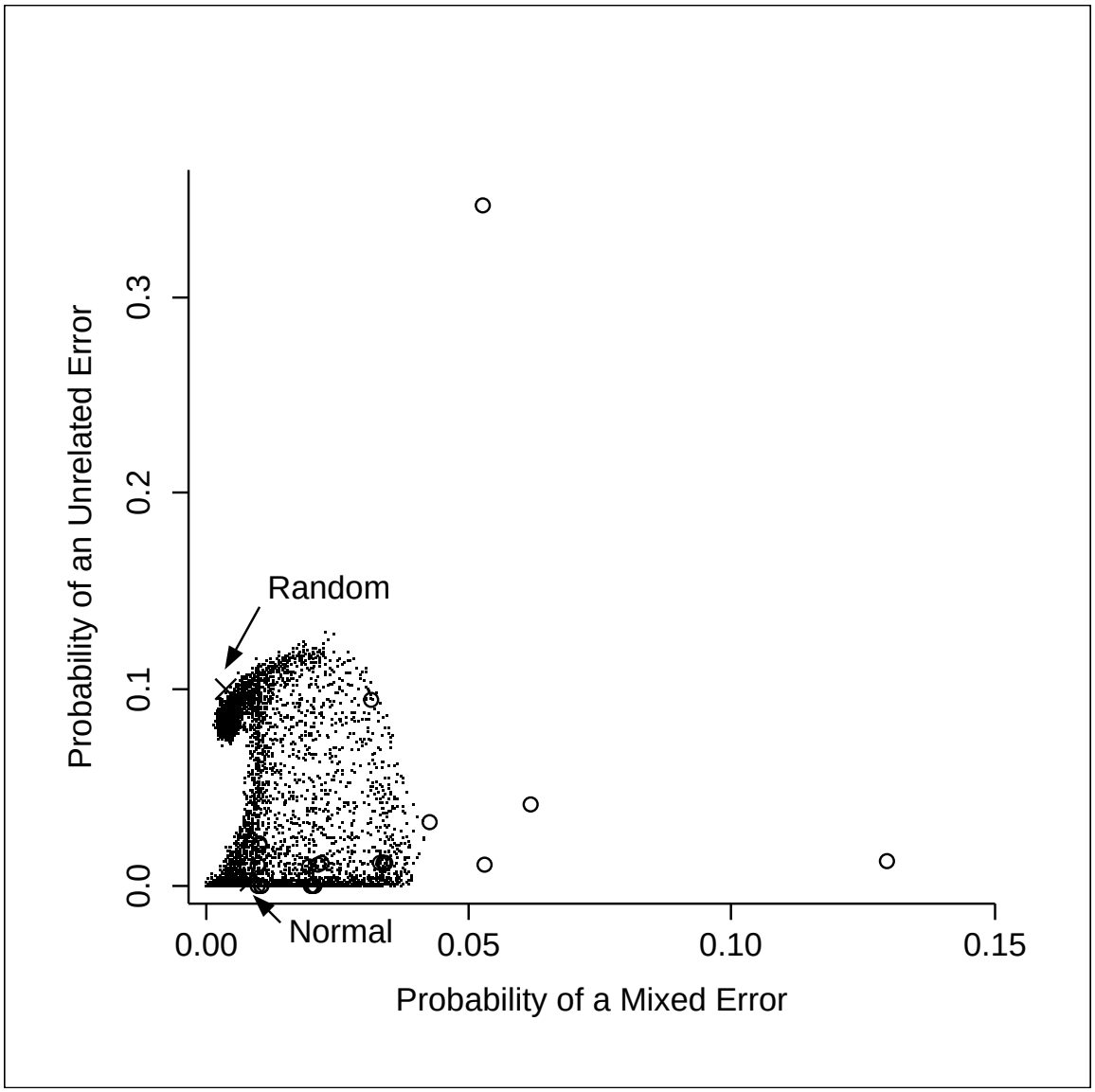

Figure 5. The relations between mixed and unrelated errors that are attainable using Dell et al.'s two-parameter representation. Patients are plotted as circles. Both the random and normal points are obscured by the model. Note the difference in scale between the two axes.

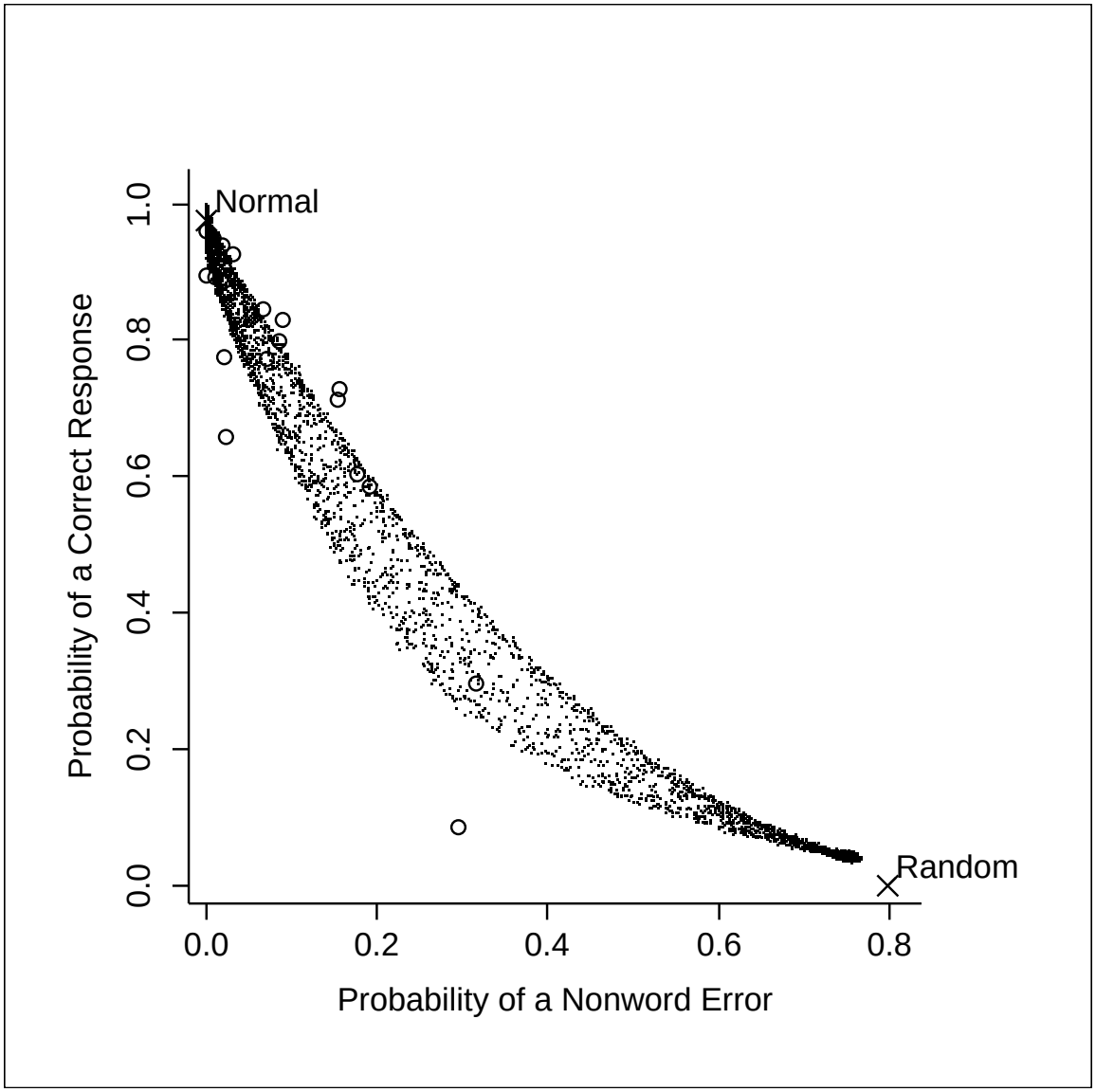

Figure 6. The relations between correct responses and non-word errors that are attainable using Dell et al.'s two-parameter representation. Patients are plotted as circles.

access. In this paper, we will merely assume that Dell et al.'s continuity thesis refers to the particular error probabilities they calculated.

Although a literal reading of the definition of the continuity thesis implies that error patterns must lie directly between the normal and random points, and therefore along a straight line, we will interpret it more loosely as claiming that error patterns should form a continuous curved surface which must start and end at the prescribed points, but which may deviate somewhat from a straight line. Since the continuity thesis only refers to the position of error patterns, it could be true or false of either the model or the patients independently, and we will consider each in turn.

Judging from Figures 3 through 6, the model does seem to admit a description in terms of a single curved surface (of intrinsic dimensionality two) flowing from the point of normal performance to the random distribution, rather than forming a scattered cloud in three or more dimensions. This is not particularly surprising, since we generated the points by varying only two parameters. Figures 3 and 6, especially, give the impression of a spinnaker filling out from the two points. While the model's possible patterns do not, strictly speaking, lie between the normal and random points (i.e., on a line or within a box), those points do seem to define the endpoints of a curved surface (somewhat like the peel of a single section of an orange, twisted) on which the model's patterns lie.

However, the wide dispersal of patients in most of the figures indicates that this is a poor characterization of aphasic naming. In Figures 3 and 5, patients lie far to both sides of the random point, and in Figure 4, patients far from the normal point lie in many directions. The position of W.R. at 8% correct and 30% nonword in Figure 6 serves as a reminder that low-correctness need not imply high nonword rates. To provide a clearer view, Figure 7 shows only the patients. Although they seem to follow a clear pattern, they do not seem to tend toward the random point as the percentage of either nonword or phonological errors increases. Figure 8 echoes Figure 5, although this time using semantic errors. We see that while patients vary widely in their frequency of semantic errors, they are not constrained to lie between the random and normal points in any obvious way. Of course, some patients may exhibit random naming behavior. But there seems to be such a variety of different patterns of breakdown that it would be misleading to speak of possible aphasic behavior as constrained by random error opportunities. The continuity thesis does not find support in Dell et al.'s data.

