Running Head: THEORY SELECTION AND EVALUATION IN CASE SERIES

Theory selection and evaluation in case series research

Matthew Goldrick
Department of Linguistics
Northwestern University

Address for Correspondence:
Matthew Goldrick
Department of Linguistics
Northwestern University
2016 Sheridan Rd.
Evanston, IL 60208 USA
Phone: (847) 491-8053
Fax: (847) 491-3770
Email: matt-goldrick@northwestern.edu
Theory selection and evaluation in case series

Abstract

Using empirical data to develop theories requires not only evaluating how well a theory accounts for data; it requires using the data to select the best theory from among a set of alternatives. Current case series research is examined in light of these two issues. Theory selection requires that theories make contrasting predictions. In the first section of this commentary, I present novel simulation results showing that existing theories of language production do not make contrasting predictions for the overall distribution of responses over a set of responses categories (e.g., correct response, semantic error, etc.; Dell, Schwartz, Martin, Saffran, & Gagnon, 1997). Given such results, in order to be theoretically productive case series research must focus on those aspects of data that serve to contrast theoretical alternatives. The second section considers evaluation of claims regarding individual differences. Such claims are typically under-constrained. Two approaches to addressing this issue are discussed. I argue that case series research should provide independent evidence for hypothesized individual differences. Second, parametric approaches might provide a means of constraining theories of individual differences. The plausibility of this approach is examined through novel analyses of empirical distributions of individual differences in impairments to lexical access (Schwartz, Dell, Martin, Gahl, & Sobel, 2006).
**Theory selection and evaluation in case series research**

As we develop and test theories of cognition based on empirical data, it is critical that we be able to precisely assess our theories. An important issue that has received substantial attention in the case series literature is *theory evaluation*. How well does a given theory account for the data? For example, can it account for the distribution of patterns of impairment? Equally critical is addressing the issue of *theory selection*. Out of a range of theories, which provides the best account of the empirical data? For example, does a theory that assumes neurological impairment affects particular subcomponents of the speech production system provide a better account than a theory that assumes impairment always affects global properties of processing?

The first section of this commentary examines the issue of theory selection in case series research. To be able to select among alternative theories A and B, it must be the case that the theories make contrasting predictions for a given type of data. If this is the case, there must be some set of observations X that theory A predicts should occur and theory B predicts should *not* occur. In other words, theory B will be unable to account for certain data predicted by theory A. The question of whether theories actually make contrasting predictions is frequently unexamined in case series research. To demonstrate that this presents a serious issue for cognitive neuropsychological investigations, simulations are used to show that extant, prominent theories of speech production fail to make contrasting predictions regarding the overall distribution of responses over a set of categories (utilized by Dell, Schwartz, Martin, Saffran, & Gagnon, 1997, et seq., to test theories of speech
production). In order to support theory selection, case series research must focus on those aspects of data that discriminate theories.

The second section examines a pervasive issue in evaluating whether the theories examined in case series research account for a given set of data. Individual variation—both pre- and post-morbid—has played a critical role in many theories built on case series research. However, in many cases, case series research has failed to provide any independent evidence to constrain hypotheses regarding individual variation. How can this issue be addressed? Existing methods for providing independent evidence are reviewed. Finally, the possibility of using a parametric approach to help constrain theories of individual variation is discussed.

Contrasting Predictions and Theory Selection

Many case series studies use simulations to evaluate the ability of a single theory to account for existing sets of data (e.g., Dell et al.’s 1997 analysis of picture naming; Plaut, Seidenberg, Patterson, & McClelland’s 1996 analysis of reading). However, such studies do not allow us to determine if this theory provides a better account of the data than an alternative account.

Case series research inspired by Dell et al. (1997) has taken steps in this direction. To select among contrasting theories, work inspired by Dell et al. has examined the degree to which simulations of different theories can fit the overall response distribution of each participant; that is, the proportion of their responses that fall into each one of a set of response categories (e.g., correct production, semantic error, phonologically related error, etc.). For example, one commonly used measure (root mean square deviation, defined below) examines the average
deviation, across all response categories, of the rate at which the simulation (with the best-fitting parameter settings) produces responses in that category vs. the rate at which a participant produces responses in that category.

The specific domain that this work has focused on is impairments to speech production processes. Foygel and Dell (2000) and Schwartz, Dell, Martin, Gahl, and Sobel (2006) examined the degree to which simulations of contrasting theories of functional impairment (global vs. local disruptions to the speech production system) could match the overall response distributions observed in case series data. Dell, Lawler, Harris and Gordon (2004) used similar metrics to contrast alternative theories of how omission errors arise during speech production. Dell, Martin, and Schwartz (2007; see also Dell et al., 1997) used measures of fit to overall response distributions to contrast models of repetition. Finally, Rumel, Caramazza, Shelton, and Chialant (2000) and Rumel, Caramazza, Capasso, and Miceli (2005) used such measures to examine contrasting assumptions regarding the nature of functional impairment and the degree of interaction between functional components of the speech production system.

