Cognitive Science (2009) 1–27
Copyright © 2009 Cognitive Science Society, Inc. All rights reserved.
ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/j.1551-6709.2009.01060.x

Global Model Analysis of Cognitive Variability

David L. Gilden

The University of Texas at Austin

Received 12 August 2008; received in revised form 26 March 2009; accepted 27 April 2009

Abstract

Residual fluctuations produced in typical experimental methodologies are examined as correlated noises. The effective range of the correlations was assessed by determining whether the decay over look-back time is better described as a power law or exponential. Both of these decay laws contain free parameters and it is argued that it is not possible to distinguish their models on the basis of simple measures of goodness-of-fit. Global analyses that evaluate models on the basis of how well they generalize are conducted. The models are examined in terms of three constructs that all bear on generalization: cross-validity, flexibility, and representativeness. Quantitative assessment of a large ensemble of data suggests that the correlations decay over time as a power law. The conclusion is that human residual fluctuation is a correlated fractal.

Keywords: Memory; 1/*f* noise; Cognitive variability; Bayesian model selection

1. Introduction

This article concerns perhaps the most basic attribute of human behavior: that it is variable. Some of this variability arises as a consequence of the fact that different states of the world demand different responses. In a designed experiment this form of variability is drawn out by the inclusion of different treatment conditions. There is also a component of human variability that is more endogenous and is not systematically linked to states of the world. In the language of experimental design, this component is the within cell or unexplained variance—unexplained, that is, by the treatment conditions. In most applications of experimental design the first type of variability is intentionally evoked while the second is an unintended consequence that is more or less tolerated as the cost of doing experiments. In this article, we shall be concerned with circumstances where the unexplained residual variability is not a cost but is of theoretical interest in its own right.

Correspondence should be sent to David L. Gilden, Department of Psychology, The University of Texas at Austin, 1 University Station A8000, Austin, TX 78712. E-mail: gilden@psy.utexas.edu

The perspective that leads to a revaluation of variability begins with a slight reconceptualization of what an experiment is and then what data are. In cognitive psychology, the field that we shall be most concerned with, it is almost always the case that a number of responses have to be collected in every treatment cell. Paradigms that involve discrimination or speeded response typically involve scores or hundreds of trials so that differences in cell means can be resolved through statistical averaging. Trials are delivered in large blocks with the different treatments being delivered at random, each cell eventually accumulating enough data to permit the resolution of whatever differences happen to exist. The change in perspective begins with thinking of the trial block not as a collection, but as a process, one that moves the observer through a series of states. Accordingly the data are not to be thought of as piecemeal instances of response awaiting delivery into various cell histograms, but as a time series. The time series is the exact historical record of what happened in the experiment and it is produced by every experiment that is organized around the concept of blocked trials. The dissection of the data time series back into the cells that form the experimental design is typically where data analysis begins and it is required for the most common of statistical models, the analysis of variance (anova). This dissection is rarely questioned but its application does depend upon the assumption that the time series consists of a sequence of independent deviates and that the trial ordering is immaterial. As the treatments are in fact typically delivered in random order and are truly independent, this assumption requires that the residuals be random independent deviates. This is where the time series perspective becomes interesting because this assumption is demonstrably false; the residuals are almost always observed to be sequentially correlated. This is not to say that the residuals have an immediate and transparent structure. Residual time series are to be understood as forming correlated noises and uncovering the structure in correlated noise is not trivial. Developing methods that actually do succeed in describing residual structure is essentially what this article is about.

The time series perspective that recasts human data as correlated noise is not undertaken as a novel but ultimately esoteric mathematical exercise. In the first place it is not novel. This perspective is an integral part of the physical and biological sciences, where an understanding of how systems evolve in time is crucial to understanding the natural laws that govern them. All of the work in chaos theory, for example, derives from this perspective. In this regard, it is noteworthy that the principal hurdle in the application of chaos theory to real data is distinguishing motion on a strange attractor from correlated noise (Sugihara & May, 1990). Secondly, correlated noises come in many varieties and knowing the variety may have tangible implications. Recent work in cardiology is one notable example where it has been demonstrated that the correlated noises formed by heartbeat can be used to distinguish healthy from diseased hearts (Norris, Stein, Cao, & Morris, 2006; Richman & Moorman, 2000). In the present case, knowing the variety will allow us to stipulate the kind of memory system that organizes the chain of cognitive operations leading to judgment and response. Thirdly, all fields of inquiry that examine historical records are implicitly in the business of studying correlated noise. What constitutes a history may be quite general. A musical passage is a history, as is a speech utterance. When viewed as correlated noises these two forms of human production were revealed to imitate nature in ways that were not anticipated within linguistics or music theory (Gardner, 1978; Voss & Clarke, 1975). This is essentially the final point; the description of behavior that focuses only on the states that the system occupies misses all of the information available in the state transitions. The transitions inform on the dynamics, and there is no way to think about dynamics without encountering correlated noise.

Sequential correlation in a time series can be mathematically described in either of two equivalent ways: in terms of the autocorrelation function (the correlation of a sequence with itself displaced by a variable lag) or in terms of its Fourier twin, the power spectrum. The spectral approach is generally preferable in the analysis of noise because complex functional dependencies in the time domain often resolve as very simple features in the spectral domain. Deciding the presence or absence of one feature in particular, the existence of low frequency plateau, motivates the present work on global model analysis. However, prior to this investigation, our interest in residuals was spurred by the most obvious feature of residual spectra, that they are not flat as required by ANOVA. We found instead that spectral power tends to increase with wavelength, and often appears to follow a 1/frequency law, suggesting that residuals are forming what is called in physics a $1/f$ noise. The basic phenomenon has been observed in speeded response paradigms (Beltz & Kello, 2006; Gilden, 1997, 2001; Gilden, Thornton, & Mallon, 1995; Kello, Beltz, Holden, & Van Orden, 2007; Van Orden, Holden, & Turvey, 2003, 2005), in two-alternative-forced choice (2AFC) (Gilden, 2001; Gilden & Gray, 1995), and in production tasks (Gilden, 2001; Gilden et al., 1995; Lemoine, Torre, & Delignieres, 2006). These were interesting results not only because they were unanticipated but also because they created connections to fields outside of psychology. A few examples of $1/f$ noise are fluctuation in heartbeat (Kobayashi & Musha, 1982), ecology (Halley & Inchausti, 2004), pitch and loudness of music and speech (Voss & Clarke, 1975), quasar light emission (Press, 1978), ocean temperature (Fraedrich, Luksch, & Blender, 2004), and this list is not remotely complete. All of these disciplines are now recognized to be relatable at a deep formal level, and this has helped create the modern conception of system complexity.

The noise perspective in human psychology is especially provoking because it is far from obvious why correlated noise would turn out to be so common and so similar across paradigms and tasks. There are circumstances where correlations might be expected to develop in the residual time series, but these mostly have to do with sequential priming. In reaction time methods, for example, it is well known that latencies are influenced by stimulus attributes and motor outputs made on previous trials. However, these effects do not extend over more than a few trials (Maljkovic & Nakayama, 1994) and are easily disentangled from the correlations of concern here (Gilden, 2001). Moreover, correlations suggestive of $1/f$ noise are observed where there is no obvious role for priming. In production methods, for example, there may be a single target stimulus and a single response. Whatever sequential effects are observed in this situation cannot be due to the kind of priming where response or stimulus repetition matters. Odder still are the correlations observed in 2AFC response outcome, where correct trials tend to follow correct trials. Streakiness in signal detection occurs even when every trial is identical, the only variation being whether the target is on the left or right (Gilden, 2001; Gilden & Gray, 1995). Since target position is randomized there are no correlations in the stimulus sequence, and hence it is not possible for the stimulus time series to prime a correlated signal in response outcome.

There are in fact no psychological theories of why human response is correlated as observed. There are at least three reasons why this is so and it is worth mentioning them to motivate the Bayesian modeling perspective that is offered in this article. The first is that the error signals in psychophysics are composed of quantities that are at some remove from the cognitive and perceptual processes that produce them. A reaction time fluctuation, for example, hardly specifies the aspects of attention, memory, and decision making that create that fluctuation. Secondly, even the most sophisticated theories of psychological process do not contemplate the formation of *spontaneous* correlations in the error signal. Theories of reaction time, an example of a behavioral measure that has been intensively studied (Luce, 1986), generally treat only its distributional properties. Dynamic theories of reaction time have been proposed (Vickers & Lee, 1998), but they focus on how latencies are influenced by systematic temporal variation in objective stimulus properties. Similarly the most well-known models of production involve timing behavior and these are built around the notion of independent random deviates (Gibbon, Church, & Meck, 1984; Wing & Kristofferson, 1973). And finally, the development of correlation is intrinsically a problem of some subtlety. $1/f$ noises have, in particular, been a source of considerable theoretical controversy because it is not clear if they arise from general system principles (Bak, 1996) or through a proliferation of individual mechanisms (Milotti, 2002). So even if psychology had foundational theories that were articulated in specific biological terms, it is not guaranteed that the observed correlations in human behavior would be any less problematic.