The Effects of the Parameters

Figures 3 through 6 gave us some intuitions about the response distributions that the model could generate, but ignored the relation between those error patterns and the parameter values used to generate them. Looking at how the parameters affect the model's output is another way of understanding its behavior. To see how the parameters affect the model's response distribution, we can measure the similarity of the model's distribution to a reference distribution as we try different values of the parameters. In Figure 9, we use patient G.S. as our reference distribution, and we measure similarity using RMSD. Lighter shading indicates a distribution more similar to that of G.S. While most parameter settings are unsuitable, there seems to be a diagonal band of settings that produce distributions similar to G.S.'s. In this region, one can use various settings of *decay*, as long as one modifies *connection* accordingly. This helps to explain why our fitting algorithm was able

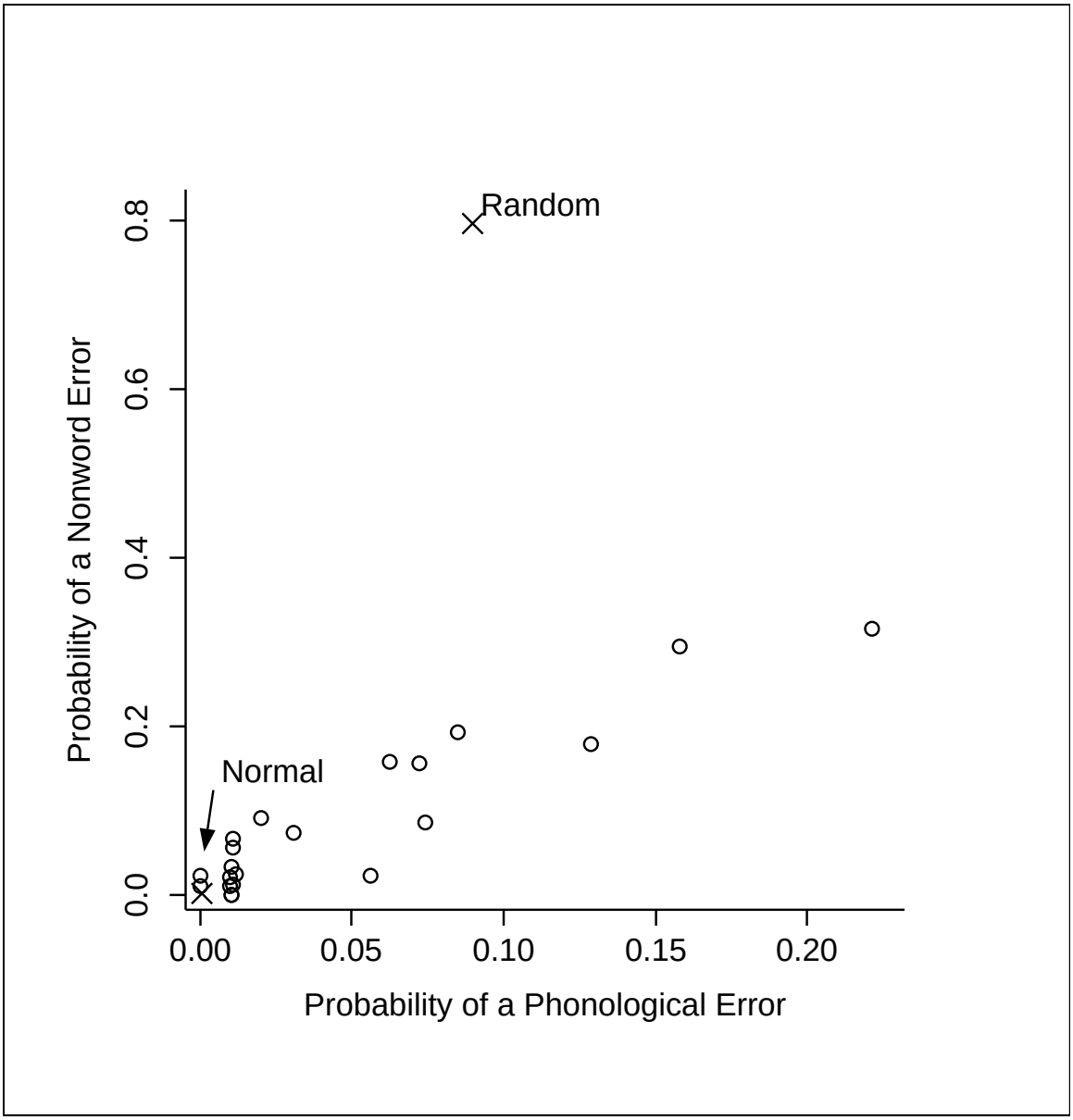

Figure 7. The percentages of nonword and phonological errors of Dell et al.'s patients.

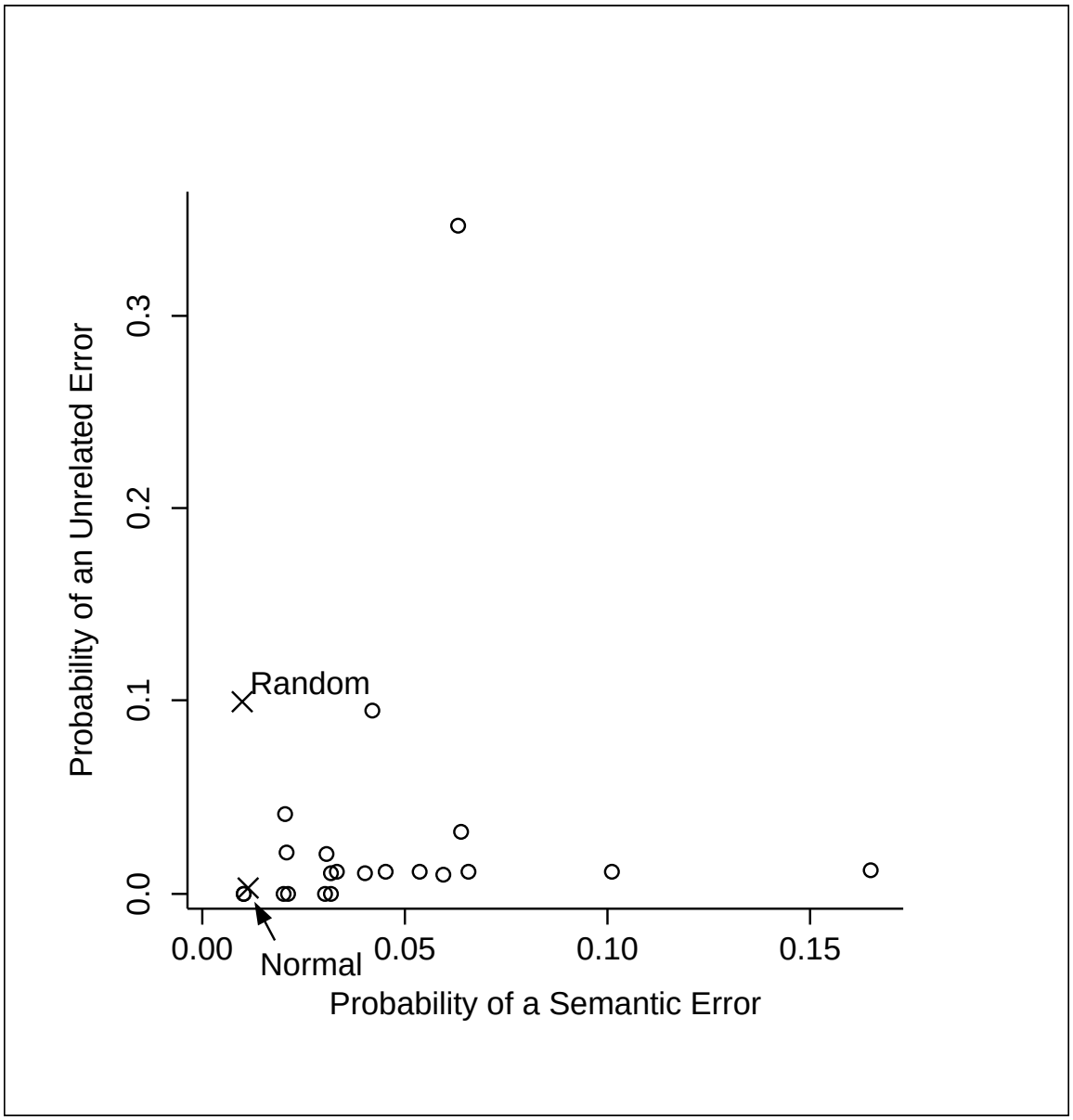

Figure 8. The percentages of unrelated and semantic errors of Dell et al.'s patients.

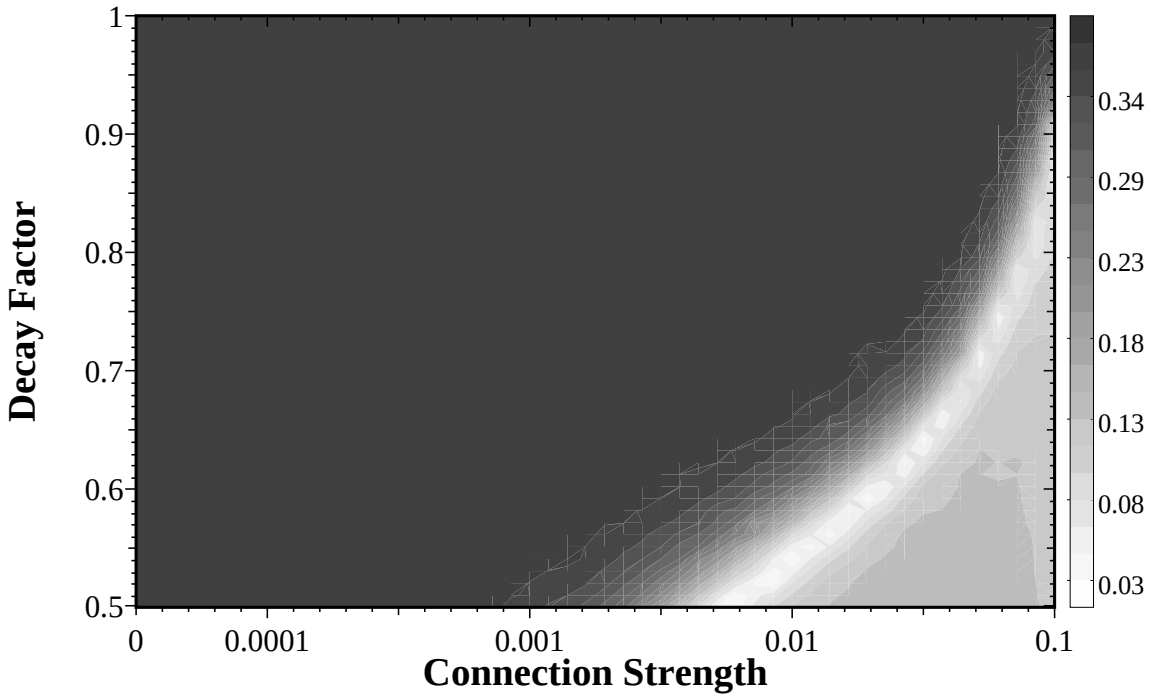

Figure 9. The RMSD of the model to patient G.S. at various parameter settings.

to improve on many of Dell et al.'s manual fits, since much effort is required to distinguish the true optimum from the many reasonably good values. The shading suggests that the model's fit to G.S. improves as *decay* is lowered, which agrees with our regression algorithm (which chose a *decay* of 0.5 and a *connection* of .0056 for G.S.).

One experiment this suggests is trying a *decay* value less than 0.5, to see if the fit could be further improved. Although Dell et al. do not consider them, there is no inherent computational reason why lower values of *decay* would be inappropriate. In Figure 10, we show the model's behavior relative to G.S. when we allow *decay* and *connection* to both vary from 0 through 1, rather than stop at 0.5 and 0.1, respectively. Figure 9 would occupy the upper left portion of this expanded view. We can see from the figure that not only does the model get even closer to G.S.'s distribution with very low values of *decay* and *connection*, but that there is a second band of similar distributions when *connection* is above 0.1. The values of the parameters that allow the model to come close to G.S. seem to form a wishbone-shaped region. This general shape occurred for all patients we examined, although the right-hand high *connection* region varied in quality, relative to the left-hand region. Figure 11 shows another example, using patient I.G.