This body of work makes several critical assumptions. Following Dell et al. (1997), it assumes that one can adjudicate amongst theories based on the performance of the corresponding simulations. The analyses below also adopt this assumption (but see McCloskey, 1993, for a critical discussion). The assumption that is critically examined here is that distinct theories predict distinct, contrasting patterns of overall response distributions. This is certainly true of some theoretical contrasts. Foygel and Dell (2000) find that simulations of theories with global vs.
local disruptions to the speech production system predict distinct overall response distributions. For example, empirically observed cases where only phonological errors are produced can be generated by a simulation incorporating only local disruption to the production system—but not by a simulation utilizing global disruption. But does the assumption that contrasting theories predict contrasting patterns of overall response distributions hold more generally?

To examine this question, a baseline theory and two contrasting theories (drawn from the literature) are considered. Using patterns of overall response distributions, could a case series study successfully determine if the baseline theory is correct or incorrect? Rather than utilizing real data—where the true, correct theory is not known—this question can be considered using simulated data. Specifically, we can generate artificial case series corresponding to the baseline theory and the two contrasting theories. We can then assess whether fit to overall response distributions serves to distinguish cases where the baseline theory is known to be correct (i.e., artificial data generated by the baseline theory) vs. cases where the baseline theory is known to be incorrect (i.e., artificial data generated by the two contrasting theories). The results reveal that this is not the case; fit to overall response distributions is comparable when the baseline theory is correct vs. incorrect.

**Contrasting theories of speech production and speech production deficits**

Figure 1 illustrates the three theoretical alternatives considered here. The baseline theory is the two-step interactive account of impairments to lexical access in speech production (Dell et al., 2004; Foygel & Dell, 2000; Schwartz et al., 2006).
This theory uses a spreading-activation architecture to model speech production processes that occur subsequent to conceptual processing (see below for further discussion of this stage). Representations are instantiated by simple processing units corresponding to units of linguistic representation: semantic features, lexical items, and phonemes. The theory is interactive in that activation spreads both forward and backward between ‘adjacent’ levels of representation. There are bidirectional connections between semantic features and lexical items as well as between lexical items and phonemes. Following other work in speech production, this theory claims that subsequent to conceptual processing speech production involves two steps or stages of pre-articulatory processing. At the end of the first stage, a representation of a lexical item is selected; when processing is intact, this almost always corresponds to the intended semantic representation (e.g., <CAT> for <furry>, <pet>, <feline>). At the end of the second stage, sound representations are selected; during intact processing, this almost always corresponds to the selected lexical item (e.g., /k/ /ae/ /t/ for target <CAT>). Speech production impairments are accounted for within this particular theory by reducing the flow of activation between adjacent levels of representations—weakening the bidirectional connections between semantic features and lexical items and/or the connections between lexical items and phonemes. This allows random noise in the system to make non-target representations more active than those of the target—resulting in errors.
Figure 1. Baseline theory and two contrasting theories. Arrows indicate the direction of activation flow between representations connected by lines; size of arrows indicate the strength of activation flow. Dotted lines indicate weakening of activation flow between representations, which leads to errors. Dashed circles highlight the differences between the contrasting theories and the baseline theory.
The predictions of this baseline theory were contrasted with those of two alternatives. The first alternative violated the baseline theory's damage assumptions by assuming a conceptual processing impairment. Prior to the two stages of processing outlined above, speech production theories assume that conceptual processes activate the appropriate semantic features. Proponents of the two-step interactive account have claimed that disruption to conceptual processes results in distinct patterns of overall response distributions than those produced by disruption to the two stages outlined above. For example, the two-step interactive account exhibited a poor fit to the overall response distributions of the “JF” subgroup of Schwartz et al.’s (2006) case series. Based on this poor fit—and the predominance of semantically-related errors in the responses of these individuals—Schwartz et al. argued that these individuals suffered from deficit to conceptual processing. Is the logic of Schwartz et al. appropriate? Do theories that allow for conceptual processing impairment predict overall response distributions that are distinct from those predicted by the baseline theory? To examine this question, the simulation architecture of Schwartz et al. (2006) was augmented to include conceptual processing (following Rapp & Goldrick, 2000). The theory's predictions were then estimated by disrupting conceptual processing in the simulation along with the two steps of the baseline account.

The second alternative theory—Rapp and Goldrick’s (2000) Restricted Interaction Account (RIA)—violated the baseline theory’s architectural assumptions. In contrast to the two-step interactive account, this alternative theory assumes (a) feedback is absent from lexical representations to semantic features
and (b) feedback from phonemes to lexical representations is weaker than the feedforward flow of activation from lexical representations to phonemes. As discussed in more detail below, Rapp and Goldrick (2000) showed that RIA made a number of contrasting predictions from the two-step interactive account. However, the extent to which it predicts overall response distributions that contrast with those of the two-step interactive account has not been examined. To examine this, a simulation embodying RIA’s assumptions regarding lexical access was constructed; processing was then disrupted as in the baseline simulation.

**Simulation Results: Contrasting Patterns of Performance**

As detailed in the Appendix, a simulation framework was constructed and used to implement all three theories. Artificial case series were then generated by damaging process within each simulation. In the simulation of the baseline theory, two sets of connection weights were weakened by random amounts: connections between semantic and lexical representations and connections between lexical and phonological representations. The simulation of the theory that violated the baseline theory’s assumptions regarding damage involved additional random reductions to the connections between conceptual and semantic representations. Finally, simulations of the theory that violated the baseline theory’s architectural assumptions involved altering the “normal” strength of connection weights in the simulation of the baseline theory. Damage in these simulations then followed that of the baseline theory. The artificial case series generated by these three simulations (each consisting of 1,000 simulated patients) covered a wide range of accuracy.
levels (baseline: 3.1-96.5%; violation of damage assumptions: 2.9-96.1%; violation of architectural assumptions: 3.2-92.7%).

