Published in final edited form as:

Cogn Neuropsychol. 2011 October ; 28(7): 515–520. doi:10.1080/02643294.2012.691423.

Who's in and who's out? Inclusion criteria, model evaluation, and the treatment of exceptions in case series

Gary S. Dell and Myrna F. Schwartz

Moss Rehabilitation Research Institute Elkins Park, Pennsylvania, USA

Several years ago, Eleanor Saffran, Nadine Martin, and the two of us began our collaborative project using aphasic speech errors to test production models that had been inspired by the errors of unimpaired speakers. At the time, though, one of us nearly balked upon seeing the data. Compared to everyday speech errors, transcripts of aphasic speech, even in controlled situations such as picture naming or word repetition, revealed an overwhelming amount of between- and even within-patient variation. To the seasoned CN researchers among us, this was no surprise. The researcher's art consists in finding ways to deal with the variability.

The case series method is valuable because it allows the researcher to embrace, and ultimately, to tame the neuropsychological variability in service of testing theory. We are pleased that several distinguished CN researchers, some of whom have considerably more experience than we do with the method, generally agreed with the positive assessment of case series that we presented in Schwartz and Dell (2010). But commenters offered specific recommendations about the method, and pointed to lacunae in our paper.

Is there is a right way to do a case series?

In Schwartz and Dell (2010), we offered examples of case series studies in our work and in other research on topics that we had some familiarity with to show what one *could* do with the method, and made some recommendations about what one *should* do. The commenters augmented both the “could” and the “should” parts. Nickels, Howard, and Best remind us that case series could be used in treatment studies, and Shallice and Buiatti point to studies in which the selection of cases is based on anatomical rather than behavioural criteria. The studies they refer to seem to be clear examples of case series methodology. And they also remind us once more that it is hard to draw the boundaries between classic CN in which the goal is to draw inferences about normal cognition and the many other reasons for studying impaired cognition.

A lot of the commentary concerned how one could (or should) treat individual differences in CN. Lambon-Ralph, Patterson, and Plaut consider the possibility that these differences can be due to pre-morbid factors explicable within the theory (e.g. the triangle theory of reading). They and Bub argue that appealing to this kind of individual variation had been “close to taboo” in the Aristotelian inspired single-subject approach to CN. We can agree with them that case series methods are well suited to the investigation of the contribution to performance of post-morbid individual differences (and pre-morbid differences, when data are available). Goldrick, however, tells us that appeals to individual differences are often too pat. When assumed to arise pre-morbidly, they usually lack pre-morbid supporting data. When assessed post-morbidly and simply attributed to differing degrees of damage,

individual differences are theoretically inert, even though they may appear to have explanatory power. Note that Shallice and Buiatti cite the power of contrasting anatomical case series to rebut appeals to pre-morbid individual differences.

Pre- and post-morbid individual differences can have value, however, if they are then used to generate predictions about other behaviour. Moreover, one can go a step further and attempt to explain the differences themselves. Goldrick shows how constructing a model of the distribution of the individual differences assessed in the case series of Schwartz et al. (2006) can provide insight into their source. Statistically speaking, what Goldrick is doing is treating patients as a random, rather than a fixed, effect. If one has 20 patients, each with some measured value on some test, a fixed-effect model of those values is unparsimonious. It uses 20 “parameters” to model the data, each parameter representing one value. A random effects model hypothesizes an underlying distribution, reducing those 20 parameters to the number of parameters of the distribution (e.g. 2 for a normal or log-normal distribution) thus gaining in elegance and providing insight, particularly when one considers that the shape of the distribution points to potential processing mechanisms, as it did for Goldrick’s log-normal model.