These figures raise the possibility that many patients, such as G.S., could be fit better by parameters that lie outside of the parameter region we have been considering. To test this hypothesis, we refit all the patients, allowing the fitting algorithm to try parameter values from 0 through 1 for both *decay* and *connection*. For seven of the twenty-one patients (33%), the most likely fit of the model resulted from parameter values outside the usual

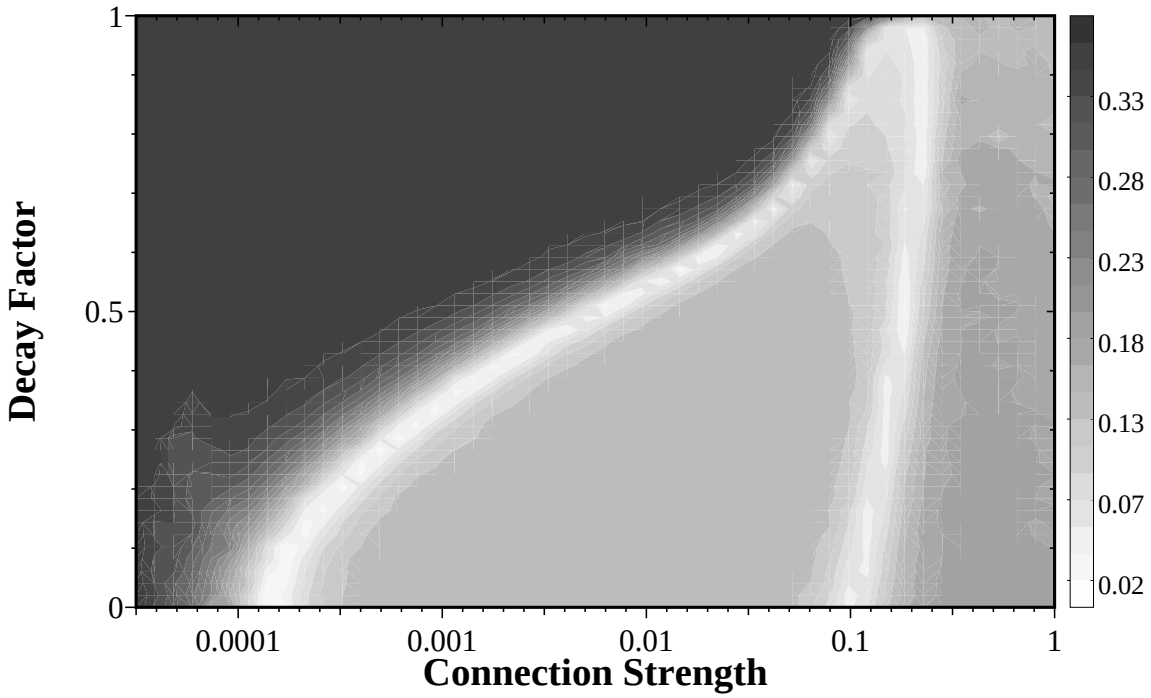

Figure 10. The RMSD of the model to patient G.S. at a wider variety of parameter settings.

region. The improved fits are shown in Table 5. This result is surprising, given Dell et al.'s intuitive interpretations of the model's parameters. Dell et al. consider the parameter called *connection* to represent ease of transmission of signals in the human brain, or the degree to which different levels of representation are communicating effectively. The parameter called *decay* is seen as reflecting the integrity and accuracy of each mental representation. If *decay* corresponds to representational integrity, then we would expect that a *decay* value less than 0.5 (Dell et al.'s setting for normals) would result in behavior even more accurate than that of normals. Similarly, if *connection* corresponds to ease of transmission, we would expect that brain damage would be better modeled by values lower than that used to model normals (0.1). (Dell et al. use these same intuitions to derive predictions regarding recovery that we will discuss in the next section.) The new fits imply that seven of the brain-damaged patients have mental representations that are either more accurate or communicating more effectively than those in normal participants. This contradiction suggests that we must use caution when assessing the extent to which elements of the model truly correspond to relevant concepts from neuropsychology.

By exhaustively sampling the possible parameter settings of Dell et al.'s core model, we have gained a more intuitive understanding of its capabilities and limitations. For example, any patient who exhibits significant phonological or semantic errors without making many nonword responses cannot be represented. While the model seems to reflect the continuity thesis, the patients do not. These analyses complement the results we obtained when fitting particular patients. Not only does the model have a poor fit to individual

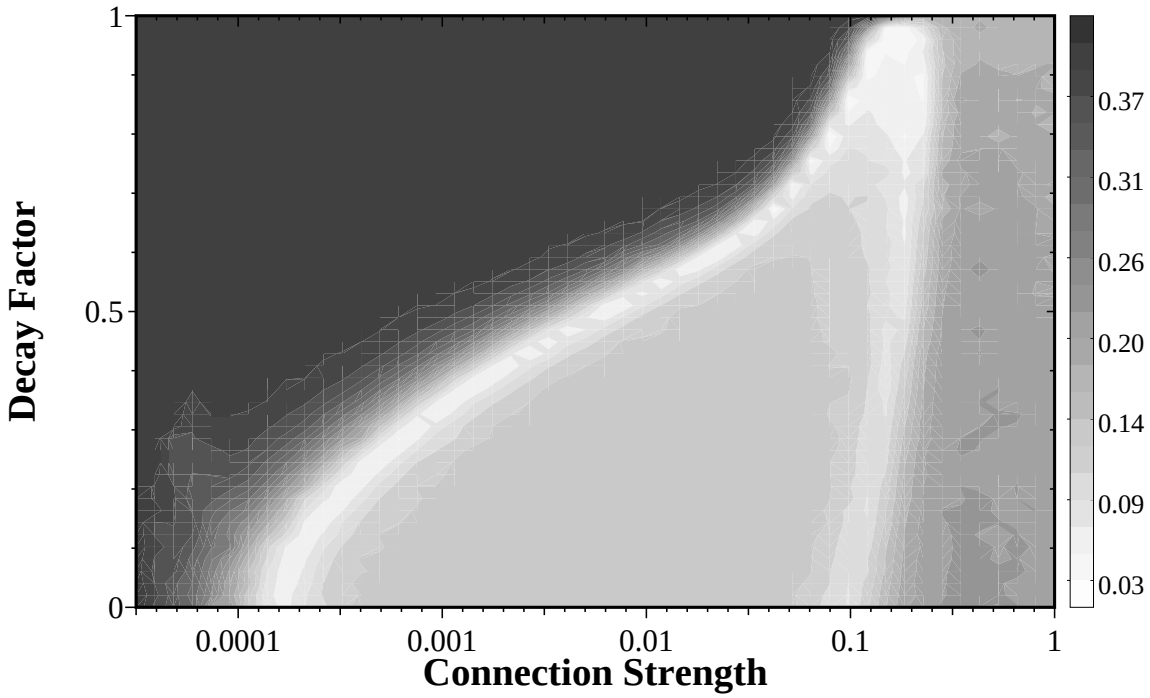

Figure 11. The RMSD of the model to patient I.G. at a wide variety of parameter settings.

patients, but it inherently excludes classes of error patterns that are known to be exhibited by patients.

The Model's Predictions

Dell et al. credit their model not only with the ability to account for the pattern of naming errors observed in aphasic patients, but also with the ability to make predictions about other aspects of patients' performance. Since the model makes detailed assumptions regarding the mechanisms involved in lexical access, those assumptions can be used to generate predictions regarding related tasks or phenomena that share those mechanisms. (It is this breadth that underlies the intuitive appeal of cognitive theories.) In Dell et al.'s theory, the predictions are obtained by first fitting the core model to each patient's naming performance, and then using the resulting values of *connection* and *decay*. Dell et al. derive four predictions:

Frequency of nominal phonological errors If one assumes that selection of lexical nodes in the human lexical system respects syntactic category, and therefore that only lexical nodes representing nouns are ever selected during picture naming, then those responses that are phonologically related to the target but are not nouns probably result from misselection at the phonological level.[4] With a normal *connection*

[4]This assumes that syntactic category is not specified by communication from the semantic representation to the lexical level, and is thus exempt from damage in the model.

Table 5: Patients that are better fit by parameter settings outside the area considered by Dell et al.

Patient and	Naming response						Fit				
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	X^2 conf.	p	p conf.
L.B.	.82	.04	.02	.01	.01	.09					
conn .0002, *dec* .07	.83	.04	.03	.01	.01	.08	.005	1.4	0.8–2.5	.922	.775–.975
conn .0104, *dec* .53	.81	.06	.03	.01	.02	.08	.013	2.8	2.1–3.9	.734	.560–.835
G.S.	.73	.02	.06	.01	.02	.15					
conn .0002, *dec* .04	.72	.05	.06	.01	.02	.15	.013	3.5	2.8–4.6	.621	.464–.725
conn .0057, *dec* .50	.69	.07	.06	.01	.03	.14	.025	6.4	5.5–7.4	.274	.193–.362
L.H.	.71	.03	.07	.01	.02	.15					
conn .0002, *dec* .04	.70	.05	.06	.01	.02	.16	.012	2.9	2.2–3.9	.709	.561–.817
conn .0055, *dec* .50	.67	.07	.07	.01	.03	.15	.023	5.2	4.4–6.2	.394	.285–.487
E.G.	.94	.03	.00	.02	.00	.01					
conn .1286, *dec* .51	.93	.03	.00	.02	.00	.01	.001	0.1	0.0–0.3	1.000	.998–1.000
conn .0703, *dec* .71	.93	.05	.00	.02	.00	.01	.009	2.9	2.1–4.1	.715	.534–.836
N.C.	.80	.03	.07	.01	.00	.09					
conn .0002, *dec* .03	.81	.04	.04	.01	.01	.10	.017	8.5	6.8–10.9	.129	.053–.233
conn .0065, *dec* .50	.77	.06	.05	.01	.02	.09	.021	8.8	7.6–10.3	.117	.067–.181
H.B.	.61	.06	.13	.02	.01	.18					
conn .0001, *dec* .06	.58	.06	.09	.01	.04	.23	.031	9.3	7.8–11.4	.096	.044–.167
conn .0375, *dec* .67	.57	.10	.11	.02	.05	.16	.030	9.7	8.5–11.0	.085	.051–.130
J.F.	.66	.16	.01	.13	.01	.03					
conn .1411, *dec* .97	.70	.15	.03	.04	.01	.07	.045	30.6	26.6–35.8	**.000**	.000–.000
conn .0984, *dec* .86	.69	.14	.05	.04	.01	.07	.046	36.5	31.9–42.4	**.000**	.000–.000

Note: The three response distributions in each block correspond to the patient, the new parameter settings, and the best parameters that are inside the usual region. RMSD stands for root mean squared deviation. Boldface indicates significant mismatches.

value, feedback from the phoneme layer should result in additional misselections at the lexical level, and hence a higher percentage of phonological errors would be nouns than one would predict by chance. If *connection* has a very low value (below 0.05), then there would be almost no feedback from the phoneme layer to the lexical layer, and we should predict noun responses to occur at chance rates.

Frequency of mixed errors Similarly, a lower *connection* value should also reduce the feedback that boosts the occurrence of mixed errors. So if *connection* is less than 0.05, mixed errors should occur no more frequently than one would predict by chance.

Naming during recovery By assuming that recovery from brain damage implies movement of *connection* and *decay* toward normal values, one derives a loose constraint on the possible naming error mixes observable during recovery: only those mixes obtainable using parameter settings between the currently fit values and the normal values should be observed.