As discussed above, the critical question for our analysis is whether fits to overall response distributions could be used to determine if the baseline theory is correct or incorrect. Before addressing this point, it’s important to demonstrate that the theories do, in fact, differ in some way. If there’s no possible way to distinguish the theories, the issue is not specific to overall response distributions—the theories are simply indistinguishable.

Analysis of these case series showed that the simulations did exhibit contrasting patterns of performance (reflecting their distinct assumptions). Deficits to conceptual processing primarily lead to semantically related errors; during this stage of processing, form representations are not strongly activated. Because of this, simulations that incorporated damage at this level—violating the baseline theory’s damage assumptions—had a higher rate of semantic errors (mean rate of purely semantic and mixed semantic/phonological errors: 12.6%) than simulations corresponding to the baseline theory (mean: 7.9%; t(1998) = 21.6, p < .001).

With respect to the simulations violating the baseline theory’s architectural assumptions, Rapp and Goldrick (2000) documented how this theory made contrasting predictions from the two-step interactive account with respect to the mixed error effect. Errors that have a mixed semantic and phonological relationship to the target (e.g., “shirt” → “skirt”) are found to occur more often than predicted based on the rates of purely semantic (e.g., “shirt” → “pants”) or purely phonological errors (e.g., “shirt” → “hurt”). Rapp and Goldrick found that the size of this mixed
error effect at lexical and semantic levels of processing was related to the degree to which phonological representations were able to influence activation of units at these levels. For example, at the lexical level, the relative advantage of <SKIRT> over <PANTS> (for target <SHIRT>) will be a function of how much of an activation boost the former receives via feedback from the phonemes it shares with the target.

These predictions were confirmed in the present simulations. Due to the reduction of the strength of feedback from phoneme to lexical units (and the absence of feedback to semantic representations), simulations of the theory violating the baseline theory’s architectural assumptions had a weaker mixed error effect than the baseline simulations. The size of the mixed error effect was approximated by comparing the relative proportion of semantic errors that were mixed errors. Across simulations violating the baseline’s architectural assumptions, a mean of 19.3% of semantic errors were mixed; in the baseline simulations, the mean was 25.3% (t (1998) = 14.6, p < .0001). (Note that both of these well exceeded the simulation architecture’s chance level of 10%.)

Do these theories predict distinct overall response distributions?

The results above show that these theories do in fact make contrasting predictions with respect to certain behaviors. Does this also hold for overall response distributions? If so, the simulation of the baseline should provide an excellent fit to the overall response distributions of simulated individuals in the artificial case series generated under its assumptions. In contrast, the baseline simulation should have marked difficulty fitting artificial case series generated by the simulations of the contrasting theories. To examine this question, the
parameter-fitting procedure of Dell et al. (2004) was used to fit a simulation of the baseline theory to the overall distribution of each simulated individual in each of the artificial case series. The appendix provides details of this fitting procedure. Note that in contrast to the simulations used to generate the artificial case series, this baseline simulation did not implement conceptual processing mechanisms. It assumed that conceptual processes were completely intact and activated only the semantic features of the target.

Following Schwartz et al. (2006), three measures of fit to the overall distribution of each simulated individual were utilized. For each simulated patient \( i \), root mean squared deviation (RMSD) was defined as:

\[
RMSD_i = \sqrt{\frac{\sum_{c} (s_{ic} - m_{ic})^2}{6}}
\]

where \( s_{ic} \) is the proportion of responses in category \( c \) (out 6 possible response categories: correct, semantically related, mixed, unrelated, formally/phonologically related, and nonword) for simulated patient \( i \). \( m_{ic} \) is the proportion of responses in the best fitting baseline simulation in category \( c \) for simulated patient \( i \). Intuitively, RMSD can be thought of as the average deviation in response proportions between the baseline simulation and the simulated patient.

The mean RMSD across the entire set of simulated patients provides one measure of the typical fit of the baseline simulation to the overall response distribution of a member of the artificial case series. A second means of assessing overall fit to a case series using RMSD is the number of deviant or outlier data
points. Following Schwartz et al. (2006) these were defined as simulations where the RMSD met or exceeded 0.041.

A third measure of fit for a particular artificial case series is total variance accounted for (VAF):

\[
VAF = 1 - \frac{\sum_{i,c} (s_{ic} - m_{ic})^2}{\sum_{i,c} (s_{ic} - \bar{s}_c)^2}
\]

(2)

where \( \bar{s}_c \) is the mean proportion of responses in category \( c \) across all simulated patients within this artificial case series. Intuitively, the amount of variance that any model could account for is the amount of variance that simulated patients exhibit around the average response proportion in each category. VAF characterizes how much of this total possible variation the baseline simulation accounts for. A high VAF therefore means that the simulation provides a good fit to the data.