In this article, we contrast two models of the correlating process in an effort to specify its most basic properties. The models, described in detail below, attempt to distinguish whether the correlation process decays with look-back time as an exponential or as a power law. This same distinction has long been at issue in descriptions of the forgetting function in long-term memory (Navarro, Pitt, & Myung, 2004), and the two fields have much in common. The problem that both fields face is that neither is fundamentally understood and so both employ free parameter models. Were it possible to specify the constant terms in the exponential and power-law formulations, the decision problem would simply boil down to whether the sampling error in the data is sufficiently small to be discriminating. The scaling term in the exponential and the power in the power law, however, are not given by psychological theory, and this makes the decision problem much more difficult.

What it means for a model to fit data is not obvious when the model specifies the functional form without also specifying the numerical values of whatever constants are required to algorithmically compute model values. Free parameters give models flexibility, and goodness-of-fit may simply reflect a model's ability to produce shapes that look like data—data that often contain a substantial amount of measurement error. Consequently it is essential to determine, to whatever extent possible, if a model is a true representation of psychological process or whether it is merely flexible and so able to bend with the measurement error in producing good scores on goodness-of-fit. Regardless of how small the minimum chi-square is for a particular set of parameter values, one will eventually have to reckon with the fact that the model did not predict that specific outcome; it predicted a range of outcomes, one of which may have happened to look like the data. Issues of model selection cannot be settled by optimizing goodness-of-fit on a data set by data set basis. Global analyses that assess model structure beyond the selection of best-fitting parameters are required. The theoretical impotency of assessing models on the basis of good fits has been discussed persuasively by Roberts and Pashler (2000). Their position on the issue is quite clear: ‘‘we did not find any support in the history of psychology for the use of good fits to support theories.’’

In this article, we demonstrate that global analyses can decide the power law versus exponential issue for correlation in response. That this is possible is testimony not only to the power of global analysis but also to the quality of the error signals that are routinely received in cognitive assessment. The corresponding issue in forgetting was found not to be decidable by Navarro et al. (2004) when global model analysis was applied to a large corpus of relevant data.

2. Two models of correlation

The correlations observed in any aspect of behavior will generally decrease with look-back time. That is, the more events that have occurred and the more time that has elapsed between past and present behavior, the less correlated will they be. The decay law, or more formally the autocorrelation function, is the central experimental observation, and one of the core theoretical questions has been whether it is best described as an exponential or a power law (Wagenmakers, Farrell, & Ratcliff, 2004—hereafter WFR; Thornton & Gilden, 2005). The two laws have very different meanings and therefore have different entailments for theories in either domain. Exponential laws have a scale. The scale is needed to make the exponent dimensionless, and in physical settings it expresses some intrinsic property of the system. For example, if a temporal process is at issue, the scale might be a cooling time (Newton’s law of cooling), a decay time, a transition probability per unit time, a diffusion time, or a crossing time. The main point is that the scale provides information about the system in question. Power laws do not have scales and this also has theoretical implications. If the autocorrelation function were a power law, then it could be asserted that the memory process responsible for correlation somehow manages to shed the physical scales of the brain. While scale freedom has hardly been an issue in psychological theory, how systems lose their scales has been at the forefront of modern physics. Scale freedom arises in the theory of phase transitions, where thermodynamic quantities are observed to be governed by power laws. Scale freedom as exemplified by self-similarity is also the defining property of fractal structure. Connections between fractals and power laws arise in a variety of contexts (Schroeder, 1992) with applications that span from economics (Mandelbrot, 1997) to physiology (Bassingthwaighte, Liebovitch, & West, 1994).

Once the shape of the autocorrelation function has been established, the deeper issue of the meaning of the constant terms can be addressed. If the decay law were exponential, then we would be in possession of a decay time scale that might have meaning beyond the experimental design in which it was observed. The numerical value of the scale may reflect some ecological or physiological constraint that sets the memory span of the implicit correlating dynamic. We might view this time scale as an adaptation that reflects an attunement to a regularity in environmental variation, or as the expression of a limiting physiological capacity. Alternatively, if the decay proves to follow a power law, then the mechanisms that produce the observed exponents become an issue. Previous studies (Beltz & Kello, 2006; Gilden, 1997, 2001; Gilden et al., 1995; Kello et al., 2007; Lemoine et al., 2006; Van Orden et al., 2003, 2005) have interpreted correlations in timing and RT data as reflecting power-law decay and have calculated exponents consistent with interpreting the fluctuations as 1/*f* noise. To the extent that this interpretation can be sustained, the derivation of the exponent is key because 1/*f* noise is produced by a handful of specific mechanisms and so the exponent highly constrains the range of theoretical models.

2.1. Short- and long-range models of temporal fluctuation

Exponential decay functions approach zero more rapidly than do power laws. For this reason fractal time series are generally referred to as having long-range correlations, while time series with exponentially decaying correlations are referred to being of short range. The most widely employed short-range models are based upon autoregression (Box, Jenkins, & Reinsel, 1994). The simplest autoregressive model, AR(1), is the leaky integrator:

$$Obs(n) = \phi Obs(n-1) + \varepsilon(n),$$

where *Obs(n)* is the observed value at time (trial) *n* and *ε(n)* is a random perturbation at time *n*. ϕ is the autoregressive parameter specifying the fractional carryover of signal from trial to trial, and in this application is bounded $0 < \phi < 1$. The AR(1) process has an exponentially decaying autocorrelation function that satisfies the recursion relation, $\rho_k = \phi\rho_{k\text{-}1}$, for $k \geq 1$. The lag over which the correlation drops by a factor of *e,* the e-folding scale, is $\tau = -1/\ln(\phi)$.

Leaky integration by itself is not a viable candidate model for describing human residuals. Reaction time latencies, in particular, tend to be characterized by a roughness that cannot be captured by simple autoregression. For this purpose it is necessary to consider a generalization of the AR(1) process that includes an independent component of uncorrelated white noise. The more general model, AR(1) plus white noise, may be written as an autoregressive moving average model, ARMA(1,1) (Pagano, 1974):

$$Obs(n) = \phi Obs(n-1) + \varepsilon(n) + \theta\varepsilon(n-1),$$

where ϕ and θ are the two parameters[1] of the ARMA(1,1) model specifying respectively the autoregressive and averaging parts of the model. In this formulation, the autoregressive piece is roughened up by employing the averaging part of the model in its opposite sense as a differencing operator. Consequently θ is a negative constant in the interval $-1 < \theta < 0$ in applications of ARMA models to psychophysical data. The autocorrelation function of the ARMA(1,1) process satisfies the recursion relation $\rho_k = \phi\rho_{k\text{-}1}$ for lag $k \geq 2$, leading to an exponential decay law with the same e-folding scale as the pure AR(1) process.

Although ARMA processes are defined in the trial domain, their parameters may be better estimated in the Fourier domain where the spectrum has a characteristic inverted-S shape. In the Fourier domain, the lag between trials is replaced by a frequency that has the units of inverse trial number. Low frequencies refer to large trial scales, and power at low frequencies corresponds to broad hills and valleys in the plots of residual time series. Because the correlations drop off exponentially in an ARMA process, the trial scale corresponding to $f_{\mathrm{crit}} = 2\pi/\tau$ marks the broadest hills and valleys that the process produces. As a result, the power spectrum whitens, becomes flat, at frequencies lower than $\mathrm{f}_{\mathrm{crit}}$, and the empirical question in fitting ARMA processes to data is whether the observed spectra also flatten at low frequency. The mathematical expression for the ARMA(1,1) spectrum is given by

$$S(f) = \frac{\left(1 + 2\theta\cos(2\pi f) + \theta^2\right)}{\left(1 - 2\phi\cos(2\pi f) + \phi^2\right)}.$$

Although not obvious, this expression generates an inverted *S*-shape in the parameter range of interest for modeling residual fluctuation, $0 < \phi < 1, -1 < \theta < 0$.

The long-range processes that we consider in this article are not defined in the time domain but rather are defined only through their spectra. Physical mechanisms that produce noises with long-range correlations typically involve complex dynamical systems, and there is no general simple formula for expressing them in the ''*Obs(n* = ….'' sense above. Stationary long-range processes, referred to as fractional Gaussian motions, have a power-law power spectrum, $S(f) = 1/f^{\alpha}$, with a corresponding autocorrelation function, $\rho(k) = 1/k^{1-\alpha}$, where f is frequency in the sense of inverse trial number, and k is the lag in trial number. For $\alpha > 1$, the $1/f^{\alpha}$ spectrum defines a class of nonstationary processes referred to as fractional Brownian motions. The special case $\alpha = 2$ is the well-known random walk. $1/f$ noises form the boundary between stationary and nonstationary self-similar statistical processes. It is remarkable that so many instances of fluctuation are observed to occupy this highly unlikely outpost—it is after all a boundary between two broad classes. The fact that natural systems are often $1/f$ has suggested that the boundary might be a kind of dynamical attractor (De Los Rios & Zhang, 1999), or that it expresses an outcome of the central limit theorem (West & Shlesinger, 1989).