Repetition If we assume that word repetition is carried out by the same mechanism as picture naming, but without the initial processing from semantic input, then we can

simulate repetition as the second half of naming. Input is represented by setting the activation level of the lexical node corresponding to the target word, and after allowing the usual number of time steps for activation to spread, phonemes are then selected as usual. By using the core model parameters derived from a patient's naming data, we can generate a predicted distribution of error types for the patient's word repetition performance.

Although Dell et al.'s model does a relatively poor job of matching patient naming data, it could be that its additional abilities as a predictive model are valuable. This would be the case if its predictions were particularly accurate or difficult. In this section, we will determine if this is, in fact, the case.

The Mixed Error Effect

The predictions of the frequency of nouns among phonological errors and the frequency of mixed errors are handled in the same way. We predict higher than chance levels of both phenomena if the value of *connection* that best fits the patient's naming data is greater than 0.05. We will consider the mixed error effect first. As Dell et al. point out, predicting the frequency of mixed errors given a patient's error pattern is less impressive that predicting the frequency of nouns among phonological errors, since the error pattern explicitly specifies the mixed error rate, while the noun rate is not included among the input to the model. However, recall that the model only accounts for 11% of the variance in the mixed category. This low reconstruction accuracy suggests that information regarding the mixed response category is, to a large extent, lost during the compression of the six error categories into the two values for *connection* and *decay*. Since this information is not necessarily available, it is not obvious that the model can predict the mixed error effect correctly based on the *connection* value alone.

Dell et al. divide their patients into two groups according to each patient's value of *connection*. To increase the amount of relevant data, they add two additional patients (G.B. and V.P.) to the high *connection* group who were not considered when evaluating the model's naming fits due to their frequent nonresponses and circumlocutions. Considering the semantic and mixed responses from each group, they calculate the frequency with which phonemes in the first, second, or third position of the response are shared with the target. Comparing these rates with estimates of chance, they find that the patients in the high *connection* group show a significant mixed error effect in all three positions ($p < 0.004$ for each position, by our calculation[5]), as predicted. However, the model's prediction that the effect will be absent in the low *connection* group is not borne out by the rates reported by Dell et al. The group's mean rate in second position of .122, estimated from 74 trials, is greater than the chance rate of .060 with $p = .014$. (Although Dell et al. discuss this problematic discrepancy, use of a poor statistical approximation when computing confidence intervals seems to have misled them into thinking it was statistically insignificant.) When we recalculated the rates using the patient groups implied by the *connection* values found by our optimization algorithm, the mixed effect in second position grew larger (.139 from

[5]Tests in this section were performed using data for individual patients that were generously supplied to us by Deborah Gagnon. Most of the data were published by Gagnon, Schwartz, Martin, Dell, and Saffran (1997).

101 trials, greater than .060 with $p < .0005$). And when we removed from the analysis the five patients that the model had failed to fit and the two that Dell et al. considered unsuitable for fitting (G.B. and V.P.), the mixed error effect remained in second position for the low-weight group, and was no longer detectable in third position for the high *connection* group. In each variation of the analysis, the model's predictions regarding the mixed error effect are not consistent with Dell et al.'s data.

The Noun Effect

The second prediction derived from the *connection* values concerns the frequency of nouns among the patients' phonological errors. As with mixed errors, Dell et al. perform a grouped analysis of the noun effect using two additional patients they initially considered unsuitable for fitting by the model. (If we exclude these unsuitable patients and those that the model failed to fit, the noun effect is no longer significant ($p > .068$).) The low number of responses from most of the patients creates uncertainty as to the correct classification of each individual. Only W.R.'s proportion of formal errors that are nouns is significantly different from chance ($p = .048$, double-sided test) and only I.G., W.R., G.B., and V.P.'s are significantly above chance (single-sided test). Twelve of the twenty-three patients had three or fewer formal errors. This means that we cannot directly assess the accuracy of Dell et al.'s classification (by estimating the number of patients misassigned to each category, for instance). We can, however, ask how impressive it is to find the model's predictions confirmed. If such a situation is unlikely, then the model's predictions should be considered evidence in its favor. Otherwise, we should be unimpressed by its performance, since it may reflect random chance.

To evaluate the significance of the noun results, we constructed 10,000 random partitions of Dell et al.'s patients into groups of size 13 and 10 (the sizes of the groups Dell et al. constructed), and tested how frequently it was the case that one group had formal errors that were nouns significantly more often than chance while the other group's rate was indistinguishable from chance.[6] This turned out to be true for 4,750 of the classifications (48%). If we were to guess at random which group would exhibit the effects, we would therefore succeed with a random partition 24% of the time. This suggests that such a criterion is not a useful indicator of empirical support for a model of naming, and that the prediction of the grammatical class effect cannot be taken as evidence in support of their model.

Recovery Predictions

Dell et al. also use their model to predict patient naming performance after a recovery period (which varied between 1.5 to 9 months for their patients). The prediction they derive from their model is that the parameter values used to fit the patient's improved naming will each lie somewhere between the value used to model the original naming performance and the value used to model normals (0.1 for *connection* and 0.5 for *decay*). Dell et al. reported that this was the case for seven of the ten patients from whom they gathered naming data after a recovery period. However, they regarded the model's predictions as clearly confirmed

[6]Dell et al. note that their particular partition creates groups that exhibit the noun effect at rates that are significantly different from each other, but since their model only predicts that a difference will exist and not that it will necessarily be large enough to be significant, we do not consider this criterion when calculating the percentage of random partitions consistent with the model's predictions.

because, for nine of the ten patients, the value of *connection* remained on the same side of 0.05 in the fit to the post-recovery data as it was in the fit to the original data. But this represents a much weaker claim than the model's prediction, and results consistent with this weaker claim are also consistent with situations in which the parameter values of patients do not always move toward normal values during recovery. Thus, the weaker claim is useless as a test of the model's prediction. Dell et al. were presumably motivated by concerns regarding the sampling error associated with having only 175 responses from each patient and the difficulty of finding optimal parameter settings. These sources of noise might conceivably cause fits to some patients to fall outside the predicted range even if the model's predictions regarding recovery were accurate. However, if sampling and regression errors cause significant inaccuracies in determining the relationship between the original and the later parameters, then the recovery prediction is inherently untestable and could not be used to support the model no matter what results were obtained. We will assume that these errors are not significant, and that the model's recovery predictions should be evaluated.

Before we try to assess the significance of 70% accuracy, we should see if our improved fitting procedure can be used to increase it. New fits of Dell et al.'s model to the post-recovery patient naming data are presented in Table 6, along with the fits generated from the parameters suggested by Dell et al. As we saw with the original naming data, the optimization algorithm is able to find improved fits of the model, although four of the ten patients still cannot be fit by the model. Now we can use these parameter values to test the recovery predictions.

To test the model's predictions, we need to compare the fitted parameter values (from Table 6) with those found for the original data (from Table 3). This comparison is distilled in Table 7. The new parameters lie in the rectangle between the original fit and the normal values of 0.1 for *connection* and 0.5 for *decay* in only four of the ten cases (J.G., H.B., J.F., and W.R.). For two of those four, the patient's naming error pattern changed so little after the recovery period that parameters almost identical to the original ones were returned by the fitting algorithm. These patients clearly don't provide a good test of the model's predictions. But even if we include them, the model's accuracy has dropped to 40%, rather than improving. This is not statistically distinguishable from the 25% accuracy one would expect if the recovered parameters lay in a random direction from the original ones ($p > 0.15$). Although we have data from only ten patients, it seems that the recovery predictions cannot be taken as evidence in support of the model.

Let us now turn to the most detailed predictions that Dell et al. derive from the model, those for repetition performance.

Repetition Predictions

Dell et al. predict a patient's distribution of repetition errors by running a simulation of repetition using the parameters derived from the patient's naming data. As with naming data, Dell et al. quantify the fit of the predicted distribution to the patient's actual performance using RMSD. The average RMSD for the predictions reported by Dell et al. is 0.048 (median 0.024). To evaluate the accuracy of these predictions, we can forgo the χ^2 tests we used to evaluate the model's naming performance and use a simple comparison. In a manner similar to the VAF analysis we performed earlier, we will choose a constant model

Table 6: Fits of the core model to the performance of Dell et al.'s patients who were tested after a recovery period.

Patient and	Naming response							Fit			
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	X^2 conf.	p	p conf.
J.B.	.92	.01	.01	.03	.00	.03					
conn .0563, *dec* .69	.90	.06	.01	.02	.00	.02	.022	10.2	8.9–12.1	.069	.034–.112
vs. .0085 and .50	.89	.04	.02	.00	.01	.04	.019	31.1	23.9–44.3	**.000**	.000–.000
J.L.	.95	.02	.01	.01	.00	.01					
conn .0129, *dec* .52	.93	.03	.01	.00	.00	.02	.011	5.2	3.8–7.7	.395	.171–.575
vs. .1000 and .50	.97	.02	.00	.01	.00	.00	.010	247.0	240-262	**.000**	.000–.000
G.S.	.91	.00	.02	.01	.00	.05					
conn .0105, *dec* .52	.89	.04	.02	.01	.01	.04	.020	10.5	9.4–11.9	.063	.036–.094
vs. .0090 and .50	.91	.04	.02	.00	.00	.03	.020	14.5	12.7–17.9	**.013**	.003–.027
L.H.	.76	.01	.09	.02	.01	.10					
conn .0344, *dec* .65	.71	.09	.07	.01	.02	.10	.040	15.7	14.3–17.4	**.008**	.004–.014
vs. .0065 and .50	.77	.06	.05	.01	.02	.09	.029	21.8	18.6–26.0	**.001**	.000–.002
J.G.	.91	.02	.01	.03	.00	.03					
conn .0563, *dec* .69	.90	.06	.01	.02	.00	.02	.017	7.2	6.0–8.9	.206	.114–.304
vs. .0090 and .50	.91	.04	.02	.00	.00	.03	.013	31.0	22.1–46.1	**.000**	.000–.001
A.F.	.93	.01	.01	.02	.01	.02					
conn .0347, *dec* .62	.89	.05	.02	.01	.01	.03	.025	10.2	8.9–12.3	.068	.030–.114
vs. .0100 and .50	.94	.03	.01	.00	.00	.02	.011	15.7	11.8–22.8	**.008**	.000–.037
H.B.	.76	.05	.06	.02	.01	.09					
conn .0422, *dec* .67	.74	.09	.05	.02	.02	.08	.019	4.3	3.5–5.4	.511	.373–.621
vs. .0540 and .71	.74	.10	.06	.02	.02	.07	.025	6.3	5.4–7.7	.281	.176–.373
J.F.	.76	.09	.01	.09	.02	.02					
conn .0984, *dec* .86	.69	.14	.05	.04	.01	.07	.045	27.6	24.3–31.0	**.000**	.000–.000
vs. .1000 and .85	.77	.12	.02	.04	.00	.05	.030	55.1	41.8–80.9	**.000**	.000–.000
G.L.	.38	.02	.20	.03	.03	.34					
conn .0422, *dec* .71	.33	.09	.15	.02	.09	.32	.048	20.2	18.4–22.2	**.001**	.000–.002
vs. .0510 and .74	.35	.10	.15	.02	.09	.28	.051	21.6	20.0–23.4	**.001**	.000–.001
W.R.	.20	.08	.22	.01	.28	.20					
conn .0969, *dec* .93	.21	.10	.19	.03	.12	.35	.092	47.9	42.9–53.5	**.000**	.000–.000
vs. .1000 and .94	.18	.09	.18	.03	.12	.40	.105	51.5	47.3–56.5	**.000**	.000–.000

*Note:*The three response distributions in each block correspond to the patient, the parameters found by our optimization algorithm, and the parameters suggested by Dell et al. RMSD stands for root mean squared deviation. Boldface indicates significant mismatches.