Speaking of models, in our article we emphasized the value for case series analysis of what Olson and Romani call process models, that is, computational models of the cognitive processes that underlie performance. Like us, most people associate CN case series studies with models of cognitive computations such as the DRC model, PDP models of lexical and semantic processing, or our models of aphasic lexical access. Olson and Romani, however, are right when they point out that statistical models are much easier than processing models to work with when making quantitative comparisons among competing models. The complexities of alternative statistical models are precisely reflected in the number of parameters and the space of possible data sets allowed, thus enabling model comparisons in terms of both fit and parsimony. Moreover, as Goldrick shows, it isn’t always easy to compare closely related process models. You need to know *where* to look in your data when there is a specific contrast at issue, such as the degree to which lexical retrieval during production is interactive. Clearly, there is room for skepticism when researchers trot out a case series accompanied by a single model that is reported to “fit” the data. That’s why you need alternative models. You can say that A fits better than B, more easily than you can say absolutely that A fits, just like you can tell that a musical note is higher than another more easily than you can call one high or low. Although model fit comparison is important, it is only part of the process of model evaluation. For example, it is difficult to quantify the degree to which one model is superior to another in the extent to which its assumptions are motivated. Moreover, aside from how well a model fits some data, we can value it simply for its role in making a theory’s ramifications concrete, hence leading to better understanding of potential cognitive mechanisms. In this regard, process models have the edge over statistical models.

Black swans in case series

We proposed that case series and case studies complemented one another particularly when a case series study identifies individuals who deviate from expectations from a statistical or process model. The commenters had a good deal to say about this procedure and the conclusions that can be drawn about outliers. Lambon-Ralph and colleagues ask whether neuropsychological cases can truly function as white and black swans in the classic metaphor for scientific falsifiability—seeing a single black swan falsifies “all swans are white” more effectively than seeing many white swans supports it. Let’s explore this further, by imagining a case series study of swan colour. First, we select birds that are all swan-like in shape. We then gather a number of measures on all the cases including their colour

(varying shades of grey that we can discretize to “white” and “black”) and other factors such as how old the swan is. Through our case series analysis, we discover a relationship between swan colour and swan age (Figure 1). Because of our careful measurements on all cases and our inclusion of many swans, we can identify a rough, but truly curvilinear, relation in which both very young and very old swans tend to be greyer than average, enough so that one might even call them black. Thus, we have a useful generalization that explains some of the black swans seen. Discovering such complex association is impossible without combining the cases. But the generalization is far from perfect. First of all, consider the two swans in the small box in the figure. Here, the younger one is black and the older one white. This dissociation, however, does not excite us. The case series data allow us to conclude that, not only is the dissociation numerically small, but it is fully consistent with the trend in the data.

What does draw our attention are the two black swans, A and B. A formal outlier detection technique such as that used by Fischer-Baum and Rapp (2012) would identify these as unusual, thus meriting closer scrutiny. In a case study of Swan A, it turns out to be covered with soot, and so it's not really black after all. It is, in Nickels and colleagues' words, an “outlier due to a confounding factor.” Swan B, though, is truly a swan and truly blacker than it should be—perhaps one that flew in from Australia. It is a real outlier—a “dissociating pattern that can falsify a theory.” For Swan B, it is clear that we need to amend our theory, as happened to the claim that all swans are white when Australian black swans were first taken note of in the 18th century. But if the only outlier were Swan A, it is clear that we do not. In cognitive neuropsychology, though, it is not so easy to tell whether we have truly black or just sooty swans. That is why we emphasized the need for thorough cognitive assessment of exceptions. Nickels and colleagues claim that that's not good enough; you also have to reassess all the unexceptional swans. To some extent, we agree. It is extraordinarily easy to come up with believable post hoc accounts of why a case is exceptional, and so any such explanation is suspicious until it is bolstered with more data. And such additional data could come from reassessing the white swans. This is just what Fischer-Baum and Rapp did in their study of perseveration. After identifying an individual whose perseveration rate was far greater than expected from the theory that perseverations are the natural consequence of weak retrieval, they did additional tests, and applied some of these tests to the unexceptional patients, as well. This allowed them to conclude that, in their sample, the exceptional case uniquely had a general inhibition deficit. Similarly, when we found individuals in our case series study of aphasic naming (Schwartz et al., 2006) whose pattern of overt errors deviated from the model, we examined other aspects of their behaviour, finding that several of these individuals also made a lot of omission errors. This led us to hypothesize that omission errors, which were not part of the set of modeled error types, may arise on trials in which overt errors are suppressed, thus distorting the overt pattern. We then tested this hypothesis by looking at omissions and other relevant measures in the entire sample. The point is that doing case studies of only the exceptions will not be enough. Whatever hypothesis is developed to explain the exceptions will require additional testing of some sort. This is true whether explaining the exception inspires a significant change in the theory (e.g. Fischer-Baum & Rapp's general inhibitory mechanism) or merely points to another factor that needs to be considered (e.g. the possibility that omission errors obscure overt error patterns). In other words, explaining Swan B and explaining *away* Swan A are both after-the-fact explanations deserving of skepticism in the absence of other data.