The particular formulation of the long-range process that is treated here was initially created to model timing data. In our original work on $1/f$ noise (Gilden et al., 1995), the Wing-Kristofferson model of timing (Wing, 1980; Wing & Kristofferson, 1973) was modified to accommodate the patent observation that the estimates were correlated and their spectra ascended with decreasing frequency. The Wing-Kristofferson model expresses the notion that timing variability arises from two sources; a central timekeeper (that aspect of cognition that permits a sense of time passing), and the mechanisms of motor output. In their model both sources contribute uncorrelated error from trial to trial. Our modification was merely to replace the white central source with one that emitted $1/f$ noise. The relative contributions of the two sources to the total variation was left as a free parameter. In subsequent work we attempted to apply the same model to the time series formed from RT residuals (Gilden, 1997, 2001). We found that RT residuals could not be modeled using a strict $1/f$

spectrum, but that they could be accurately rendered if the exponent were allowed to be a free parameter, replacing the $1/f$ spectrum by a $1/f^{\alpha}$ spectrum. We shall refer to this model throughout as a whitened fractional Brownian motion (fBmW), although it will turn out that the data are mostly stationary and $\alpha < 1$ in general. The power spectrum of the fBmW model is written constructively in the same spirit as the ARMA model, correlated part plus uncorrelated part, as

$$S(f) = 1/f^{\alpha} + \beta^2,$$

where β is the amplitude of a white noise source. Fluctuation data are generally fit by parameter values in the range $0 < \alpha < 1$, $0 < \beta < 2$ (Gilden, 2001; Thornton & Gilden, 2005). The model should be understood simply as a description of the correlational structure in the data. It in no way stipulates a constructive process in the time domain.

2.2. Previous work on deciding the range of residual fluctuation

The decay law for residual fluctuation was first posed by WFR in terms of null hypothesis testing. Two models were considered. A two-parameter ARMA(1,1) process served as the null hypothesis and was pitted against a three-parameter autoregressive fractionally integrated moving average [ARFIMA(1,*d*,1)] process. The ARFIMA process nests the ARMA process as a special case, such that when its long-range parameter d is greater than zero, it becomes a long-range process with correlations that decay as a power law of look-back time. Three experimental paradigms served as test-beds for testing whether $d > 0$; simple RT, choice RT, and temporal estimation. Model selection was decided for individual time series on the basis of goodness-of-fit, with the larger model being penalized for its extra parameter. WFR concluded that the power-law description of the fluctuation data was supported by their data: ''In all three tasks (i.e., simple RT, choice RT, and temporal estimation), ARFIMA analyses and accompanying model comparison techniques show support for the presence of LRD [long-range dependence, i.e., fractal structure]'' (WFR, p. 581). This finding was in substantial agreement with earlier work (Gilden, 2001; Gilden et al., 1995), where choice RT and estimation data were construed as containing $1/f$ noise. Interestingly, this method also succeeded in establishing long-range correlations in simple RT, a paradigm that had earlier been dismissed as producing essentially white noise (Gilden et al., 1995). Subsequently, Farrell, Wagenmakers, and Ratcliff (2006) reexamined the same data employing a different usage of goodness-of-fit. Farrell et al. used a spectral classifier (Thornton & Gilden, 2005) that pits the fBmW model against the ARMA model in a straight-up goodness-of-fit contest. On the basis of this classifier, Farrell et al. concluded that there were numerous counterexamples to the claim that psychophysical fluctuations had long-range memory.

A mixture of results is not an unanticipated outcome of using goodness-of-fit to referee power law and exponential models of individual time series. Over their parameter ranges there is a great deal of shape overlap between the two models. The central problem is that neither the power law nor exponential model is fundamentally derived from a theory of cognitive fluctuation. In a physical application, say in radioactive decay, the derivation of the decay law would be attended by a derivation of the time scale—in this case through the quantum mechanical calculation of the transition probability per unit time. In psychological theorizing about memory, there is obviously no such calculation, and the time scale is inevitably posed as a free parameter. Similarly, the power in the power law must also be a free parameter. There is no recourse but to fit the models to the data, allowing the parameter values to achieve specific values through optimization. Although the procedure bears superficial similarities to theory testing in the physical sciences, fitting free-parameter models to data in psychology is actually quite different and much more subtle.

3. Global model analysis: Theory

A variety of techniques have been proposed that deal with the problems raised by free parameters in model selection. The most well known of these, cross-validation (Browne, 2000; Mosier, 1951; Stone, 1974), requires models to fix their parameters by fitting a subset of the data, and then to predict data for which it has not been trained. To the extent that a model overfits the training data, the parameter values selected will not generalize to the validation sets—even though the model could possibly fit these sets as well were it allowed to reset the parameters. In this way cross-validation allows variations in the sample statistics to expose models that overfit data. One of the virtues of cross-validation is that it allows models to be tested in the absence of specific knowledge about their parameters. In contrast to the more powerful Bayesian techniques, the prior probabilities of parameter values are not required to effect a concrete application. The cost is that the technique has low power compared to Bayesian methods (Myung, 2000). Cross-validation also has the odd property that it is less informative at large sample size because both training and validation data have the same sample statistics (Busemeyer & Wang, 2000).

A more principled perspective on model selection has been developed (Kass & Raftery, 1995; Myung, 2000; Myung & Pitt, 1997) by focusing squarely on the uncertainty about what is actually being achieved in curve fitting. All signals in psychological data are accompanied by sampling error as well as by individual and idiosyncratic trends. When the parameters of a model are adjusted to optimize goodness-of-fit, both the signal and nonsignal sources contribute variation; the distance (y_{pred}-y_{obs}) is undifferentiated as to its source. Models with free parameters inevitably bend to the noise in achieving best fits, and this presents a delicate problem for model selection. The preference for one model over another ought to be based on which model provides a better representation of the signal component of data. However, the procedure of curve fitting is designed to recommend models that are flexible and so can bend more effectively in response to whatever noise components are present. Goodness-of-fit must be handicapped in some way that punishes flexibility and rewards rigidity. Accordingly goodness-of-fit criteria have been developed, the Akaike information criterion (AIC, Akaike, 1973) for one, that adjust raw goodness-of-fit on the basis of the number of free parameters that the model uses in fitting data. This was the procedure used by WFR to negotiate the additional parameter in the ARFIMA model. However, not all *n*-parameter models are equally flexible and measures of model complexity are required that take into account how models use their specific functional form to overfit data. It is in this context that the shape of the fitting surface, the goodness-of-fit index plotted as a function of free parameters, becomes an issue.

In modern conceptions of flexibility (Hochreiter & Schmidhuber, 1997; Myung, 2000; Myung & Pitt, 1997), the shape of the likelihood surface is used as a key diagnostic. The more flexible a model is, the more peaked will be its likelihood surface. This is because flexible models change their shapes rapidly as their parameters are varied. Even if the best fitting model instance closely resembles the data, the model instances in its neighborhood will not if the model is too flexible. The existence of a peak in the likelihood surface is in itself good evidence that a model is overfitting data—using its flexibility to fit random variation. Inflexible models may not fit as well in the sense of maximum likelihood, but their likelihood surfaces will drop off gradually, presenting a larger region of their parameter space for potential candidates. From a Bayesian perspective, these other candidates should be considered as contributing to the likelihood of the data given the model (Myung & Pitt, 1997). Every potential set of parameters contributes to the *probability(data|model)*, each weighted by the prior probability of that set obtaining in nature. The mean likelihood, more generally referred to as the marginal likelihood, may be relatively small even when the model is able to produce a set of parameters that give an excellent fit to the data. Models that overfit data will tend not to be selected when marginal likelihood referees the competition.

Related to model complexity is the notion of model representativeness (Navarro et al., 2004; see also Wagenmakers, Ratcliff, Gomez, & Iverson, 2004). Both constructs ultimately measure how facile models are in twisting themselves into the shapes of observed data, but representativeness gets at the problem from a different vantage point. In the analysis of representativeness, models are assessed in terms of typicality-of-fit, a construct that has a better defined statistical meaning than goodness-of-fit in the realm of free parameter models. The assessment of typicality begins with the formation of prediction spaces that Navarro et al. refer to as landscapes. Landscapes come in pairs and they are formed as two models compete for their own and for the other's simulated data. In this way landscapes provide a global portrait of how data would be fit under the hypothesis that one or the other model is true.

The specific recipe for measuring representativeness is as follows. First, one model (A) generates an ensemble of simulated data sets across its parameter range. Then both it and a competing model (B) produce their best fit for each ensemble member. Each best fit is associated with a goodness-of-fit measure, say a minimum chi-square, so that every ensemble member can be thought of as being indexed by a pair of minimum chi-square. These pairs serve as coordinates of simulated data sets in the landscape. The simulated data are intended to exemplify what might be observed in nature, and so they express the kind of sampling error that is appropriate to the data domain under consideration. Consequently neither model will fit any of the data perfectly, and the simulated ensemble fills the landscape as an elongated cloud (see e.g., Navarro et al., 2004). The procedure is repeated by allowing the other model (B) to form the ensemble of simulated data sets. To the extent the models are distinctive and make different predictions, the clouds in the two landscapes will have different shapes.