Table 7: A comparison of the fitted parameter values for each patient's original naming performance and the performance obtained after a recovery period.

	Original		Recovery	
Patient	*Conn.*	*Decay*	*Conn.*	*Decay*
J.B.	.0453	.67	.0563	**.69**
J.L.	.0453	.67	**.0129**	.52
G.S.	.0057	.50	.0105	**.52**
L.H.	.0055	.50	.0344	**.65**
J.G.	.0470	.70	.0563	.69
A.F.	.0531	.71	**.0347**	.62
H.B.	.0375	.67	.0422	.67
J.F.	.0984	.86	.0984	.86
G.L.	.0734	.83	**.0422**	.71
W.R.	.0969	.94	.0969	.93

Note: Boldface indicates values contrary to the model's predictions.

as a baseline hypothesis. In other words, we will assume that all aphasic patients have the same repetition performance. For convenience, we will use the data as Dell et al. present it, rather than renormalizing without miscellaneous responses. Computing the mean of all the patients' responses in each category yields a baseline hypothesis of

correct	93%
semantic	0%
phonological	2%
mixed	0%
unrelated	0%
nonword	3%.

(Error types don't sum to one due to uncodable responses and rounding.)

This constant model yields an average RMSD of 0.016 (median 0.015). The fit for each patient is listed in Table 8. Removing from the analysis that patient for whom Dell et al.'s model's fit is worst (W.R.) improves its mean RMSD to 0.026 (median 0.023), while the performance of the baseline hypothesis remains unchanged. When predicting repetition performance, Dell et al.'s model fails to capture any of the phenomena in their data—it is less accurate than ignoring the patient's naming performance and guessing the mean. Its performance is comparable to what one would obtain by generating random patterns according to a distribution centered around the mean patient pattern. Clearly, this level of performance fails to provide support for Dell et al.'s model.

Summary of Empirical Results

Dell et al. (1997) propose a two-step, interactive model of lexical access in aphasic patients, and apply it to oral word production. They argue that global alterations of parameters in the model successfully mimicked the patterns of naming errors produced by

Table 8: Predicting repetition performance using a constant model.

	Repetition Response						
Patient	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD
Mean	.93	.00	.02	.00	.00	.03	
T.T.	.98	.00	.02	.00	.00	.00	.024
V.C.	.95	.00	.01	.00	.00	.04	.008
L.B.	.91	.00	.03	.00	.00	.06	.015
J.L.	.89	.00	.02	.00	.00	.03	.018
J.G.	.91	.00	.02	.01	.01	.05	.012
E.G.	.94	.00	.03	.00	.00	.01	.011
B.Mi	1.00	.00	.00	.00	.00	.00	.032
J.A.	.90	.00	.02	.00	.00	.08	.023
I.G.	.95	.00	.02	.00	.01	.02	.010
J.F.	.94	.00	.02	.01	.00	.03	.004
W.R.	.90	.00	.03	.01	.00	.06	.018

Note: RMSD stands for root mean squared deviation.

fluent aphasic patients. They interpreted the putative success of the lesioned model in fitting patient naming data as providing support for the central assumptions underlying the model, specifically, the interactivity assumption, the globality assumption, and the continuity thesis. However, we have now seen that Dell et al.'s model of aphasic naming cannot be justified by any of the empirical data they presented. We evaluated the model's fit to naming data in three different ways, ranging from formal statistical tests to informal visual displays:

Matching patient patterns: Even when Dell et al.'s model is fitted using a numerical optimization procedure, it cannot fit five of the twenty-one patients (24%). It also failed to match four of the ten patients who were tested after a recovery period, and the behavior of control participants.

Relative comparisons: A simple two-parameter linear model can match more patients, has a lower mean RMSD (root mean squared deviation), and has a higher weighted VAF (variance accounted for) and mean category VAF, indicating that Dell et al.'s model does a relatively poor job of accounting for the patient data, even if we don't insist on a match.

Global qualitative analysis: Plotting the range of patterns that the model can represent reveals that it cannot model several well-known patient classes, including any patient with only a moderate percentage of correct responses yet few nonword errors. Its adherence to the continuity thesis seems problematic.

We also evaluated the predictions that Dell et al. draw from the model, and found that:

The mixed error effect was not borne out in Dell et al.'s data,

The noun effect was correctly predicted by 24% of random models and is therefore not a useful criterion,

Recovery predictions from the model were correct in only 40% of the cases, which is not distinguishable from random prediction, and

Repetition predictions were worse than guessing the mean of the data, and comparable to guessing random patterns.

In short, patient data and simulation results do not provide support for Dell et al.'s two-step interactive model of aphasic naming. While its performance may be suggestive of avenues for future work, it cannot be used as validation of Dell et al.'s theoretical assumptions. (And even if we could identify a particular assumption that might be to blame for the model's failures, this would not provide validation for the remaining assumptions.) Our evaluation has provided a precise characterization of the model's performance, which will facilitate comparisons with future models (see, for example, Ruml, Caramazza, Shelton, & Chialant, submitted).

Methodological Considerations

We are now ready to turn to the other aspects of our investigation. In the introduction, we noted that besides a direct assessment of the model's fit to the patient data, there are two other ways we could evaluate Dell et al.'s claims. We can ask whether there are other empirical facts about naming deficits that are inconsistent with the model, and we can ask whether the inferences drawn by the authors on the basis of their results would be justified even if the model's fits to the data were good. We turn to these issues next.

Implications of Simulation Results

As we noted above, Dell et al. argue that the putative ability of their model to simulate the patterns of naming errors in aphasic patients can be taken as support for the assumptions of the model, such as the interactivity assumption, the globality assumption, and the continuity thesis. Given our demonstration that the model's fits to the patient data are poor, can we conclude that the central assumptions of the model are false? It seems clear that no such conclusion can be drawn from the model's poor performance. The failure could be due to any one of the model's major or minor assumptions, including but not limited to Dell et al.'s three central claims. But if one cannot conclude that empirical failure removes support for any specific aspect of the model, can one claim that empirical success would have provided it?

Consider Dell et al.'s claim that the putative success of their simulation research supports the assumption of interactivity (in the context of a two-step model of lexical access). Would such a conclusion have been justified by positive results? We think not. Empirical success would only have allowed the conclusion that the model as a whole is consistent with the observed results. In order to be able to begin to assign credit or blame to any one of the many assumptions of the model, we would have had to have much more specific evidence than Dell et al. have provided. In order to draw conclusions specifically about the assumption of interactivity, for instance, we would need to be certain that relaxing that assumption (by removing the connections allowing activation to spread backwards, for

example) would have adversely affected the model's ability to fit patients' naming error profiles. Given the deviations between Dell et al.'s model and the patient data, it is not clear that a model incapable of producing a mixed error effect would do worse. Similarly, in order to conclude that the simulation results support the globality assumption, the authors would have had to show that failure to implement the assumption (by testing different parameter values for different layers of the model, for instance) would have resulted in poorer fits to the data. (Indeed, broad comparative studies along these lines have now been carried out by Rapp and Goldrick (in press) and Ruml et al. (submitted).)

Furthermore, any conclusions about the necessity of an assumption, such as interactivity or globality, must necessarily be tentative unless it can be shown that no modification of the other, non-central assumptions of the model (such as the number of semantic nodes connected to each lexical node) would allow a model lacking that assumption to fit the data. Without such direct evidence that success depends critically on a specific feature or set of features of the model, no strong inferences can be made specifically about assumptions such as interactivity or globality. The only conclusions that are possible from evaluating a single version of the model is that the tested model as a whole is either consistent or inconsistent with the data. Thus, Dell et al.'s claim that the putative success of their simulation research provides support for both the interactivity assumption in lexical access and the globality assumption in naming deficits would be unwarranted even if the fits of the model to the data had been good.

Of course, this cautionary stance must also apply to those assumptions of the general theory that the authors describe but never subject to test in their model. For example, in describing their model, Dell et al. claim that

> lemma access is concluded by a selection process. The most highly activated word node of the proper syntactic category is selected. ... In the case of object picture naming, we assume a degenerate frame consisting of a slot for a single noun. In our implementation of the naming task, the most highly activated noun is selected. (p. 806)

While it is useful to know how Dell et al.'s model relates to their general theory of language production, it is important to note that no such mechanism was implemented in the model. All lexical nodes included in the model are nouns. Another example is provided by the authors' claim that

> ... the model assumes the existence of a layer of word nodes that is actively selected and controlled by syntactic processes. These processes also create a sizable nonlinearity in the network... (p. 829)

However, in the implemented model, a jolt of activation is applied whenever a node is selected, and no process beyond selection acts on the nodes. The only thing syntactic about the selection of lexical nodes is its label. Assumptions not implemented by mechanisms in the model cannot be tested by matching simulation results to patient data. Thus, no meaningful inferences are possible about putative syntactic properties of lexical nodes. The authors' allusions to such properties must be recognized as suggestive intuitions, rather than theoretical claims that could have been validated had the model been successful.

The Globality Assumption and Cognitive Neuropsychology

The globality assumption, as instantiated in the context of naming deficits, is the claim that aphasic errors result from global damage to all levels of the lexical access system. Dell et al. instantiated this assumption by altering parameters that affected all layers of their model equally. Although it is clear how the assumption was instantiated in Dell et al.'s model, it is less clear what substantive claim is being made by invoking this assumption. There are at least two possible readings. The claim could be that if a patient is classified as a fluent aphasic then his naming errors must reflect damage to all layers of the lexical access system. Alternatively, the claim could be much weaker: it could simply assert that if a patient is classified as a fluent aphasic then his naming errors could (though not necessarily) reflect damage to all layers of the lexical access system. We will examine both versions.

The Strong Globality Assumption.

We argued above that the failure of the model to fit the naming data of some patients does not allow us to reject the globality assumption, because of the possibility that it is some other aspect of the model that is responsible for such failures. When considering oral naming performance alone, certain patients do seem problematic. For instance, patients who make exclusively semantic errors, such as R.G.B (Caramazza & Hillis, 1990), would seem to pose difficulties. Dell et al. dismiss such problematic cases on the grounds that "...pure semantic patients are often associated with high rates of failure to name: no responses and semantic descriptions" (p. 832). However, patient R.G.B. always produced a response. He produced 69% correct responses, 15% semantic errors, and 16% descriptions. Dell et al. further attempt to circumvent the apparent problems raised by such cases for their globality assumption by arguing that

> ...this pattern does not necessarily require a nonglobal lesion. If nonattempts are construed as events in which the patient has retrieved a nonword or a word that is semantically unrelated to the picture, but has elected to suppress output, a global lesion is entirely consistent. (p. 832)

However, it is not clear what in Dell et al.'s theory of lexical access has the function of monitoring output in order to screen out phonological errors. What device would have the capacity to determine that the set of phonemes selected for output do not correspond to a word? And if we were to grant the feasibility of an output filter (a notion similar to that of Levelt (1989)), why wasn't it available to the patients who did produce phonological errors? Are we to assume that in the latter cases the filter was damaged? But if we were to make such a move, we would violate the globality assumption since we would have postulated differential damage in order to explain the two types of patients. That is, the patients who make semantic errors and phonological errors would have both the global damage hypothesized by Dell et al. as well as damage to the filter mechanism (so as to allow the production of phonological errors); the patients who make only semantic errors would only have the hypothesized global damage (so that the intact filter mechanism would be able to detect and suppress phonological errors). If instead the filter were operational in both cases, one would need to explain the difference in its effects without allowing such an explanation to function as an additional free parameter in the theory. In addition, it would be more difficult to motivate the interactivity assumption, since as Levelt (1989) has argued, a filter

Table 9: The modality-dependent naming performance of patients R.G.B. and R.C.M.