The fits of the baseline simulation to each artificial case series are shown in Table 1. As expected, the fit to the case series generated by the baseline simulations was extremely good. However, the baseline simulation also had extremely good fits to the other artificial case series: one generated by simulations that violated its assumptions regarding damage and another that violated the baseline theory’s assumption regarding the architecture of the lexical access system. Although these fits were slightly lower than those of the baseline set, they were very close to perfect. In fact, relative to results of other studies, these fits are extremely good. For example, in the set of 94 patients analyzed Schwartz et al. (2006) the mean RMSD of the baseline simulation was 0.024, the proportion of deviant cases 0.17 and the total
VAF was 94.4%. Relative to the fits to empirically observed case series, in this work the baseline simulation exhibited excellent fits to each of the simulated case series.
<table>
<thead>
<tr>
<th></th>
<th>Mean RMSD</th>
<th>Proportion Deviant</th>
<th>Total VAF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>0.01</td>
<td>0.005</td>
<td>98.8%</td>
</tr>
<tr>
<td>Violation of damage assumptions</td>
<td>0.017</td>
<td>0.01</td>
<td>97.6%</td>
</tr>
<tr>
<td>Violation of architectural</td>
<td>0.017</td>
<td>0.01</td>
<td>97.5%</td>
</tr>
</tbody>
</table>

Table 1. Mean root mean square deviation (RMSD; best fit = 0), proportion of deviant simulations (RMSD >= 0.041) and total Variance Accounted For (VAF; best fit = 100%) for the baseline simulation fits for each set of artificial case series.

Figure 2 provides a graphical representation of the baseline simulation’s fit to each individual simulated patient’s responses within each category (following the method of Schwartz et al., 2006. Figures 4 and 5; note scale matched to their figures). The difference between observed and predicted response proportion in each category is shown. In all three artificial case series, the baseline simulation slightly underpredicted the rate of semantic errors. This may be due to the fact that the simulations that generated the artificial case series included conceptual selection processes. As noted above, the simulation that was used to fit the overall response distributions did not implement these mechanisms, assuming instead that in all cases only the target semantic features were activated. For other error categories, there were no systematic deviations, although the fit to formal and
unrelated errors categories was somewhat more noisy than other categories.

Critically, the baseline simulation’s fit to the baseline case series was not systematically, qualitatively better than its fit to case series generated by simulations violating its assumptions.
Figure 2. Observed – predicted proportion of responses in each response category for each artificial case series. A) Baseline simulations. B) Violation of damage assumptions. C) Violation of architectural assumptions.
Discussion

Selecting among competing theories is a critical part of evaluating whether data support a given cognitive hypothesis. If some type of data is to be used select among alternative theories, it must be the case that the theories make contrasting predictions with respect to this type of data. The simulation results illustrate this general point for the specific measure of overall response distributions. The contrasting theories examined here do not make contrasting predictions for this type of data; it therefore cannot be used to distinguish these theories.

To utilize data to select among competing theories, it is necessary to identify those aspects of impaired performance along which theories make distinct predictions; one must focus in on the behavioral *contrasts* that distinguish contrasting theories. For example, Rapp and Goldrick (2000) examined theories of speech production varying along the discreteness-interactivity dimension. Simulations of contrasting theoretical architectures were utilized to examine whether certain measures did or did not discriminate theories with contrasting assumptions. For example, the simple presence of a mixed error effect was consistent with a wide range of theoretical architectures. In contrast, the observation of mixed error effects following deficits to *particular* levels of the processing system (i.e., stage 1 of the two-step account; lexical selection) was consistent with a much smaller range of theories. Similarly, Nozari, Kittredge, Dell and Schwartz (2010) found that distinct theoretical architectures made similar predictions regarding the magnitude of word frequency effects in naming and
repetition—revealing that this measure was not an appropriate way to discriminate among the specific hypotheses they considered.

**Evaluating Theories’ Claims Regarding Individual Variation**

In case series research, individual variation plays a key explanatory role. How can we evaluate whether theories’ claims concerning individual variation are empirically accurate? To frame this discussion, the two types of individual variation (pre- and post-morbid) utilized in the studies discussed by Schwartz and Dell (2010) are reviewed. As shown by consideration of a number of case series studies, there is frequently a lack of independent evidence for such claims; these studies utilize no constraints on hypothesized individual variation beyond the observed impaired performance.

How can this issue be addressed in case series research? One is to directly address the issue by providing independent evidence for individual differences; current approaches to this problem are reviewed. A complementary method would serve to constrain the range of hypotheses regarding individual variation. Instead of relying solely on the observed data to determine the nature of individual variation, such methods would provide an additional source of constraint on inferred individual differences. Using empirical data from Schwartz et al. (2006), the final section examines the possibility of using a parametric approach to provide this type of constraint on individual variation.