Representativeness in this context is determined by the location of observed data within the respective landscapes. If an observed data set occupies a dense region of a particular cloud, then that data set is representative of the kind of data that is typically produced by the simulating model. To the extent the observed data are informative, they will tend to occupy dense regions in one landscape, and relatively empty regions in the other. A strong case for model selection can be made if there is a large ensemble of experimental data that can be placed into the respective landscapes. The observations themselves will then also form a cloud, and it may be quite obvious whether the observational cloud sits atop one or the other of the simulated clouds. If the observational cloud is not uniquely associated with either model landscape, then it is clear that the data will not decide any questions of model selection. In their analysis of forgetting functions, Navarro et al. (2004) unfortunately found that the bulk of their observed data was not particularly informative in distinguishing power laws from exponentials. In the application here the models are better differentiated and the data sets more numerous. In this context, we will be able to make a stronger case for model preference.

4. Global model analysis: Application

Here we shall determine to what extent global model analysis provides perspective on the nature of psychophysical fluctuation. First, the test-bed will be introduced. It consists of 107 residual time series, each having 1,024 trials. The spectra associated with these time series will be fit by both the ARMA and fBmW models in ways that will permit assessments of flexibility, cross validity, and representativeness. As we shall see, these constructs provide compelling evidence for a fractal interpretation of the residual fluctuations—even though both models generally do a pretty good job at fitting individual data sets. The discussion will be supported by clear graphical evidence with the hope that the power of these analyses will be apparent.

4.1. Data sets

Two ensembles of data will be treated in these analyses, although we wish to emphasize that any set of trials that are generated sequentially in a block may be examined as to the character of their residual fluctuation. The data discussed here are not special in any way. The first ensemble consists of the choice RT and temporal estimation (TE) data collected by WFR, the data that were used by WFR and Farrell et al. (2006) in their treatment of model selection. The third paradigm examined by WFR, simple RT, will not be treated here because the correlations associated with latency to detect the onset of a visual stimulus are quite weak (Gilden et al., 1995 concluded that they were nonexistent) and model selection is not a compelling issue where there is little to be accounted for. The second ensemble consists of choice RT data that we (DG) have collected as part of a general investigation into spectral shape.

There are salient methodological differences in the formation of these two ensembles having to do with the timing of trial delivery. In all experimental conditions WFR imposed a random interval of time between response and delivery of the next trial—the response to stimulus (RSI) interval. A random RSI is required in simple RT as the task is compromised when the onset time is certain. However, there is no such requirement in the presentation of choice RT trials nor in the estimation of temporal intervals. The use of a random RSI in these latter paradigms appears to have been motivated by a desire to make them commensurate with simple RT with respect to temporal uncertainty. In our ensemble of choice RT data the RSI is always constant within a block. Temporal uncertainty could conceivably influence the correlation law and so the two ensembles will be kept separate throughout all phases of the analysis.[2]

The WFR ensemble consists in detail of a 24-time series of length 1,024 trials produced by six participants. In both the choice RT and TE paradigms there were two conditions of RSI; long (mean = 1.35 s) and short (mean = 0.75 s). The DG ensemble consists of an 83-time series of length 1,024 trials produced over the course of eight separate studies; mental rotation (see Gilden, 1997, 2001; Gilden & Hancock, 2007) in two conditions of blocked (and fixed) RSI (500 and 3,000 ms), singleton visual search (set size 2, 4, 8, at most one target present) with the same two RSIs, self-paced absolute identification of two colors with two response keys, self-paced absolute identification of three colors with three response keys, and both serial and parallel forms of multiple-target visual search with self-paced trials (using the aperture bounded gradients described in Thornton & Gilden, 2007). All of our observers were unpaid undergraduates from the University of Texas at Austin

The spectrum of residuals from these 107-time series were estimated using a window averaging version of the fast-Fourier-transform (Press, Teukolsky, Vetterling, & Flannery, 1992) that reduces the variance of spectral estimates well below that which is obtained in the raw transform. In this technique, the time series is divided into partially overlapping windows. The spectrum is then computed separately in each window, and the average over windows at each frequency is obtained. As in any averaging procedure, the variance of each estimate is reduced by a factor proportional to the number of windows that form the average. One downside to this method is that the window sizes are necessarily much smaller than the sequence length and the lowest resolvable frequency is thus shifted from 1/sequence-size to 1/window-size. A second downside to this method is that the spectral estimates that it produces are correlated at neighboring frequencies and this leads to a covariance matrix with off-diagonal elements that must be taken into account in computing chi-square and likelihood. In our construction of the spectrum (Appendix A of Thornton & Gilden, 2005), the power at each frequency is estimated using a different window size, allowing each estimate to have the lowest possible variance. The window sizes were powers of 2 to facilitate the use of the fast Fourier transform, leading to frequencies (inverse trial number) of $1/2^n$, $2 \leq n \leq 9$, where spectral estimates were obtained.

4.2. Numerical definition of models

For the purpose of model comparison both the ARMA and fBmW models were resolved on a uniform 100×100 grid of free parameters. For the ARMA model the grid boundaries were $[-1 < \theta < 0,\ 0 < \phi < 1]$, and for the fBmW model the boundaries were set at $[0 < \alpha < 2, 0 < \beta < 3]$. At each point in the parameter grid 2,000 exemplars were computed as exact spectra at 512 frequencies (the maximum number given observed time series of 1,024 trials), and each was then perturbed using the appropriate exponential distribution at each frequency. These 2,000 spectra were then transformed into the time domain and their spectra were recomputed using the 8-point minimum variance method. The ensemble of 2,000 8-point spectra were reduced to just eight expectation values and an 8 × 8 covariance matrix, all that is needed to compute chi-square goodness-of-fit and likelihood for any set of parameter values (see Thornton & Gilden, 2005, Appendix C).

The covariance matrices of the ARMA and fBmW models have diagonal elements (variances) that are highly nonuniform, a fact that has critical implications for model selection. Spectral estimates at the lowest frequency have standard deviations several hundred times larger than estimates at the highest frequency. This circumstance arises both from the way variability intrinsically appears in the raw spectrum, and by the arithmetic constraints on window averaging. In the raw spectrum, estimates of spectral power are exponentially distributed and the variance grows linearly with increasing power. It happens to be the case that power increases markedly at low frequencies in the application of interest here—residual time series. In addition, there is the numerical constraint that for a fixed amount of data, the number of windows available for averaging is inversely proportional to the window size. In our version of the Press et al. (1992) method of spectral estimation, larger window sizes are used to compute power at lower frequencies. Consequently, statistical averaging offers less variance reduction at lower frequency. Both effects conspire to make the low frequency part of the spectrum the least certain. Unfortunately, deciding whether data are short or long range involves determining whether a spectrum has a knee, and this decision basically comes down to a determination of whether there is a discernible plateau or shelf at the lowest frequencies. This is the key reason why simple curve fitting has not been able to produce a clear consensus on the nature of residual fluctuation.

4.3. Marginal likelihood

The issues surrounding model flexibility and free parameter modeling are well illustrated by contrasting marginal likelihood with maximum likelihood. In this section, the ARMA and fBmW models will be evaluated in terms of both constructs so as to expose their respective tendencies to overfit data. As an illustration of the method, averaged spectra from the WFR collection will first be considered. The averaged spectra provide a clear illustration of both the difficulties posed by fluctuation data and their resolution by the marginal likelihood. Following this introduction, all 107 data sets will be treated individually.

From prior experience with temporal estimation (TE) and choice RT data (Gilden, 1997, 2001; Gilden et al., 1995; Thornton & Gilden, 2005), we expected that the spectra from these two paradigms would be sufficiently different to warrant their being averaged separately. Choice RT data tends to have a higher coefficient of variation (standard deviation/mean) than TE data given the experimental designs that are typically employed, such is the case here, and consequently RT spectra tend to have shallower gradients than TE spectra (Gilden, 2001; provides numerous examples). It was also unknown what effect short or long RSIs would have. Consequently we initially computed average spectra over the six observers in each of the four separate task x RSI conditions. Within each task the RSI variable was found to create only minor differences in the spectrum, of order 10–15%, and so we have averaged over the RSI conditions as well. The averaged choice RT and TE spectra were then evaluated by computing the likelihood values for both models across their respective parameter ranges. The results are shown in Fig. 1, where the first column of panels refers to choice RT data and the second column refers to TE data.

The averaged spectra and best fitting models are shown in panels A and D of Fig. 1. These spectra are consistent with what is generally found in both paradigms (Beltz & Kello, 2006; Gilden, 1997, 2001; Gilden et al., 1995; Kello et al., 2007; Lemoine et al., 2006; Van Orden et al., 2003, 2005; Wagenmakers et al., 2004). The choice RT data are heavily whitened and generate a spectrum (panel A) that is relatively shallow (fBmW model fits at $\alpha = 0.44$, $\beta = 0.82$). The estimation data generate a steep spectrum (panel B), with a slope close to −1 in the log-log plane, the expectation from a *1/f* process that is not substantially whitened (fBmW model fits at $\alpha = 0.95$, $\beta = 0.55$). The best fitting models are superimposed upon the data, and they offer a practical introduction to the difficulties in distinguishing power laws from exponentials.