Traditionally, truly black swans are supposed to *falsify* theories. CN theories and their instantiations in models, however, are complex collections of assumptions and hence quite different from simple statements about swan colour. Rarely does the whole edifice collapse when contrary data are found, in the manner that is suggested by the term “falsify”. Moreover, instead of a hard and fast distinction between theory-harmless soot covered white

swans and theory-damning Australian black swans, exceptional observations in CN fall on a continuum with regard to their impact on how much the theory is changed in response to the observation. This is because the assumptions that make up CN theories and models vary in how central they are. In practice, a model's developers and advocates routinely change the model incrementally in response to data. The dual-route and triangle models of today, for example, are not what they were 20 years ago. The most central assumptions of the models remain, however, testifying to the difficulty of finding falsifying evidence that is widely regarded as convincing.

Inclusion criteria in case series

Commentators agreed that CN case series should be used to address theoretical questions. Some, like Romani and Olson, went further, suggesting that the inclusion criteria should always be theoretically motivated. In contrast, Shallice and Buiatti allowed that use of pragmatic behaviourally-defined criteria is acceptable in some instances, so long as deviating cases are followed up. We concur, and we applaud their efforts to distinguish among different types of inclusion criteria. As for their enthusiasm for sometimes basing inclusion criteria on anatomy, we wonder whether some of the complications Shallice and Buiatti allude to are more serious than they acknowledge.

Consider two anatomical areas, *x* and *y*, both relevant to the cognitive architecture of functions under study but for different reasons. If one's anatomical criteria fail to differentiate patients with *x* and *y* lesions, then inclusion of patients with lesions that overlap *x* and *y* will distort the conclusions. This looms as a big problem when we consider that *x* and *y* may be close neighbours and thus often, if not always, damaged together by naturally occurring lesions. For example, residing in close proximity within the left posterior superior temporal lobe (Wernicke's area) are neural nodes that appear to make separate contributions to the dorsal and ventral language pathways (e.g., Hickok & Poeppel, 2004; Wise et al., 2001). As part of the dorsal pathway, Wernicke's area contributes to auditory-motor integration in speech and auditory-verbal short-term memory (AVSTM). As part of the ventral pathway, this area contributes to the mapping of speech to semantics. So is it quite likely that a naturally occurring lesion in Wernicke's area will compromise a task like auditory comprehension in two, quite different ways (two functional lesions). Note the violation of the hypothesis that at least initially guides the anatomically-based case series according to Shallice and Buiatti: "when lesions of a particular area of cortex impair performance of a task, they are giving rise to the same functional lesion." While possibly true on a microanatomic scale, the assumption is on shaky ground when it comes to the macroanatomic areas identified by naturally occurring lesions.