**Types of individual variation in case series research**

Post-morbid individual variation is most commonly utilized in case series research. (n.b. Such variation has been implicitly assumed by case studies research
that allows for varying degrees of severity of damage to functional components.) For example, in studies utilizing the two-step interactive theory (e.g., Dell et al., 1997), the distribution of empirically observed cases reflects not only variation in the loci of impairment but quantitative variation in damage to cognitive processes (e.g., variation in the strength of weights connecting levels of representation). Related work has argued that post-morbid individual variation provides a superior account to theories assuming categorical differences in the locus of functional damage. For example, Mirman, Yee, Blumstein and Magnuson (2011) used the visual world paradigm to examine how recognition of a target word was influenced by words that share sounds (lexical neighbors; e.g., cap, hat for target cat). At a group level, they found that individuals with Broca’s aphasia profiles had difficulty processing targets in the presence of neighbors sharing final sounds (e.g., hat for target cat) whereas individuals with Wernicke’s aphasia profiles had difficulty with neighbors sharing initial sounds (e.g., cap). This pattern could be accounted for by assuming a categorical difference in functional deficits across the two groups. One deficit leads to enhanced competition from neighbors sharing final sounds; another deficit to a distinct functional component leads to competition for neighbors sharing initial sounds. However, Mirman et al.’s case series analysis suggested that across individuals there was a negative correlation between the degree of interference from the two types of neighbors. A theory that assumes the two effects arise due to independent functional deficits cannot explain such an relationship. Mirman et al. offer an alternative account that assumes these performance patterns reflect continuous variation along a single dimension of processing impairment. Simulation
results show that variation along this dimension leads to shifts in the relative
strength of interference from different types of neighbors—providing an account of
the correlation observed in their case series.

Other case series research has claimed pre-morbid individual differences in
the cognitive architecture strongly influence patterns of impaired performance.
Jeffries, Rogers, and Lambon-Ralph (2011) used pre-morbid differences in category
expertise to predict patterns of semantic impairment. In most semantic categories,
two individuals with semantic dementia showed an impairment in knowledge of
subordinate category distinctions (e.g., the distinction between *dachshund* and
*poodle*). However, they showed relatively selective sparing of subordinate
knowledge in a domain in which they had pre-morbid expertise. One individual
showed selectively preserved knowledge of cars whereas another showed
selectively preserved knowledge of plants.

Other work incorporates both pre- and post-morbid individual variation to
account for impaired performance patterns. Plaut et al. (1996) find that their
simulations are only able to account for the full range of impairment patterns in
reading if there are individual pre-morbid differences in the contribution of
semantically-based processes to reading aloud and post-morbid differences in the
degree of damage to semantic and non-semantic processes. This pre-morbid
difference is hypothesized to be related to a number of factors, including “extent of
reading experience... the nature of reading instruction, the sophistication of
preliterate phonological representations, relative experience in reading aloud
versus silently, the computational resources devoted to each pathway, and the
reader’s more general skill levels in visual pattern recognition and in spoken word comprehension and production (pp. 97-98).” Wollams, Lambon-Ralph, Plaut, and Patterson (2007) show that this proposal provides a good quantitative fit to a separate set of individuals with semantic dementia. Plaut (1997) extends Plaut et al.’s (1996) proposal to include pre-morbid individual variation in weight decay, a network parameter that he links (p. 796) to “the degree to which the underlying physiology can support large numbers of synapses and, hence, strong interactions between neurons.” Dilkina, McClelland, and Plaut (2008) adopt a distinct approach to pre-morbid differences. Following Plaut et al. (1996), they incorporate pre-morbid differences in reading experience; in contrast to Plaut (1997), they assume pre-morbid variation in the computational complexity of non-semantic processes in reading. After combining these pre-morbid factors with post-morbid variation in degree of damage, Dilkina et al.’s models provide a good quantitative fit to the performance patterns of 5 individuals with semantic dementia.

(Lack of) constraints on hypothesized individual differences

A common issue with much of case series research is that hypothesized individual differences—the pre- or post-morbid factors that cause variation in performance—are severely under-constrained. Typically, such work fails to provide independent evidence supporting the pre- and/or post-morbid variation attributed to individuals. As discussed above, research utilizing the two-step interactive theory (Dell et al., 1997; Schwarz et al., 2006) determines post-morbid variation based solely on post-morbid performance. Specific parameters of the two-step interactive simulation are determined for each individual in the case series by an algorithm that
attempts to closely match the observed overall response distribution of that individual. The theory is then assessed via measures of fit to overall response distributions on the very same case series data. For example, the RMSD measure compares each individual's overall response distribution to that of a simulation whose parameters have been set based on the very same overall response distribution. Similarly, Mirman et al. (2011) account for the gradient variation in their data by fitting a statistical model of post-morbid variation to the very same data. In each case, researchers are free to hypothesize any degree of individual variation to provide the best fit to the data they observe; there are no other theoretical or empirical consequences of postulating individual variation.

Theories of pre-morbid variation also typically lack independent evidence of individual variation. Plaut et al.'s (1996) postulation of pre-morbid differences in experience is based solely on post-morbid variation in individual performance patterns. As Plaut et al. explicitly note (p. 99), this account would be considerably strengthened by an independent source of evidence—perhaps by connecting such differences to individual differences in reading skill measures. Although such information could in principle be gathered, it has not been systematically examined in research building on this work (e.g., Woollams et al., 2007). Similarly, Dilkina et al. (2008) propose that pre-morbid variation in the computational capacity of non-semantic reading processes would be reflected by patterns of metabolic activity in brain regions associated with the integration of orthographic and phonological structure. However, such data are not actually available to constrain their case series modeling. Their parameters are based solely on the very same post-morbid
performance patterns from the case series used to assess their theory’s accuracy. This general issue presents a perhaps insurmountable challenge to Plaut’s (1997) proposal for pre-morbid variation in weight decay. How could we independently assess individual variation in the capacity to support large numbers of synapses?