When we frame model selection in terms of best fit, the contest for the average choice RT spectra is won by the ARMA model (max-like[fBmW]/max-like[ARMA] = 0.53), while the TE spectra is more clearly fractal (max-like[fBmW]/max-like[ARMA] = 2.4). The goodness-of-fit scores that these models achieve arise from the interplay of three factors; the steepness of the spectrum to be fit, the large model variability at low frequencies, and the shapes of the model spectra. First consider model shape. The fractal model because it is long range must use spectral shapes that increase with decreasing frequency. The ARMA model because it is short range must use spectral shapes that turn over and produce the contour of an inverted S. The question of fit then begins with whether the observed spectra turn over at low frequency. This matter is negotiated on data where the models are forced to fit the high frequency part of the spectrum because that is where the model variability is smallest. Fit is primarily determined by the range of shapes available to each model given that the high-frequency part is pinned down. This is where spectral steepness comes into play. A shallow spectrum such as observed in choice RT can accommodate both ascending and *S*-shapes. The *S*-shaped ARMA spectrum that fits this data is sufficiently compressed so that it is difficult to distinguish from an ascending spectrum. The two shapes diverge at the lowest frequencies but here the model variability is so high that the ARMA pays no penalty for a poor fit and the fBmW derives little benefit from a good fit. A steeper spectrum such as typically found in temporal estimation provides better discrimination. The best *S*-shape that the pinned ARMA can find to fit the TE data has regions of curvature in the midrange of frequency where the observed spectrum does not. The model variability in this region is sufficiently small that a substantial fitting penalty is exacted. The observed spectrum smoothly ascends at low frequency, but this is not a problem for the fBmW because all it can produce are smoothly ascending spectra. This is as much as can be learned from maximum likelihood. In order to further resolve the nature of choice RT data it is necessary to take into account that the two models are not equally flexible and that they do not compete for sampling error equally well.

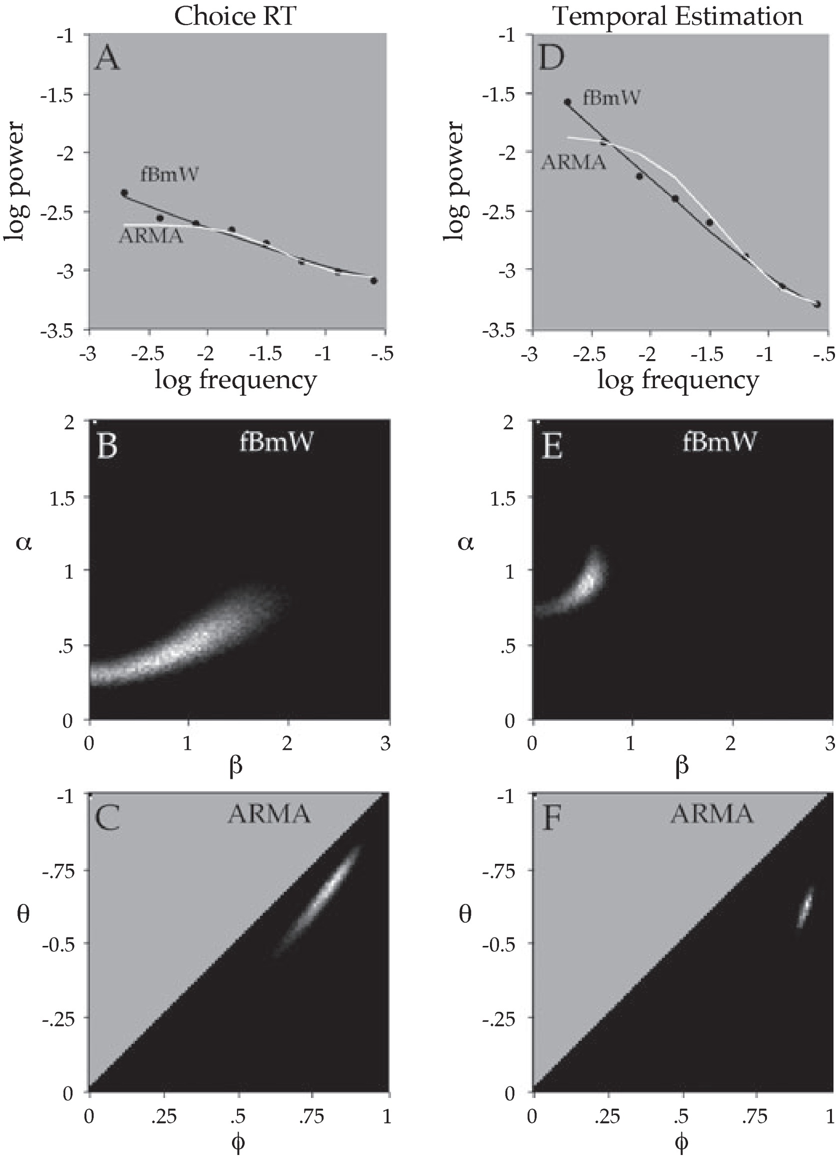

Fig. 1. Models and their fitting spaces are shown in the analysis of flexibility. Panel A shows the averaged choice RT spectrum and the best-fitting fBmW (black line) and ARMA (white line) models. Panel D shows the same for the averaged temporal estimation spectrum. Panels B and C show the likelihood surfaces of the fBmW and ARMA models over their range of parameter variation for the choice RT spectrum. Panels E and F show the same for the temporal estimation spectrum. Gray scale illustrates the likelihood of the model given the data calculated at each pair of parameter values. Lighter grays depict greater likelihood.

A picture of relative model flexibility may be had by contrasting the likelihood surfaces generated by each model when fitting the same set of data. These surfaces are illustrated in panels B, C, E, and F, where likelihood is depicted in grayscale as continuous function of two variables; α and β for the fBmW, θ and ϕ for the ARMA. Only values within a factor of 10 of the maximum likelihood are shown. In panels C and F the upper triangle of the fitting space is grayed out, indicating that this is a region of the ARMA model where spectral power increases with frequency, the opposite of that observed in human data, and so is not relevant to model fitting. In panels B and E the parameter space has been continued into the region $\alpha > 1$ where the fBmW model is nonstationary in order to make it clear that the bulk of the relevant parameter space is stationary, even when fitting the TE data where the best-fitting models have an exponent α close to unity. For both RT and TE data the ARMA likelihood surface is sharply peaked relative to the fBmW surface fitting the same data. To reiterate, a peaked surface occurs when a model shifts its shape rapidly as its parameters are perturbed, achieving a good fit only at highly specific parameter combinations. It is this kind of behavior that a flexible model displays when it overfits data. Although the impressions generated by these surfaces are partly due to the parameter ranges plotted, the surfaces provide a fair visual representation of how rapidly these models lose fit.

The shape of the likelihood surface enters model selection through the Bayesian calculus of probability (Myung, 2000; Myung & Pitt, 1997). Bayesian model selection is based upon how probable the data are given the totality of parameter variation within the model, not on how well the model fits the data at any particular setting of its parameters. Formally, the mean or marginal likelihood is calculated by summing the likelihood at each set of parameter values, with each set weighted by the prior probabilities of those parameter values. The maximum likelihood contributes only a single term to this sum. In order for the sum to be appreciable, the model must offer up a range of parameter values with substantial likelihood. In this way the marginal likelihood penalizes flexible models that have peaked likelihood surfaces.

In practice, the calculation of the marginal likelihood may not be straightforward because while the likelihoods are found from rote calculation, the prior probabilities of the parameter values are unlikely to be known. Realistically, the reason that parameters are free to vary in the first place is that they are not specified by the theory. The calculation of marginal likelihood within experimental psychology will generally require a judgment about how this ignorance should be expressed. The approach taken here will be to constrain the range of prior distributions on the basis of empirical observation and to assume uniform prior distributions within these bounds.[3]

There are three characteristics of fluctuation data that are sufficiently regular that they may be used to place limits on the range of free parameter variation. The first is that fluctuations in both interval estimation and reaction time tend to be positively correlated. Positively correlated signals have a waves-within-waves appearance and this is generally the case in natural systems; trends over large scales have greater amplitude than the point-to-point jitter. Spectra tend to rise at low frequency and there is no reason to consider models that exhibit the reverse. This observation limits the ARMA parameters to $|\phi| > |\theta|$ and fBmW parameter $\alpha > 0$. The first inequality has been incorporated into panels C and F of Fig. 1 by eliminating half of the fitting space. A second observation is that psychophysical residuals are rarely found to be completely white (Beltz & Kello, 2006; Gilden, 1997, 2001; Gilden et al., 1995; Kello et al., 2007; Van Orden et al., 2003, 2005; WFR). The fBmW model is almost completely white beyond $\beta = 2$, and this serves as a practical limit on this parameter. Finally, normal human fluctuation appears to be relatively stationary. Drift in RT would indicate perhaps that the participant is experiencing profound fatigue (RT secularly increasing) or is developing new perceptual sensitivities to the stimuli (RT secularly decreasing). Such trends are a potential problem in any RT paradigm, and reaction time latencies must always be inspected in order to ascertain if they are present. The RT data discussed here do not contain manifest secular trends. Drift in TE data are expected only when the target intervals exceed 1.5 s (Gilden & Marusich, 2009; Madison, 2001), and this is not an issue here. Stationarity in the fBmW model limits $\alpha < 1$, and in the ARMA model it limits $\phi < 1$.