	Oral naming			Written naming		
Patient	Corr.	Sem.	Descr.	Corr.	Sem.	Other
R.G.B.	.687	.153	.16	.94		.06
R.C.M.	1.0			.53	.382	.088

Note: Data from Caramazza and Hillis (1990) and Hillis, Rapp, and Caramazza (in press), respectively.

can account for similar effects. It seems that the patients Dell et al. call pure semantic cases pose a serious challenge to their model even if we invoke an ad hoc filter mechanism.

But arguing about the model's ability to fit the naming performance of patients such as R.G.B. is beside the point. The oral naming performance of such patients cannot be used on its own to reject the strong version of the globality assumption. Dell et al. argue that

> globality as a substantive claim turns on whether the patterns observed in our sample are compatible with the model and whether the observed patterns represent a fair sampling of those present in the population at large. (p. 814)

But we would argue that the strong version of the globality assumption can be falsified only by demonstrating conclusively that there exist fluent aphasic patients in whom at least one level of the lexical access system is intact, a criterion that cannot be evaluated by considering a single task in isolation or by computational modeling, and has little to do with the incidence of such patients in the general population.

Such patients have been found, using the techniques of cognitive neuropsychology. An example can be constructed by comparing aphasic patients who make semantic errors in all modalities of input (visual, tactile, auditory, definition) and output (spoken, written) with those who make semantic errors in only one output modality. Since the former also make semantic errors in word comprehension tasks (Butterworth, Howard, & McLoughlin, 1984; Hillis, Rapp, Romani, & Caramazza, 1990), their configuration of performance invites the inference that the patients have damage to the semantic system—a component of processing shared by all the tasks tested. However, we cannot rule out that they also have damage to other levels of lexical access. But there are also fluent aphasics who make semantic errors in naming only in one modality of output and who show normal performance in word comprehension tasks. For example, Table 9 shows two such patients. Patient R.G.B. made semantic errors only in spoken production tasks. By contrast, patient R.C.M. (Hillis et al., in press) made semantic errors only in written production tasks. For these cases, we can be reasonably confident that the locus of damage cannot be the semantic system since it has to be intact in order to support normal performance in the spared output modality and in the comprehension tasks. This means that the naming errors in these patients must arise from damage to modality-specific lexical access mechanisms in the context of spared semantic processing.[7] This single example is enough to falsify the strong globality assumption: there

[7]This argument assumes a single semantic store for speaking and writing. We are unaware of evidence inconsistent with this hypothesis.

are fluent aphasics whose naming profile cannot be explained by the assumption of global damage to the lexical access system.

This example also illustrates a powerful and general procedure for establishing the locus of damage to a cognitive system. The approach is based on a core assumption in the cognitive sciences: the same cognitive mechanism may be recruited in the performance of many different tasks. By testing performance across tasks that are assumed to share some components of processing but not others, it may be possible to establish which portion of the system of interest is damaged. In our example, the assumption is made that naming with either oral or written responses recruits the same lexical access mechanisms and that, therefore, selective difficulties for one task must mean that large portions of the lexical access system are intact. If we had considered only performance on oral naming, we would not have been able to establish the locus of damage responsible for the semantic errors produced on that task. It is evident that the strong globality assumption is unfalsifiable if we restrict ourselves to the types of observations and research methodology adopted by Dell et al. The failure of the computational model cannot tell us anything about the globality assumption in particular. By focusing on performance on a single task (oral naming), we have no way of independently establishing the intactness of hypothesized components of a complex processing system. Dell et al. do not consider the kind of evidence that would have allowed them to empirically assess their claim.

The Weak Globality Assumption.

We have presented clear evidence that the strong globality assumption of naming errors in fluent aphasics is false. However, Dell et al. might have meant to advance a weaker claim. At one point in their paper, they seem to adopt the position that the globality assumption may only hold for some patients:

> ... we cannot fully endorse the globality assumption as a substantive claim about the functional basis of lexical retrieval disorders in aphasia. However, we do not find that the existing evidence compels rejection of the assumption either. Our view, based on the data reported here, is that the globality assumption works for a large enough segment of the population to merit further investigation. (p. 832)

If one were to adopt the latter position, the globality assumption would lose its substantive status and become a methodological prescription that is widely observed in cognitive neuropsychology research. On this view, Dell et al. would merely be saying that the fact that a particular patient makes disproportionately different rates of error types does not necessarily reflect differential damage to the levels of his lexical access system, but could equally well reflect damage to all layers equally.

The problem, then, becomes one of trying to distinguish whether a specific patient's error mix reflects a global lesion or arises from differential lesions to the different levels of his lexical access system. As already noted, this problem cannot be solved by considering oral naming performance alone. To distinguish between the two alternatives we would have to have independent evidence about the relative integrity of each of the hypothesized mechanisms in the lexical access process. And we have seen that this kind of evidence can only be obtained by considering each patient's performance across different tasks so as to allow us to converge on the locus or loci of damage. We conclude that whether one adopts

the strong or the weak version of the globality assumption, the evidence provided by Dell et al. is not adequate for its stated purpose.

The Continuity Thesis

Although we saw, when looking at possible error patterns, that patient naming behavior does not seem to form a continuum between normal performance and random error opportunities, it could be that Dell et al. meant only to make one of two more abstract claims. Dell et al. emphasize a quantitative aspect of the continuity thesis. They note that different patterns of impaired performance could reflect different degrees of damage to the normal system. Presumably this position is to be contrasted with the view that brain damage only affects cognitive mechanisms in an 'all-or-none' fashion, in which partial damage would render a mechanism completely non-functional. Although this is a logical possibility, it is not clear that anyone has ever championed this view, and it is certainly not a position that has any currency in modern cognitive neuropsychology.

Alternatively, the continuity thesis could be interpreted as the claim that impaired performance reflects the functioning of the mechanisms that are used in normal performance but that these ordinary mechanisms have been damaged. This view contrasts with the alternative claim that impaired performance involves mechanisms that are not used in normal processing (Kosslyn & Kleek, 1990, but see Carmazza, 1992). As Dell et al. see it,

> [the] continuity [thesis] implies that the model should provide a complete account of aphasic naming performance, without recourse to mechanisms that have sometimes been invoked by aphasiologists, such as neologism-producing devices or editors that are not part of normal production. (p. 830)

In the latter case, impaired performance would only have an indirect, perhaps impenetrably opaque relation to normal mechanisms and could not be used to constrain theories of normal functioning. In other words, it would violate the 'transparency assumption' that undergirds the cognitive neuropsychology enterprise (see Caramazza (1986) for discussion). Clearly, it is an empirical matter whether or not impaired performance reflects only the functioning of mechanisms used in normal performance that have been altered in some way. However, it is not the kind of empirical issue that is routinely tested by individual experimental projects the way we might test a specific theoretical claim. Rather it is the kind of background assumption that is used to motivate a whole enterprise, and is assessed not by the successes of individual projects but by the long-term success of the enterprise as a whole in providing theoretical insight to its domains of interest. In other words, we don't claim to have provided support for the transparency assumption (or the continuity thesis) each time we have a plausible explanation for a pattern of impaired performance; similarly, we don't claim to have reasons for rejecting the transparency assumption each time we are not able to provide an explanation for some pattern of performance. In the latter case, we are more likely to find fault with our theories. Thus, even if the model's fits to the aphasia data had not been poor, it would have made little sense to claim that the results of the research provided support for the continuity thesis.

Conclusions

We have evaluated Dell et al.'s computational model of lexical access in three different ways. First, we investigated its fit to the experimental data it was designed to explain. Using an automated regression procedure, we tried to match Dell et al.'s patient data. Even though we found many fits that were better than those reported by Dell et al., 24% of the patients could not be fit by the model. By comparing against a simplistic linear model of naming, we saw that this level of performance was relatively easy to attain, as the mathematical model was more accurate by every measure. And by sampling the behavior of Dell et al.'s model throughout the full range of its parameters, we saw that it was incapable of representing well-known patterns of patient naming. Although we considered the predictions that could be derived from the model, they failed to raise our evaluation, as we saw that the predictions were easy to make or inaccurate. We must conclude that Dell et al.'s model of aphasic naming enjoys little support from the empirical data.

Second, we investigated the central claim of globality, and noted that well-documented patients in the existing literature already disconfirm it. The continuity thesis found little support when taken literally, and became an untestable axiom of cognitive neuropsychology when interpreted loosely.

And third, we saw that the support that Dell et al. claimed to draw from their simulation results was not fortified by any investigation of models that lacked interactivity or globality, and hence the fit of their model to the data could not have bolstered those assumptions even if it had been good.

However, none of the flaws we have identified in Dell et al.'s core model is necessarily insurmountable. Although the globality assumption and the continuity thesis seem untenable, we have no reason to think that assumptions such as interactivity or localist representations are to blame for the model's failures. Indeed, our evaluation should be merely the first step in improving the model. Computational theories of cognitive behavior allow a precision of prediction that gives them great power. However, precision is not the same as accuracy, and the thorough evaluation methodology we advocate here will only become more important. Both formal tests of models' abilities to match human data and relative evaluations comparing against simple baseline models will enable quantitative measures of future progress. It is encouraging that Dell et al.'s model of aphasic naming was developed from the same general theory that can be used to construct individual models of many other behaviors. We hope that such models might, in the future, be unified to provide a comprehensive computational theory of the lexical system.

References

Badecker, W., Miozzo, M., & Zanuttini, R. (1995). The two-stage model of lexical retrieval: Evidence from a case of anomia with selective preservation of grammatical gender. *Cognition, 57*, 193–216.

Bock, K., & Levelt, W. (1994). Language production: Grammatical encoding. In M. A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 945–984). San Diego, CA: Academic Press.

Burke, D., MacKay, D. G., Worthley, J. S., & Wade, E. (1991). On the tip of the tongue: What causes word finding failures in young and older adults? *Journal of Memory and Language, 30*, 542–579.

Butterworth, B. (1979). Hesitation and production of verbal paraphasias and neologisms in jargon aphasia. *Brain and Language*, *8*, 133-161.

Butterworth, B. (1980). Evidence from pauses in speech. In B. Butterworth (Ed.), *Language production. volume 1: Speech and talk* (pp. 155–176). London: Academic Press.

Butterworth, B. (1992). Disorders of phonological encoding. *Cognition*, *42*, 261–286.

Butterworth, B., Howard, D., & McLoughlin, P. (1984). The semantic deficit in aphasia: The relationship between semantic errors in auditory comprehension and picture naming. *Neuropsychologia*, *22*, 409–426.

Caramazza, A. (1986). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*(1), 41–66.

Caramazza, A. (1992). Is cognitive neuropsychology possible? *Journal of Cognitive Neuropsychology*, *4*(1), 80–95.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology*, *14*, 177–208.