**Empirical constraints on theories of individual differences**

Although in some cases it is unclear how one could obtain empirical support for hypothesized individual variation, other situations are clearly amenable to empirical validation. The most direct method is to directly document individual variation along the proposed pre-morbid variable. This is clearly tractable for some types of pre-morbid variation. For example, as discussed above Jeffries et al. (2011) use documented pre-morbid differences in category expertise to predict patterns of semantic impairment.

Other work has used post-morbid individual differences to make novel predictions about performance—providing independent data on which to test the predictions of individual differences. For example, Dell et al. (1997) and Schwartz et al. (2006; discussed above) use overall response distributions to determine a set of simulation parameters for each specific individual in their case series analysis. These simulation parameters make novel predictions—for example, concerning the probability that syntactic category of the target will be reflected in the syntactic category of individuals’ formal errors (where a target is replaced with a phonologically related but semantically unrelated word; e.g., cop→cot, noun vs. cop→chop, verb). In work like Dell et al. and Schwartz et al., hypothesized individual differences go beyond capturing variance in the range of performance
patterns. Successful *novel* predictions greatly increase our confidence in the validity of these hypotheses.

**Constraining hypotheses concerning individual variation**

As discussed above, in many cases the postulation of individual differences is determined solely by variation in individual performance patterns. Within the space of possible parameter values for a simulation, the hypothesized individual differences are limited only by the data and the techniques for setting parameters. If these individual differences were instead subject to strong constraints, data might be more informative in distinguishing different hypotheses regarding individual differences.

To provide such constraints, case series research might be able to capitalize on insights from statistical modeling techniques that explicitly incorporate individual differences (e.g., linear mixed effects models; Baayen, Davidson, & Bates, 2008). These statistical techniques treat the idiosyncratic contributions of individuals (or items) as samples drawn from a larger population with known properties (e.g., a normal distribution). Parameters of the statistical model reflecting individual differences are therefore constrained not only by the observed data but also by the properties of the population (Baayen, 2008).

To illustrate this, consider a simple psycholinguistic task—lexical decision for frequent nouns. There are known to be individual differences in this task; some speakers will always respond quickly (their true mean reaction times are small), while others will always respond slowly (their true mean reaction times are large). We run a new set of participants using this task. Over 20 trials, we find that
participant S1’s observed mean reaction time is 750 milliseconds. What do we estimate to be S1’s true mean reaction time? That is, if we ran S1 on an additional 20 trials, would we predict that the mean reaction time for these new trials would be equal to, less than, or greater than 750 milliseconds? If we impose no constraints on individual differences, then our best estimate for her true mean reaction time is 750 milliseconds—that is, equal to what we observe in the data. We would therefore expect that given additional trials her mean reaction time would be the same. Given no other information, this is the best guess we can make. This situation is quite similar to that encountered in most current case series research.

However, suppose that based on prior research with this task, it is known that across individuals the mean reaction times of speakers are normally distributed around 1000 milliseconds with a standard deviation of 200 milliseconds. This background knowledge could help constrain our interpretation of observed reaction times. Because S1’s observed mean reaction time falls within the lower tail of the expected distribution of reaction times, we would expect that on a new set of 20 trials she would probably have slightly longer reaction times. In other words, we would estimate her true mean reaction time to be somewhat longer than 750 milliseconds—a bit closer to the population mean of 1000 milliseconds. In this way, background knowledge about the distribution of individual differences can constrain our interpretation of observed data.

Could similar methods be applied to case series data? If individual differences in neurological impairment followed a known distribution, this information could be used to constrain the interpretation of observations in a case
series study. To determine the plausibility of this approach, this section presents a preliminary analysis of the distribution of individual differences from the case series study of Schwartz et al. (2006).

**Empirical sample of individual differences.** Schwartz et al. (2006) used the two-step interactive model to characterize the picture naming performance of 94 individuals with spoken production deficits. As discussed above, in this approach the strength of connections between semantic and lexical units (S connections) as well as the strength of connections from lexical to phonological units (P connections) are adjusted to model the response distributions of impaired individuals. For 16 of these individuals, the best-fitting parameterization of the model was substantially different from the observed distribution of errors. The analysis presented here focuses on the distribution of S and P values for the remaining 78 individuals that the model successfully fit. These parameter values provide a sample of the distribution of individual differences of effects in aphasia; specifically, individual differences in the effect of neurological impairment on lexical vs. phonological retrieval in speech production.

**Analysis of distributional properties.** A histogram of the values of S and P weights fitted to the individuals in this case series is shown in Figure 3. Both distributions are skewed, with a long right tail. The ability of two extensively studied distribution functions (normal and log-normal) to fit these data was examined.

The fit of the normal distribution was explicitly assessed through the D'Agostino-Pearson omnibus test for normality (D'Agostino, Belanger, & D'Agostino,
1990). The null hypothesis that these samples of individual differences were
normally distributed was rejected; the distribution of $S$ and $P$ weights significantly
deviated from normality ($S: \chi^2(2, N = 78) = 31.0, p < .0001; P: \chi^2(2, N = 78) = 51.4, p
< .0001). This suggests that the normal distribution is not an appropriate
colorization of the distribution of individual differences.