The empirical bounds on α, β, ϕ, and θ, together with the assumption of uniform prior distributions, allows the marginal likelihoods to be calculated. For both the choice RT and TE data the marginal likelihood ratio (fBmW vs. ARMA) was 2.36, positive evidence that the average spectra from both paradigms are more consistent with a long-range process. Although the ARMA model presents a good fit in the sense of maximum likelihood to the choice RT data, its fitting surface is so highly peaked that overall it is a relatively improbable model. This is the signature of a model that overfits data.

The distinction between local maximum likelihood and global marginal likelihood was reiterated in the analysis of individual spectra. Likelihood surfaces for each of the 107-time series in the combined data sets were constructed for both models. Table 1 tallies for the WFR collection the number of times that one or the other model would be selected on the basis of having the best overall fit (maximum likelihood) or having the larger marginal likelihood. The counts tell an interesting story. On the basis of best-fit the ARMA model has an edge with 15 winners of 24 possible. However, on the basis of marginal likelihood the ARMA model had eight defectors while the fBmW model had none, so that the fBmW model claims 2/3 of the data. The same pattern is seen in Table 2, which summarizes the

Table 1
Maximum likelihood and marginal likelihood as selection criteria (Wagenmakers et al. [2004] data)

	Marginal Likelihood	
Maximum Likelihood	ARMA	fBmW
ARMA	7	8
fBmW	0	9

Table 2
Maximum likelihood and marginal likelihood as selection criteria (Gilden RT data)

	Marginal Likelihood	
Maximum Likelihood	ARMA	fBmW
ARMA	13	30
fBmW	0	40

counts for our RT collection. Here the ARMA process has a slight edge with 43 winners out of 83 possible. However, here there were 30 defections from the ARMA model when marginal likelihood is the selection criterion. These defections left the ARMA process with only 13 winners and the fBmW model claims roughly 7/8 of the data. In both data sets we find positive evidence to prefer the fBmW model, but equally important is the observation that there were no instances where the fBmW model was reversed as the focus shifted from maximum to marginal likelihood. The large number of reversals suffered by the ARMA model indicates that it fits data by being flexible and so by being able to produce shapes that resemble whatever sampling error happens to be present.

4.4. Cross-validity

In the cross-validation technique, there is some freedom in specifying how the data are to be split into training and validation sets. In the application here the models are rendered in terms of their power spectra, and the models are mostly differentiated by their spectral shapes at the lowest frequencies. This circumstance suggests that the individual data sets be kept intact at the block size in which they were collected, and that cross-validation proceed across participants. Although it is unusual to cross-validate between different individuals in this way, it makes sense if the lowest frequencies in the data are to be available for model specification. It must be recognized that splitting at the level of the individual participant implicitly assumes that the residual formation process is universal, that each participant provides a sample of general stochastic process. This assumption has been explicit in most studies of residual fluctuation where the goal has been to compute and interpret the model parameters as if they were meaningful psychological quantities, not just an outcome of individual idiosyncratic behavior. In this way the data were divided into three sets: 12 sequences of TE data and 12 sequences of choice RT data from WFR, and 83 sequences of choice RT data from our collection. The two different RT collections were again separated in order to assess the impact of different RSI procedures used in this paradigm. From the WFR data there are $(11 \times 12)/2 = 66$ predictions generated from both the RT and TE data (12 observers in each condition). From our collection the 83 observers lead to $(82 \times 83)/2 = 3{,}403$ predictions.

Each point in Fig. 2 corresponds to two values of chi-square obtained from fixing parameters for one observer and cross-validating on another observer in the same ensemble. One coordinate comes from the ARMA model, the other from the fBmW model. Panel A shows how the models competed for the RT and TE data from the WFR collection. Panel B shows how the models competed for our RT data. Points in the upper triangles depict instances where the fBmW model cross-validated with a smaller chi-square than the ARMA model. It is evident that most of the predictions from both collections fall into the upper triangle. The proportion of times fixed fBmW models cross-validated better than fixed ARMA models were 0.71 in WFR choice RT data, 0.76 in TE data, and 0.75 in our collection of choice RT data. This is further evidence that ARMA model achieves good fits by bending with the sampling error and consequently the parameters it selects are not stable across the range of data.

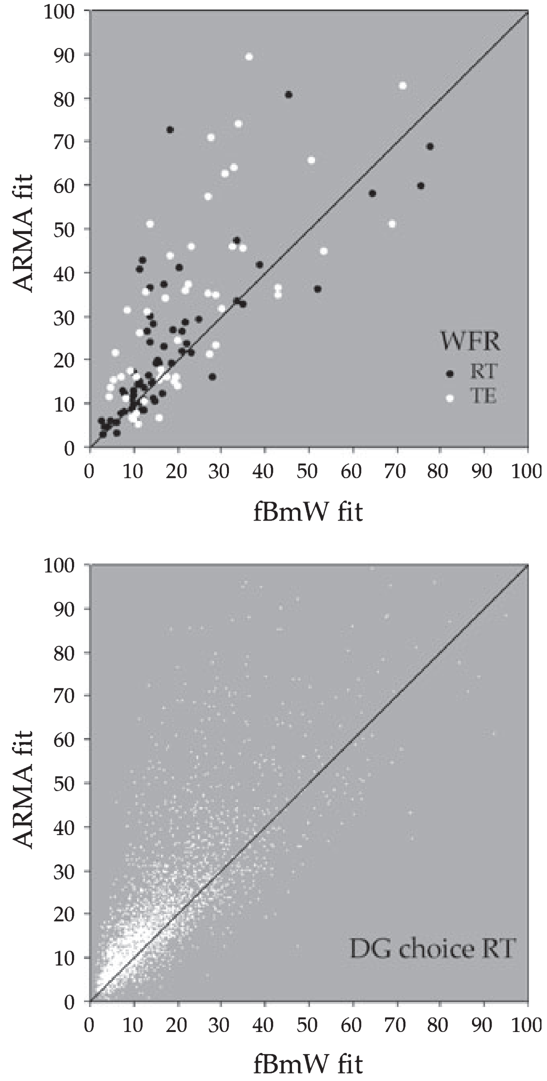

Fig. 2. The results of cross-validation between observers are shown for two ensembles of data. The top panel illustrates WFR cross-validation chi-square values plotted in the plane formed by the competing ARMA and fBmW models. As speeded choice (RT) and temporal estimation (TE) are two entirely different tasks with demonstrably different spectra, they have been cross-validated separately. The bottom panel illustrates cross-validation chi-square values for DG choice RT data. In both panels, points above the diagonal depict data sequences where the fBmW model produced a better cross-validating fit than did the ARMA.

4.5. Representativeness

The landscapes used in this technique create a model competition that is based not on which model produces the best fit, but rather upon which model can better simulate the kinds of fits that are actually observed. To the extent that a particular model is a true description of human residual fluctuation, then its simulated data sets should produce a landscape that densely covers the observed data. As in the calculation of marginal likelihood, the formation of the landscapes is rote up to the specification of the model parameter priors. The priors are needed because they influence the density of simulates in the landscape. Again, the priors distributions were treated as being uniform, with boundaries set by constraints on stationarity and overall spectral shape.

Landscapes were formed by first fixing the simulating model as either ARMA or fBmW. Data were simulated by choosing parameter values at random within the specified boundaries, and then computing spectra for these parameters that reflect sampling error and finite sequence length. Each spectrum formed in this way was interpreted as an example of the kind of data that would be observed were the simulating model true. The pairs of minimum chi-square values that result when simulated spectra are fit by the fBmW and ARMA models form one point in the landscape. The entire landscape is formed by repeated sampling from the distribution of parameter priors. Ten thousand iterates of this procedure were obtained from both the ARMA and fBmW models providing the simulated data. Fig. 3 illustrates the respective simulated landscapes as 10,000 yellow dots. Each dot marks the minimum chi-square for both models fitting one instance from the simulating model. Panel A depicts a landscape populated by data that would exist were the fBmW model true, and panel B depicts the landscape were the ARMA true. As chi-square is used here to index goodness-of-fit, points far from the origin imply objectively bad fits. This technique requires that both good and bad fitting instances be richly represented so that the models can generate alignment. Alignment is evident in the yellow dots in both landscapes, vertical in panel A, horizontal in panel B. Alignment is a generic feature of landscapes produced by potentially distinguishable models and simply reflects the fact that a given model tends to fits its own simulates better than competitor models.

Quite a bit can be learned about the relative complexity of competing models by noting how well each fits the simulates produced by its competitor. In panel A, where the simulates are produced by the fBmW model, the ARMA does on average not much worse than the fBmW; mean χ^2 (ARMA) = 6.7, mean χ^2 (fBmW) = 5.0, where the average is taken over the 10,000 fBmW simulates. This is striking in so far as in reality the data did come from the fBmW model, and it shows that the ARMA is capable of producing spectral shapes that look fractal. Such a state of affairs could merely signify that the two models have a strong family resemblance in their spectra—or it could signify that the ARMA is capable of producing a variety of spectral shapes, and can make itself look fractal where required. The second possibility is borne out in panel B, where the ARMA produces the simulates. In this case, the full range of the ARMA model is expressed and it produces a profusion of shapes, the majority of which would not be confused with the output of an fBmW. Evidence for this is that the fBmW model does quite poorly in fitting ARMA simulates, mean χ^2 = 11.4, while the ARMA fits its own simulates with roughly the same precision that the fBmW fit its own, with a mean χ^2 = 5.2. The asymmetry in how well each model fits its competitors simulates is simply more evidence that the ARMA is a more complex model than the fBmW, and that it is able to productively use its flexibility in goodness-of-fit contests regardless of whether it is a true description of the psychophysics.