Caramazza, A., & Hillis, A. E. (1990). Where do semantic errors come from? *Cortex*, *26*, 95–122.

Cox, T. F., & Cox, M. A. A. (1994). *Multidimensional scaling.* London: Chapman and Hall.

Dell, G. (1986). A spreading activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dell, G. S. (1988). The retrieval of phonological forms in production: Test of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124–142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *4*, 313-349.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science*, *17*, 149–195.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. (1991). *Psychological Review*, *98*(4), 604–614.

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Language and Verbal Behavior*, *20*, 611-629.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. A. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review*, *104*(4), 801–838.

Gagnon, D. A., Schwartz, M. F., Martin, N., Dell, G. S., & Saffran, E. M. (1997). The origins of formal paraphasias in aphasics' picture naming. *Brain and Language*, *59*, 450–472.

Garrett, M. (1992). Disorders of lexical selection. *Cognition*, *42*, 143–180.

Harley, T. A. (1993). Phonological activation of semantic competitors during lexical access in speech production. *Language and Cognitive Processes*, *8*(3), 291–309.

Hillis, A. E., Rapp, B., & Caramazza, A. (in press). When a rose is a rose in speech but a tulip in writing. *Cortex*.

Hillis, A. E., Rapp, B., Romani, C., & Caramazza, A. (1990). Selective impairments of semantics in lexical processing. *Cognitive Neuropsychology*, *7*, 191–243.

Ho, Y. C., Sreenivas, R. S., & Vakili, P. (1992). Ordinal optimization in DEDS. *Journal of Discrete Event Dynamical Systems*, *2*, 61–68.

Humphreys, G. W., Riddoch, M. J., & Quinlan, P. T. (1988). Cascade processing in picture identification. *Cognitive Neuropsychology*, *5*, 67–103.

Kearns, M., Mansour, Y., Ng, A. Y., & Ron, D. (1997). An experimental and theoretical comparison of model selection methods. *Machine Learning*, *27*(1), 7–50.

Kosslyn, S. M., & Kleek, M. V. (1990). Broken brains and normal minds: Why humpty dumpty needs a skeleton. In E. Schwartz (Ed.), *Computational neuroscience* (pp. 390–402). MIT Press.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications.* New York: Springer.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation.* Cambridge, MA: MIT Press.

Levelt, W. J. M. (1992). Accessing words in speech production: Stages, processes, and representations. *Cognition*, *42*, 1-22.

Martin, N., Weisberg, R. W., & Saffran, E. M. (1989). Variables influencing the occurrence of naming errors: Implications for models of lexical retrieval. *Journal of Memory and Language*, *28*, 462–485.

Miozzo, M., & Caramazza, A. (1997). The retrieval of lexical-syntactic features in tip-of-the-tongue states. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 1410–1423.

Moore, A. W., Schneider, J. G., Boyan, J. A., & Lee, M. S. (1998). Q2: Memory-based active learning for optimizing noisy continuous functions. In *Proceedings of the international conference on machine learning.* San Francisco: Morgan Kaufmann.

Mosteller, F., Fienberg, S. E., & Rourke, R. E. K. (1983). *Beginning statistics with data analysis.* Reading, MA: Addison-Wesley.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes in C* (second ed.). Cambridge: Cambridge University Press.

Rapp, B., & Goldrick, M. (in press). Discreteness and interactivity in spoken word production. *Psychological Review.*

Roach, A., Schwartz, M. F., Martin, N., Grewal, R. S., & Brecher, A. (1996). The Philadelphia Naming Test: Scoring and rationale. *Clinical Aphasiology*, *24*, 121–133.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition*, *42*, 107–142.

Ruml, W., Caramazza, A., Shelton, J. R., & Chialant, D. (submitted). *Testing assumptions in computational theories of aphasia.*

Schriefers, H., Meyer, A. S., & Levelt, W. J. M. (1990). Exploring the time course of lexical access in language production: Picture-word interference studies. *Journal of Memory and Language*, *29*, 86–102.

Shepard, R. N., & Arabie, P. (1979). Additive clustering: Representation of similarities as combinations of discrete overlapping properties. *Psychological Review*, *86*(2), 87–123.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, *36*(2), 111-147.

Wilde, D. J. (1964). *Optimum seeking methods.* Englewood Cliffs, NJ: Prentice-Hall.

Appendix A: Optimizing the Model's Fit

In this appendix, we discuss our automated regression algorithm for fitting Dell et al.'s core model to patient data. Fitting simulation models like theirs is a difficult computational problem, known as stochastic optimization or stochastic approximation. At every setting of the input parameters, one can only determine an estimate of the model's output distribution, and hence, only an estimate of the fit to the desired distribution at those settings. (Similar problems arise in manufacturing settings, such as controlling a chemical plant.) Since most current methods either require large numbers of trials at many settings (Kushner & Yin, 1997) or sophisticated Bayesian statistical analysis (Moore, Schneider, Boyan, & Lee, 1998), we developed our own simple algorithm, tailored to our specific, low-dimensional problem. Although the details would take us far afield from our interests in this paper, we provide a sketch of the procedure. Since it is a general algorithm, we will first describe it in terms of a generic measure of fit, and afterwards discuss how we use it to optimize the X^2 metric in particular.

Since we have already accumulated many estimates of the output distribution of the model at various parameter settings (see the discussion of possible error patterns in the main text), we start our optimization algorithm from that previously-visited parameter setting that most closely matches the desired distribution. From there, our algorithm has the basic structure of a direction-set optimization method (Press, Teukolsky, Vetterling, & Flannery, 1992, p. 412), in which we carry out successive one-dimensional optimizations. However, instead of trying to determine an approximate gradient at the current best point, we only ask if adjacent points are significantly better. Fewer trials are required to resolve such relative comparisons than would be needed for a firm estimate of the rate of change, since the latter necessitates determining the size of the difference (Ho, Sreenivas, & Vakili, 1992). Although the lack of slope information forces us to use a simplistic bisection-style search strategy, such an approach protects us from problems caused by misestimation or asymmetry in the shape of the cost function (Wilde, 1964). Thus, the core computational task of our fitting procedure is to determine only whether a new candidate parameter setting is better than the current one, and not by how much.

We answer this question by sampling the output distribution at each setting, and comparing the two resulting distributions to the desired distribution using the measure of fit. However, since we have only a finite number of samples of each distribution, there is uncertainty regarding each goodness of fit value. We compute 95% confidence intervals around each value using Monte Carlo simulation (Press et al., 1992). This technique is based on estimating the distribution of fit scores around the one we actually got, taking into account that we only took x samples. We conduct 500 random experiments in which we consider the estimated distribution at this setting to be the true underlying multinomial. In each experiment, we randomly choose x samples according to the multinomial, and construct an empirical distribution from them. Measuring the fit of these 500 experimental distributions to the desired distribution gives a distribution over fit scores. By examining the 2.5th and 97.5th percentiles of this distribution of scores, we obtain estimates of the bounds of a confidence interval for the fit of the original samples.

When we have taken enough samples from each parameter setting, the two confidence intervals will cease to overlap and we can confidently choose one setting as superior. If the

intervals become very small (0.001 from the end of one to the end of the other), yet continue to overlap, we conclude that the new point is no better than the current one. And finally, if after many trials (10,000), the intervals have neither separated nor shrunken sufficiently, we rashly trust the current estimates as accurate, and choose the seemingly better setting.

Although this simple algorithm performs poorly when started in a desolate region of the search space where all nearby possibilities look equally poor, it seems to perform well for our purposes in this paper, since the prior sampling allows us to start the algorithm in an interesting part of the space. Comparisons against manual fitting on the naming data reported by Dell et al. can be found in the body of the paper, in the discussion of improving the model's fit and the discussion of the model's predictions regarding recovery.

Optimizing the X^2 Metric

Although the above procedure works well for measures of distributional similarity such as RMSD, it stumbles on metrics such as X^2 which can vary in value according to the number of samples taken. When comparing two fits which were estimated from different numbers of samples, the X^2 metric can be biased in favor of the fit estimated from fewer samples, since a failure to fit may not be discernible from a few samples. To remove this bias, we can use the fact that, since the number of samples from the target patient distribution is fixed, the X^2 value for a particular parameter setting should asymptotically approach a fixed value as more samples are taken. This is because the samples taken from the model will come to dominate the calculation of the expected number of samples in each response category, reducing the model's contribution to the X^2 value to zero, and causing the patient's contribution to converge to a fixed value. To compare fits with different numbers of samples, we therefore compute an estimate of this asymptotic value for each distribution. This can be easily done by using only the model's samples to determine the expected number of samples in each category, rather than using both the model's and the patient's. Rather than optimizing X^2 directly, our fitting algorithm instead optimizes this estimate of the asymptotic value of X^2. This works well, except for the case in which the model predicts no samples in a category, but the patient distribution contains samples in that category. In this case, the expected X^2 value is infinity, and our estimate attempts to divide by zero. Such situations may occur even at the optimal settings, and we would like the algorithm to choose the setting with the lowest actual X^2 value (even if it seems to be heading towards infinity). Therefore, we avoid infinite error by reverting to the ordinary X^2 estimation method for the troublesome category, using both the model and the patient samples. (This modified X^2 approach also seems to be the strategy taken by Dell et al. during their informal fitting, although they revert to RMSD when the model predicts zero in a category.)

Appendix B: A Linear Model of Naming

In this appendix, we evaluate a two-parameter linear model of patient naming patterns. Rather than postulating specific mental representations, the model just specifies that aphasia involves a systematic linear breakdown of language performance. This simple model can serve as a point of comparison for evaluating the performance of Dell et al.'s model. We will use it in the same general way as Dell et al.'s model. First, the model's adjustable parameters will be set to allow the model to fit a patient's pattern as closely as possible. Then, we will test, given those parameter settings, how well the model can regenerate the patient's pattern.

The form of the linear model embodies the claim that patient performance is composed of two phenomena that interact simply by summing. It assumes that each patient will vary from the mean patient response pattern to the extent that these two phenomena are particularly strong or weak. More formally, if the mean probability of a patient response of type i is m_i, and we notate the strengths of the two phenomena as y_1 and y_2, then patient performance will be predicted to be

$$\begin{aligned}
\text{correct} &= a_1 y_1 + b_1 y_2 + m_1 \\
\text{semantic} &= a_2 y_1 + b_2 y_2 + m_2 \\
\text{phonological} &= a_3 y_1 + b_3 y_2 + m_3 \\
\text{mixed} &= a_4 y_1 + b_4 y_2 + m_4 \\
\text{unrelated} &= a_5 y_1 + b_5 y_2 + m_5 \\
\text{nonword} &= a_6 y_1 + b_6 y_2 + m_6.
\end{aligned}$$

where the a_i and b_i are fixed constants representing the impact of the two phenomena on each response category. Each equation is a simple linear combination of the parameters y_1 and y_2.[8] The model has a simple geometric interpretation. When we vary y_1 and y_2, the equations will result in error patterns that lie on a two-dimensional plane slicing through the six-dimensional space of possible patterns. The position of each patient on the plane is specified by y_1 and y_2, and the orientation of the plane itself is specified by the a_i and b_i. Put another way, a two-dimensional plane in six-dimensional space can be fully specified by three points in that space, implicitly represented here by the a_i, b_i, and m_i.