Inspection of the histograms suggests a log-normal distribution might better
colorize the distribution of these observations. To examine this possibility, the
fitdistr function from the R library MASS (Venables & Ripley, 2002) was used to
estimate the best fitting normal and log-normal distributions for the $S$ and $P$
weights. Inspection of Figure 3 suggests that the log-normal distributions provided
much better fits to these skewed distributions than the normal distributions. These
fits were quantitatively assessed using the Kolmogorov-Smirnov goodness of fit
statistic. For both the $S$ and $P$ weights, the log-normal distribution provided a better
fit ($S: D = 0.17; P: D = 0.09$) to the observed data than the normal distribution ($S: D =
1; P: D = 0.14$). These findings suggests that the distribution of individual
differences may be better characterized by a log-normal vs. normal distribution.
Figure 3. Empirical distribution of estimated weight values from Schwartz et al. (2006; successfully fit patients only), with best-fitting normal (dashed curve) and log-normal (solid curve) fits. Density indicates the number of individuals at a given weight value. A) $S$ weight values. B) $P$ weight values.
Discussion. Individual differences in lexical and phonological processing impairments (as assessed by the analyses of Schwartz et al., 2006) appear to follow a log-normal distribution. Particularly with respect to $P$ weights, this distribution appears to provide a good characterization of the distribution of individual differences. This could provide background knowledge that could serve as an additional constraint on hypotheses regarding individual differences in case series research. Rather than simply postulating individual differences based only on the specific data observed in the study, this background knowledge could allow for better estimation of true individual differences in neurological impairments.

On a more speculative note, we might ask why log-normal distributions would provide a good fit to individual differences in neurological impairments. This distributional property may be a consequence of the organization of neural processing and/or the response of neural systems to various forms of damage. Log-normal distributions are found in many empirical domains in both the social and physical sciences (Limpert, Stahel, & Abbt, 2001). In the nervous system, the distributions of dendritic spine sizes (a measure of synaptic efficiency) as well as spontaneous firing rates has been argued to follow a log-normal distribution (see Loewenstein, Kuras, & Rumpel, 2011, for discussion). Such distributions are typically argued to reflect the multiplicative interaction of multiple independent effects; in contrast, normal distributions result from the additive interaction of independent effects (Limpert et al., 2001; Loewenstein et al., 2011). In the current context, the log-normal distribution of individual differences may reflect the multiplicative nature of the source of functional impairments. For example, the
functional consequences of neural disruption may be super additive; functional impairments could be greater than the sum of the effects of each damaged neural component (see Sitton, Mozer, & Farah, 2000, for discussion).

Conclusions: Strengthening Case Series Research

Case series research has the potential to advance the development of cognitive theories. To realize this potential, it must aggressively address both theory evaluation and theory selection. With respect to theory selection, it is critical that case series research focus on data that allow us to examine the contrasting predictions of different theories. With respect to theory evaluation, it is critical that theories of individual variation be sufficiently constrained by empirical data. The challenge for conducting cognitive neuropsychological research is to develop the empirical and theoretical tools that allow us to address these critical issues.
Author Note

Thanks to Simon Fischer-Baum and Brenda Rapp for insightful discussion of these issues and helpful comments on the manuscript. This research was supported by National Science Foundation Grant BCS0846147. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the NSF.
References


Appendix: Simulation Methods

As we are examining the fit of the baseline theory—the two-step interactive account—to the simulated case series data, the simulation architecture generally followed its specifications.

Representations and connectivity

Following simulations of the two-step interactive account such as Dell et al. (1997) and subsequent work, the production of a single target word in two different lexical neighborhoods was simulated. In addition to whole-word (lexical) representations, each network had semantic feature nodes (10 per lexical item) and phonemes (3 per lexical item, corresponding to the onset, nucleus and coda of a consonant-vowel-consonant syllable). Following Rapp and Goldrick (2000), a set of conceptual nodes (one per lexical item) was also included. This extended other two-step interactive simulations to incorporate conceptual processes prior to lexical access.

Connectivity also followed previous work. Concept nodes were linked to the semantic features corresponding to the lexical item. These semantic features were then linked to the corresponding lexical unit, which was in turn linked to the corresponding phonemes. In the baseline simulations, all connections were bidirectional and of equal strength. This also held for simulations of the theory violating the baseline account’s damage assumptions. In simulations of the theory violating the baseline’s assumptions regarding the architecture of spoken production—the Restricted Interaction Account (RIA) of Rapp and Goldrick (2000)—there were no feedback connections from lexical units to semantic
features, and the feedback weight from phonemes to lexical units was set to be 25% of the feedforward weight from lexical units to phonemes.