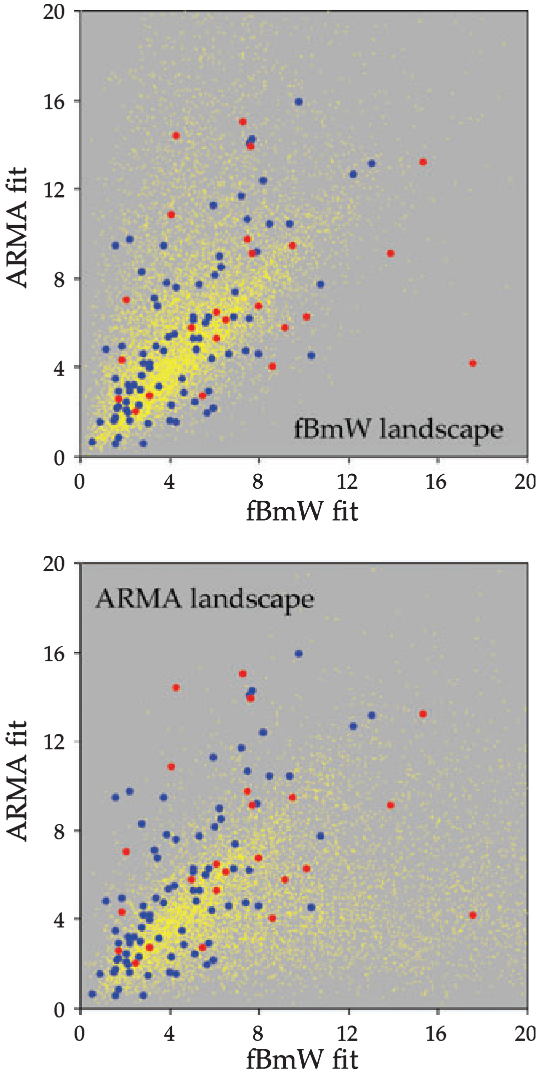

Fig. 3. Landscapes are shown for the fBmW and ARMA models along with the full ensemble of data. The axes denote values of chi-square, and each point represents the pair of best-fitting fBmW and ARMA models. In the top panel, the yellow dots illustrate 10,000 simulates of the fBmW model. In the bottom panel, the yellow dots illustrate 10,000 simulates of the ARMA model. Red dots denote choice RT and temporal estimation data from Wagenmakers et al. (2004). Blue dots denote data from our collection of choice RT data. Representativeness may be inferred from the yellow cloud density at the positions of the data.

Model selection proceeds in this technique by plotting the data in each of the two landscapes and deciding which set of yellow dots follows the data better. Fig. 3 depicts the two cases; the blue dots show the Gilden choice RT data (mental rotation, search, etc.), and the red dots show the temporal estimation and choice RT data from WFR.[4] There is no question that the data clouds appear to be more vertically than horizontally aligned, evidence that the

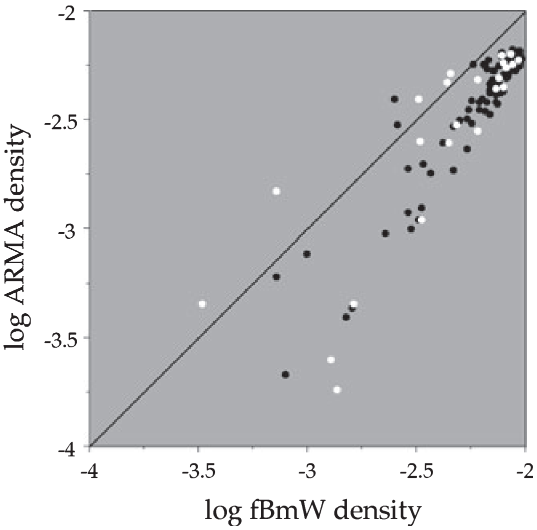

Fig. 4. Comparison between ARMA and fBmW models in terms of representativeness probability (local density in simulated landscape). White dots denote WFR data and block dots denote DG choice RT data. Points below the diagonal depict individual data sets that were more representative of the fBmW process.

data resemble the simulates of the fBmW model. This impression is placed onto a quantitative scale by explicitly computing the representativeness of the observed data. In a given landscape, the local density of the yellow dot cloud is an exact measure of how likely the model is to produce simulated data at that particular point. In this sense, the local density measures the representativeness of data. As described by Navarro et al. (2004), density in the landscape can be estimated by treating the simulated data (yellow dots) as bivariate Gaussians and summing their collective contributions at each point of data in the landscape. The width of the bivariate Gaussian is not specified by the technique and it is necessary to check that the conclusions are not sensitive to its choice.

Fig. 4 illustrates the representativeness of the data as indexed by the density of simulated data sets in the two landscapes. Data that are more representative of the fBmW model lie in the lower triangle. Ninety-eight of the 107 data sets lie in the lower triangle in Fig. 4. This illustration was computed using a width for the bivariate normal of three chi-square units. Over the range of width between 1 and 10 this value varied from 90 to 106 and so no conclusions are materially challenged by not being able to provide a theoretically motivated value of the width. Although much of the data are bunched together near the diagonal, it is nevertheless evident that the quantitative calculations support what is evident to the eye in Fig. 3; the data follow the fBmW simulates.

5. Perspective

In this article, we have used global model analysis to develop three independent lines of evidence that residual fluctuation is fractal, that it has an autocorrelation function with power law decay. This is an important result for several reasons. It implies that the dynamic that produces residual time series cannot be described by a single process operating over a single timescale. Rather it suggests that psychophysical decision making involves a *coordinated* series of processes that operate over a range of timescales. What does this mean for psychology? If nothing else, it means that the complexity of human thought and response can and should be framed within the physical conception of complexity. That conception is in a state of rapid maturation encompassing game theory, animal behavior, market behavior, evolution, and adaptive systems generally. Research in complex systems will offers new metaphors for understanding what happens when a person makes a decision, as well as new analytic techniques for framing behaviors that rely upon the coordination of interacting subsystems.

However, the more interesting result, at least from the point of view of modeling, is that we can make the argument at all. Without the perspective of global model analysis, the nature of residual fluctuation would be mired in a goodness-of-fit contest. This perspective has important consequences for theory building in cognitive psychology generally, and it is well worth summarizing. We will close with three of its counterintuitive observations on the enterprise of fitting models to data.

1. There should be no premium placed on a good fit per se. A good fit could result from gross opportunism on the part of the model. This is a difficult lesson to learn as a good fit has an enormous rhetorical force. Good fits are compelling. However, if the likelihood surface is peaked, it means that the model is probably overfitting the data—regardless of how beautiful the fit is at the top of the peak. A flexible model may be able to follow the noise from data set to data set. If so, it does not matter how many ''best fit'' contests such a model succeeds in winning if the marginal likelihood, the integral over the range of allowable parameter variation, is small compared to other models. The ARMA is, in this sense, an opportunistic model. Its fitting peaks are always observed to be quite narrow within its range of parameter variation. As a case in point, the ARMA process wins many of the goodness-of-fit contests in both the WFR and DG data sets, but it does so with a sharp peak and so loses the majority of contests based on marginal likelihood.
2. Good fits are not always desirable. In most cognitive and psychophysical applications, it is inevitable that some data sets will be fit better than others. A good model must be able to produce this kind of variation in simulation; it cannot produce uniformly good fits because that is not the observed world. A model is a good description of nature when its simulations look like both the good-fitting data as well as the bad-fitting data. It is in this light that the fractal process most clearly makes a compelling case for the data. The 107 data sets plotted in Fig. 3 line up fairly nicely with the fBmW simulates—especially in comparison with the ARMA simulates.
3. Goodness-of-fit alone cannot serve up counterexamples that falsify theories. Farrell et al. (2006) claim to have found counterexamples to the general statement that human psychophysical time series are fractal. This claim makes sense in the context of goodness-of-fit contests because it means that the ARMA model won some of the contests.

We now see that the production of counterexamples may not be so straightforward. Counterexamples must be evaluated in terms of the representativeness and complexity of the models that produce them. A good case for a counterexample would be an entire class of data that is not representative of the fractal process or more specifically, a 1/*f* noise process. Gilden and Hancock (2007) took one step toward this when we showed that a class of adults with attentional disorders produced choice RT residuals that were definitely not examples of 1/*f* noise.

Acknowledgments

Preparation of this article was supported by NIMH grants R01-MH58606 and R01-MH065272, and NSF grant BCS-0226449. We would like to thank Tom Thornton for his help with the time-series analysis.