To form a testable model from this schematic template, we need to choose specific values for the a_i and b_i in such a way that the resulting plane is as close as possible to the data. We chose these values using the downhill simplex optimization algorithm described by Press et al. (1992), attempting to minimize the sum of the X^2 values of the model from each patient. After computing the mean patient response in each category, we have the following complete model:

[8] Since the model isn't even assuming that the patient data represent probabilities, its predictions won't necessarily remain positive. To prevent the model from benefiting from negative terms during the calculation of X^2, we must force any negative response probabilities to zero. This clamping means that the model has a crease at the edge of the response space and is thus, speaking strictly, only piece-wise linear.

$$\begin{aligned}
\text{correct} &= 0.868y_1 + 0.198y_2 + 0.757 \\
\text{semantic} &= -0.071y_1 - 0.424y_2 + 0.046 \\
\text{phonological} &= -0.198y_1 + 0.296y_2 + 0.048 \\
\text{mixed} &= -0.045y_1 - 0.400y_2 + 0.031 \\
\text{unrelated} &= -0.203y_1 - 0.387y_2 + 0.031 \\
\text{nonword} &= -0.383y_1 + 0.601y_2 + 0.086.
\end{aligned}$$

To test this model, we need to determine y_1 and y_2 for each patient and then measure the model's ability to reconstruct the patient's response pattern from those two parameters. Since the model is linear, we can directly invert it to derive equations for y_1 and y_2.[9] These are:

$$\begin{aligned}
y_1 &= 0.868c - 0.383n - 0.203u - 0.198p - 0.071s - 0.045m - 0.604 \\
y_2 &= 0.601n - 0.424s - 0.400m - 0.387u + 0.296p + 0.198c - 0.172
\end{aligned}$$

where $c, s, p, m, u,$ and n represent the probabilities of correct responses and semantic, phonological, mixed, unrelated, and nonword errors for the patient, respectively. The type and amount of information lost by summarizing the six numbers as y_1 and y_2 will determine the accuracy of the model.

To see how this linear model works, let's take Dell et al.'s patient N.C. as an example. The distribution of N.C.'s picture naming responses is

correct	80%
semantic	3%
phonological	7%
mixed	1%
unrelated	0%
nonword	9%

First we need to compute our representation for N.C., using the equations for y_1 and y_2:

$$\begin{aligned}
y_1 = {} & (0.868 \times 0.80) - (0.383 \times 0.09) - (0.203 \times 0.00) - \\
& (0.198 \times 0.07) - (0.071 \times 0.03) - (0.045 \times 0.01) - 0.604 \quad = 0.039 \\
y_2 = {} & (0.601 \times 0.09) - (0.424 \times 0.03) - (0.400 \times 0.01) - \\
& (0.387 \times 0.00) + (0.296 \times 0.07) + (0.198 \times 0.80) - 0.172 \quad = 0.041.
\end{aligned}$$

Then, to assess the accuracy of this representation, we need to reconstruct the error mix that corresponds to $y_1 = 0.039$ and $y_2 = 0.041$ in the model:

$$\begin{aligned}
\text{correct} &= (0.868 \times 0.039) + (0.198 \times 0.041) + 0.757 = 80\% \\
\text{semantic} &= (-0.071 \times 0.039) - (0.424 \times 0.041) + 0.046 = 3\% \\
\text{phonological} &= (-0.198 \times 0.039) + (0.296 \times 0.041) + 0.048 = 5\% \\
\text{mixed} &= (-0.045 \times 0.039) - (0.400 \times 0.041) + 0.031 = 1\% \\
\text{unrelated} &= (-0.203 \times 0.039) - (0.387 \times 0.041) + 0.031 = 1\% \\
\text{nonword} &= (-0.383 \times 0.039) + (0.601 \times 0.041) + 0.086 = 10\%.
\end{aligned}$$

[9] The inversion depends on assumptions about the orthogonality of the two phenomena a and b. We will ignore these issues, assume orthogonality, and compute the inverse as the transpose of the matrix representing the projection. No additional parameters are required—the constants in the resulting equations are the dot products of the constants and coefficients from the original model equations.

As with Dell et al.'s model, the match is rarely exact—the reconstructed pattern must conform to the model's assumptions. Here we are assuming that N.C. lies somewhere on the plane defined by the model. We can multiply these reconstructed percentages by a large integer, such as 10,000, to obtain a distribution which we can compare to N.C.'s. This fit gives an X^2 value of 2.8 ($p = 0.73$), indicating that the linear model has successfully matched N.C.'s error pattern. The linear model is directly interpretable in terms of the patient error probabilities, and doesn't require manual fitting or numerical regression to produce each description.

When we follow the same procedure with each of Dell et al.'s patients, we find that the two-parameter linear model fits eighteen of the twenty-one patients (86%), with a mean RMSD of 0.015 (median 0.010), and a mean VAF of 78%. This is a surprisingly better fit than that of Dell et al.'s model (which fit sixteen patients with a mean RMSD of 0.028).

However, recall that we chose most of the fixed parameters of the model (the a_i and b_i) by explicitly attempting to optimize the match to the patient data. While this is methodologically sound if we merely want to show that the data are well-captured by linear equations, and hence easy to model, one could argue that it would be unfair to directly compare that model's performance to the performance of Dell et al.'s model, since the non-adjustable parameters of Dell et al.'s model were fixed without reference to the patient data that the model would be tested against. If we had a large supply of patient data, we could avoid this problem by testing the linear model on different patients than we used for constructing it. (There is nothing unsound about using patient data to construct the model, as long as different patients are used for testing.) Since we have only a small number of patients, we must use the standard testing technique of cross-validation, in which we use many partitionings of the data we have in order to estimate the accuracy we would obtain if we had new data (Kearns, Mansour, Ng, & Ron, 1997; Stone, 1974). We will repeatedly construct a linear model on the basis of twenty patients, and then test the model on the remaining patient, whose error pattern was not used in constructing the model. Since we aren't allowing ourselves to use all of the data when constructing the model, this leave-one-out technique should give us a slightly pessimistic estimate of the generalization accuracy of the full model on new data.

Results for all patients are shown in Table 10. The two-parameter linear model fits seventeen of the twenty-one patients (81%), and has a mean RMSD of 0.022 (median 0.015), a weighted VAF of 89%, and a mean category VAF of 65%. The VAF in the individual categories is

correct	99%
semantic	28%
phonological	80%
mixed	57%
unrelated	33%
nonword	92%.

Like Dell et al.'s model (the fits of which were shown in Table 3), the linear model does best in the frequent correct and nonword categories. It consistently outperforms the mean, even in the semantic category. Using each of the three metrics that we have considered (X^2, RMSD, VAF), the linear model is more accurate than Dell et al.'s two-parameter

Table 10: Fits of the linear model to Dell et al.'s patients.

Patient and	Naming response						Fit				
parameter values	Corr.	Sem.	Phon.	Mixed	Unrel.	Non.	RMSD	X^2	X^2 conf.	p	p conf.
W.B.	.94	.02	.01	.01	.00	.01					
linear model	.94	.02	.01	.02	.00	.01	.003	0.4	0.2–0.8	.996	.975–.999
T.T.	.95	.01	.01	.02	.00	.00					
linear model	.91	.04	.02	.03	.00	.00	.021	4.4	3.8–5.3	.487	.378–.583
J.Fr.	.93	.01	.01	.02	.00	.02					
linear model	.92	.01	.01	.02	.00	.03	.005	0.5	0.2–1.0	.992	.966–.999
V.C.	.92	.02	.01	.01	.00	.03					
linear model	.89	.03	.02	.02	.00	.03	.013	1.7	1.2–2.3	.894	.813–.948
L.B.	.82	.04	.02	.01	.01	.09					
linear model	.78	.04	.06	.03	.01	.09	.026	6.4	5.5–7.4	.269	.192–.358
J.B.	.83	.06	.01	.03	.01	.06					
linear model	.81	.05	.03	.03	.02	.05	.012	3.0	2.4–3.8	.694	.578–.787
J.L.	.85	.03	.01	.03	.01	.06					
linear model	.82	.04	.03	.03	.02	.06	.015	2.6	2.1–3.4	.759	.643–.840
G.S.	.73	.02	.06	.01	.02	.15					
linear model	.72	.03	.08	.01	.03	.14	.010	1.1	0.7–1.8	.952	.882–.983
L.H.	.71	.03	.07	.01	.02	.15					
linear model	.71	.03	.08	.01	.03	.14	.006	0.6	0.3–1.0	.989	.958–.997
J.G.	.59	.06	.09	.04	.03	.20					
linear model	.59	.05	.09	.03	.07	.17	.021	6.0	4.9–7.2	.311	.206–.434
E.G.	.94	.03	.00	.02	.00	.01					
linear model	.94	.03	.00	.02	.00	.00	.005	3.0	2.0–4.8	.699	.440–.854
B.Me.	.89	.03	.01	.05	.01	.00					
linear model	.89	.05	.01	.04	.01	.01	.011	5.7	4.6–7.6	.340	.183–.468
B.Mi	.88	.05	.01	.02	.01	.01					
linear model	.88	.04	.01	.04	.01	.02	.007	1.8	1.2–2.7	.879	.753–.941
J.A.	.88	.05	.00	.03	.01	.03					
linear model	.84	.05	.02	.04	.02	.03	.020	4.1	3.5–4.8	.542	.441–.627
A.F.	.78	.02	.03	.06	.04	.07					
linear model	.78	.05	.04	.03	.04	.07	.016	6.8	5.2–8.6	.238	.125–.390
N.C.	.80	.03	.07	.01	.00	.09					
linear model	.78	.03	.05	.02	.02	.10	.015	5.5	4.6–6.7	.358	.245–.461
I.G.	.77	.10	.06	.03	.01	.03					
linear model	.77	.06	.04	.05	.05	.04	.026	13.0	11.3–15.2	**.024**	.009–.046
H.B.	.61	.06	.13	.02	.01	.18					
linear model	.64	.03	.09	.01	.05	.18	.029	14.2	12.3–16.7	**.014**	.005–.031
J.F.	.66	.16	.01	.13	.01	.03					
linear model	.68	.09	.03	.08	.10	.02	.051	25.5	23.3–28.3	**.000**	.000–.000
G.L.	.29	.04	.22	.03	.10	.32					
linear model	.31	.06	.16	.02	.13	.33	.032	7.6	6.3–9.3	.181	.099–.277
W.R.	.08	.06	.16	.05	.35	.30					
linear model	.04	.14	.22	.09	.10	.41	.122	123.1	112.2–133.9	**.000**	.000–.000

Note: RMSD stands for root mean squared deviation. Boldface indicates significant mismatches.

representation.

As one would expect, a linear model with only one adjustable parameter gives a worse fit to the data. Using leave-one-out cross-validation, a linear model succeeds in matching fourteen of the twenty-one patients (67%), with a mean RMSD of 0.034 (median 0.022), a weighted VAF of 54%, and a mean category VAF of 37%. This result quantifies the increased accuracy gained by characterizing aphasic naming by two linear parameters rather than one. It also starts to give us some intuition for reasonable values of RMSD.

Of course, the simple superficial linear representation we have used as a baseline for comparison cannot be taken seriously as a theory of language production, since it does not identify any mental representations or processes. It is presented solely as a way of determining whether it is particularly difficult to match sixteen of the twenty-one patients. We have concluded that it is not, since this level of performance can be surpassed by a simple linear model.