Combining data across cases

We described two properties of case series studies: (1) the systematic collection of data from a sample with, ideally, every subject having a value on every dependent measure and (2) the analysis of covariation and other trends across the participants. In some of the research recommended by the commentators (e.g. some of the work cited by Olson and Romani), the first, but not the second, property is true. We can call this kind of research "systematic multiple-case study". There is clearly value in juxtaposing the systematically collected data from the cases without computing cross-participant correlations or other statistics over more than one individual. However, we see little principled difference between the systematic multiple-case study that informally notes trends in the data and a case series that happen to use regression-like data aggregation to identify similarities, differences, and relationships among the cases. Studies that compute statistics across aggregated data have a poor reputation in CN, because of the belief that such aggregation obscures individual variation.

In the approach to case series that we advocate, though, variation is not obscured. It is modeled in one form or another (e.g. regression or a process model), and deviating cases are noted and explored.

A conceptual divide in cognitive neuropsychology?

Bub argues that there remains a fundamental conceptual divide in CN, with case-study advocates assuming a universal functional architecture shared by all intact individuals, and those who do case series focusing on the variability of neuropsychological cases. Moreover, researchers on the two sides of this divide are said to differ on how to account for the observed heterogeneity of cases, whether appealing solely to damage to the universal functional system or also allowing pre-morbid differences in this system to explain post-morbid ones. We believe that most researchers approach the cognitive systems that we study as Bub's "Aristotelians", seeking the universal architecture that underlies diversity. But, like Bub and a number of the commentators, we see a role for lawful pre-morbid individual differences in explaining the data. In our own case series models of aphasic lexical processing, pre- and post-morbid variation are considered to be related. They both are modeled through variation in model parameters (e.g. the p parameter, the strength of connections between lexical and phonological units). Moreover, variation is not just about individuals; it can apply to items as well. For example, words that have been experienced to a greater extent have higher parameter values than those with less experience (Nozari et al., 2010). That is, pre-morbid frequency effects are assumed to be reflected in the same parameters that characterise variations in brain damage. This example reminds us that even those on the Aristotelian side of Bub's divide must and indeed do acknowledge variation in the normal system—within the population of words. In general, by assuming that there exists a common functional architecture and also expecting and seeking lawful (e.g. parametric) variation within it, we bridge Bub's conceptual divide.

Ultimately from our vantage point near the middle, we don't see a true "divide", as least with regard to the value of case series methods. As the commentaries show, CN researchers appreciate the value of case series, but we are still working out all the rules of the road. There is a clear need for motivated selection of cases, the development of quantitative accounts of the observed variability, the consideration of alternative accounts, and the thorough exploration of deviant findings. Moreover, we all agree that case series is just one of many CN research techniques, and that case series studies will flourish best when complemented by the full range of converging methods that the field has to offer.

Acknowledgments

Supported by NIH/NIDCD R01DC000191. We thank the commentators and the editor for their efforts.

References

- Fischer-Baum S, Rapp B. Underlying cause/s of letter perseveration errors. *Neuropsychologia*. 2012; 50:305–318. [PubMed: 22178232]
- Hickok G, Poeppel D. Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. *Cognition*. 2004; 92:67–99. [PubMed: 15037127]
- Nozari N, Kittredge AK, Dell GS, Schwartz MF. Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language*. 2010; 63:541–559. [PubMed: 21076661]
- Schwartz MF, Dell GS. Case series investigations in cognitive neuropsychology. *Cognitive Neuropsychology*. 2010; 27:477–494. [PubMed: 21714756]
- Schwartz MF, Dell GS, Martin N, Gahl S, Sobel P. A case-series test of the interactive two-step model of lexical access: Evidence from picture naming. *Journal of Memory and Language*. 2006; 54:228–264.

Wise RJS, Scott SK, Blank C, Mummery CJ, Murphy K, et al. Separate neural subsystems within 'Wernicke's area'. *Brain*. 2001; 124:83–95. [PubMed: 11133789]

Figure 1.
Hypothetical case series study of swan colour