Notes

1. The variance of the perturbations enters as an additional parameter in the specification of the ARMA(1,1) process, and it would be required, say, for forecasting. However, the variance does not influence the shape of the autocorrelation function, and without loss of generality we treat the case of normalized residuals.
2. The use of a random RSI clearly did have a disrupting influence on time estimation. Typical coefficients of variation for continuous tapping tasks are 5% (Allan, 1979; Fetterman & Killeen, 1990; Grondin, 1993; Wearden, 1991; Wing, 1980; Wing & Kristofferson, 1973). WFR observers generated coefficients of variation some five times larger.
3. A common criticism of the assumption of uniform priors is that the assumption does not hold under general reparameterizations of the model (Gelman, Carlin, Stern, & Rubin, 2004). This criticism is certainly true here and would be more telling were the ARMA and fBmW models uninterpreted and subject to recasting in arbitrary coordinates. These models are, however, not uninterpreted and each parameter plays a specific role that derives from a history of physical and statistical theory: α is a spectral exponent and makes sense only in relation to a spectrum, ϕ and θ are the autoregressive and averaging coefficients in a general ARMA(1,1) model, and these make sense only in relation to the notions of averaging and autoregression. There is little motivation to consider arbitrary transformations of any of these variables. Of the four parameters considered here, only β might be transformed so that we would consider the priors to be defined on β^2 or even $\beta^2/(1 + \beta^2)$, which gives the proportion of variance in pure white noise in the fBmW model.
4. We would have included temporal estimation data from Gilden et al. (1995) as well in the landscapes, but there is a slight difference in procedure that makes this impossible.

We used a continuous responding paradigm, where each response initiates the timer for the next interval. This creates an upturn in the spectrum at high frequencies for technical reasons—the motor delays end up being differenced, and the first derivative of white noise is a spectrum that *increases* at high frequency. WFR used a separate ‘‘go’’ signal for each estimate and consequently each motor delay only adds white variation. The models being tested are not designed to accommodate positive spectral slope.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox & F. Caski (Eds.), *Second international symposium on information theory* (p. 267). Budapest: Akademiai Kiado.

Allan, L. G. (1979). The perception of time. *Perception & Psychophysics*, *26*, 340–354.

Bak, P. (1996). *How nature works*. New York: Springer-Verlag.

Bassingthwaighte, J. B., Liebovitch, L., & West, B. J. (1994). *Fractal physiology*. Oxford, England: Oxford University Press.

Beltz, B. B., & Kello, C. T. (2006). On the intrinsic fluctuations of human behavior. In M. Vanchevsky (Ed.), *Focus on cognitive psychology research* (pp. 25–41). Hauppauge, NY: Nova Science Publishers.

Box, G. E. P., Jenkins, G. M., & Reinsel, G. C. (1994). *Time Series analysis: Forecasting and control*. Hoboken, NJ: John Wiley & sons.

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*, 108–132.

Busemeyer, J. R., & Wang, Y.-M. (2000). Model comparisons and model selection based on generalization criterion methodology. *Journal of Mathematical Psychology*, *44*, 171–189.

De Los Rios, P., & Zhang, Y.-C. (1999). Universal 1/f noise from dissipative self-organized criticality models. *Physical Review Letters*, *82*, 472–475.

Farrell, S., Wagenmakers, E.-J., & Ratcliff, R. (2006). 1/*f* noise in human cognition: Is it ubiquitous, and what does it mean? *Psychonomic Bulletin & Review*, *13*, 737–741.

Fetterman, J. G., & Killeen, P. R. (1990). A componential analysis of pacemaker-counter timing systems. *Journal of Experimental Psychology: Human Perception and Performance*, *16*, 766–780.

Fraedrich, K., Luksch, U., & Blender, R. (2004). 1/*f* model for long-time memory of the ocean surface temperature. *Physical Review E*, *70*, 037301.

Gardner, M. (1978). White and brown music, fractal curves and one-over-f fluctuations. *Scientific American*, *238*(April), 16–31.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis*, 2nd ed. (pp. 61–66). Boca Raton, FL: Chapman and Hall.

Gibbon, J., Church, R. M., & Meck, W. (1984). Scalar timing in memory. In J. Gibbon & L. Allan (Eds.), *Annals of the New York Academy of Sciences, 423: Timing and time perception* (pp. 52–77). New York: New York Academy of Sciences.

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science*, *8*, 296–301.

Gilden, D. L. (2001). Cognitive emissions of 1/*f* noise. *Psychological Review*, *108*, 33–56.

Gilden, D. L., & Gray, S. A. (1995). On the nature of streaks in signal detection. *Cognitive Psychology*, *28*, 1–16.

Gilden, D. L., & Hancock, H. (2007). Response variability in attention deficit disorders. *Psychological Science*, *18*, 796–802.

Gilden, D. L, & Marusich, L. R. (2009). Contraction of time in attention-deficit hyperactivity disorder. *Neuropsychology*, *23*, 265–269.

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

Grondin, S. (1993). Duration discrimination of empty and filled intervals marked by auditory and visual signals. *Perception & Psychophysics*, *54*, 383–394.

Halley, J. M., & Inchausti, P. (2004). The increasing importance of 1/f-noises in models of ecological variability. *Fluctuation and Noise Letters*, *4*, R1–R26.

Hochreiter, S., & Schmidhuber, J. (1997). Flat minima. *Neural Computation*, *9*, 1–42.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General*, *136*, 551–568.

Kobayashi, M., & Musha, T. (1982). 1/f fluctuation of heartbeat period. *IEEE Transactions of Biomedical Engineering*, *29*, 456–457.

Lemoine, L., Torre, K., & Delignieres, D. (2006). Testing for the presence of 1/f noise in continuation tapping data. *Canadian Journal of Experimental Psychology*, *60*, 247–257.

Luce, D. (1986). *Response times: Their role in inferring elementary mental organization* (Oxford psychology series). New York: Oxford University Press.

Madison, G. (2001). Variability in isochronous tapping: Higher order dependencies as a function of intertap interval. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 411–422.

Maljkovic, V., & Nakayama, K. (1994). Priming of pop-out: I. Role of features. *Memory and Cognition*, *22*, 657–672.

Mandelbrot, B. B. (1997). *Fractals and scaling in finance*. New York: Springer.

Milotti, E. (2002). 1/f noise: A pedagogical review. *Arxiv: Physics,* 0204033v1.

Mosier, C. I. (1951). Problems and designs of cross-validation. *Educational and Psychological Measurement*, *11*, 5–11.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, *44*, 190–204.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, *4*, 79–95.

Navarro, D. J., Pitt, M. A., & Myung, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, *49*, 47–84.

Norris, P. R., Stein, P. K., Cao, H., & Morris, J. J. Jr (2006). Heart rate multiscale entropy reflects reduced complexity and mortality in 285 patients with trauma. *Journal of Critical Care*, *21*, 343.

Pagano, M. (1974). Estimation of models of autoregressive signal plus white noise. *Annals of Statistics*, *2*, 99–108.

Press, W. H. (1978). Flicker noises in astronomy and elsewhere. *Comments in Astrophysics*, *7*, 103–119.

Press, W. H., Teukolsky, S. A., Vetterling, W. T., & Flannery, B. P. (1992). *Numerical recipes*. New York: Cambridge University Press.

Richman, J. S., & Moorman, J. R. (2000). Physiological time-series analysis using approximate entropy and sample entropy. *American Journal of Physiology - Heart and Circulatory Physiology*, *278*, 2039–2049.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*, 358–367.

Schroeder, M. (1992). *Fractals, chaos, power laws: Minutes from an infinite paradise*. New York: W. H. Freeman.

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

Sugihara, G., & May, M. M. (1990). Nonlinear forecasting as a way of distinguishing chaos from measurement error in time series. *Nature*, *344*, 734–741.

Thornton, T. L., & Gilden, D. L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin and Review*, *12*, 409–441.

Thornton, T. L., & Gilden, D. L. (2007). Parallel and serial processes in visual search. *Psychological Review*, *114*, 71–103.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General*, *132*(3), 331–350.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and $1/f$ scaling. *Journal of Experimental Psychology: General*, *134*(1), 117–123.

Vickers, D., & Lee, M. D. (1998). Dynamic models of simple judgments: I. Properties of a self-regulating accumulator model. *Nonlinear Dynamics, Psychology, and Life Sciences*, *2*, 169–194.

Voss, R. F., & Clarke, J. (1975). ''1/f noise'' in music and speech. *Nature*, *258*, 317–318.

Wagenmakers, E.-J., Farrell, S., & Ratcliff, R. (2004). Estimation and interpretation of $1/f^{\mathrm{a}}$ noise in human cognition. *Psychonomic Bulletin & Review*, *11*(4), 579–615.

Wagenmakers, E.-J., Ratcliff, R., Gomez, P., & Iverson, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, *48*, 28–50.

Wearden, J. H. (1991). Do humans possess an internal clock with scalar-timing? *Learning and Motivation*, *22*, 59–83.

West, B. J., & Shlesinger, M. F. (1989). On the ubiquity of 1/f noise. *International Journal of Modern Physics B*, *3*(6), 795–819.

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

Wing, A. M., & Kristofferson, A. B. (1973). Response delays and the timing of discrete motor responses. *Perception & Psychophysics*, *14*, 5–12.