**Neighborhood structure**

Following Dell et al. (1997), each lexical neighborhood consisted of the target plus 5 neighbors. In both neighborhoods, three of these neighbors shared none of the target’s phonemes. One of these was semantically related to the target (sharing 3/10 of its semantic feature nodes). These defined the pure semantic (e.g., for target “cat,” “dog”) and unrelated (e.g., “log”) error categories. The remaining 2 neighbors shared the vowel and coda of the target. In one neighborhood, both of these neighbors were semantically unrelated to the target (e.g., “hat” for target “cat”); these made up the formal error category. In the other neighborhood, one of these was semantically and phonologically related to the target (e.g., “rat”); this defined the mixed error category. Following Dell et al. (1997) the simulation’s responses were determined by averaging across these two neighborhoods, with responses from the neighborhood containing the mixed error weighted at 20% relative to 80% from the neighborhood without a mixed neighbor (Dell et al., 2004; Schwartz et al., 2006).

**Processing**

Following Dell et al. (1997), the activation of each unit was updated as specified in (A1):

\[
A_{\beta}(t + 1) = (1 - q)A_{\beta}(t) + \sum_{\gamma} P_{\beta\gamma} A_{\gamma}(t) + N_{\gamma}(0, SD1) + N_{\gamma}(0, a_{\beta}(t)SD2)
\]
where: \( a_{\beta}(t) \) is the activation of unit \( \beta \) at time \( t \); \( q \) is the decay constant (set to 0.6 in all simulations); \( P_{\beta\gamma} \) is the strength of the connection from unit \( \gamma \) to unit \( \beta \); \( N_t(0, \sigma) \) is a pseudo-random draw at time \( t \) from a normal distribution with mean 0 and standard deviation \( \sigma \); SD1 is the strength of the ‘intrinsic’ noise (set to 0.01 in all simulations); and SD2 is the strength of the ‘activation’ noise (set to 0.16 in all simulations).

Following Rapp and Goldrick (2000), processing consisted of three stages marked by “jolts” of activation to selected units. The strength of these jolts reflected those used by Dell et al. (1997; reflecting the adoption of their neighborhood structure). The first stage, conceptual processing, began with by setting the activation of semantic feature units corresponding to the target concept to 10.0. Activation then spread for 8 time steps, at which time the activation of the semantic features corresponding to the most active concept were set to 10.0 units. This marked the beginning of the second stage, lexical selection. After activation spread for another 8 time steps, the most active lexical unit’s activation was set to 100.0. After activation spread for a final 8 time steps, the most active phonemes in onset, nucleus and coda position were selected, determining the simulations’ output. These were placed into 6 response categories representing the target, the 4 types of neighbors (mixed, semantic, formal, and unrelated) and nonwords.

**Simulating case series data**

Following previous work with the two-step interactive model (Dell et al., 2004; Foygel & Dell, 2000; Schwartz et al., 2006) the effect of lesions was simulated by weakening connections between specific representational levels in the network.
To simulate intact processing, the connection strength throughout the network was set to 0.05 (note: this is weaker than in previous simulations of the two-step architecture due to the addition of conceptual processing and the concomitant increase in activation flowing through the network). Damage to the connections between two levels of processing was simulated by setting the strength of the connections to a pseudo-random draw from a normal distribution with a mean of 0.025 and a standard deviation of 0.01. The connection strength was not allowed to exceed the range (0.0001, 0.05); if the random draw fell outside this range, it was set to the corresponding minimum/maximum value. The damage to connections between different levels of representation was independent. With these damaged connections, 175 naming attempts in each lexical neighborhood were simulated. This equals the number of trials in the Philadelphia Naming Test used to assess patient performance in Dell et al. (1997) and subsequent work.

A total of 1,000 simulated patients were generated for each of three sets of simulations:

- **Baseline: Two-step interactive account.** These simulations followed Dell et al. (1997) and subsequent work in setting making all connections bidirectional and of equal strength. Connections between concept and semantic feature units were left intact; damage was restricted to connections between semantic feature and lexical units and lexical units and phonemes. This corresponds most closely to the assumptions of the simulation used to model case series data by Dell et al. (1997) and subsequent work.
• **Violation of damage assumptions: Conceptual processing impairment.** In these simulations, connections at all levels of processing were damaged. This violates the assumptions of the baseline simulations by allowing errors to arise in processes prior to lexical access.

• **Violation of architectural assumptions: Restricted Interaction Account (RIA).** To implement Rapp and Goldrick’s (2000) RIA proposal, there were no feedback connections from lexical units to semantic features, and the feedback weight from phonemes to lexical units was set to be 25% of the feedforward weight from lexical units to phonemes. This violates the assumptions of the baseline simulations by altering the flow of activation within lexical access processes. Damage followed the baseline simulations.

**Fitting the baseline simulation to simulated case series**

The procedure of Dell et al. (2004) was used to allow a simulations of the two-step interactive account to model the responses proportions from each of the 1,000 simulated patients from each set of simulations. Unlike the baseline simulation above, this simulation did not include conceptual processing; the parameters within the remaining two stages of production processing therefore reflected those utilized by Dell et al. (2004). Using the web-based version of this algorithm (http://langprod.cogsci.uiuc.edu/cgi-bin/webfit.cgi), semantic feature to lexical connections (S weight value) and lexical to phoneme connections (P weight value) in this two-step simulation were set to minimize the $\chi^2$ goodness of fit measure between the two-step model and the simulated response proportions. Following the simulations used to generate the artificial case series, this fitting
procedure was parameterized by setting the number of trials was to 175 and the mixed error opportunities to 20%.