Published in final edited form as:

Behav Processes. 2009 June ; 81(2): 205–215. doi:10.1016/j.beproc.2009.01.013.

The Generality of Empirical and Theoretical Explanations of Behavior

Paulo Guilhardi and Russell M. Church

Brown University

Abstract

For theoretical explanations of data, parameter values estimated from a single dependent measure from one procedure are used to predict alternative dependent measures from many procedures. Theoretical explanations were compared to empirical explanations of data in which known functions and principles were used to fit only selected dependent measures. The comparison focused on the ability of theoretical and empirical explanations to generalize across samples of the data, across dependent measures of behavior, and across different procedures. Rat and human data from fixed-interval and peak procedures, in which principles (e.g., scalar timing) are well known, were described and fit by a theory with independent modules for perception, memory, and decision. The theoretical approach consisted of fitting closed-form equations of the theory to response rate gradients calculated from the data, simulating responses using parameter values previously estimated, and comparing theoretical predictions with dependent measures not used to estimate parameters. Although the empirical and theoretical explanations provided similar fits to the response rate gradients that generalized across samples and had the same number of parameters, only the theoretical explanation generalized across procedures and dependent measures.

Keywords

Cross-validation; modular theory; timing theories; theory evaluation

Empirical explanations provide predictions of observed behavior without intervening variables; theoretical explanations also provide predictions of observed behavior but contain intervening variables that may represent psychological or neural processes. The goal was to compare empirical and theoretical explanations of behavior using an example of each. This paper describes an approach to determine the degree to which theories of behavior actually predict behavior.

Over the past few years we have been developing what we consider to be a theoretical explanation of behavior (Guilhardi, Yi, & Church, 2007). More recently, however, we found it necessary to better understand the specific characteristics of theoretical explanations that justify their relatively complex structures that contain modules that represent different intervening processes, and that are well-specified by equations. The following statement was

Contact address: Paulo Guilhardi, Department of Psychology, Brown University, Box 1853, Providence, RI 02912, Phone: (401) 863-3979, Fax: (401) 863-1300.

This article was based on talks presented on April 11, 2008 in Bloomington, IN at the Meeting of the Society of Experimental Psychologists (SEP) and on May 23, 2008 in Chicago, IL at the Conference of the Society for Quantitative Analysis of Behavior (SQAB).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

taken from an editor's letter rejecting a paper that was submitted to a well respected journal: "Neither reviewer is greatly impressed by your model. Its assumptions seem arbitrary, or not obviously grounded in timing theory. Is it any more than an exercise in curve fitting?"

At the time, we believed we were working with a reasonably well developed explanation of behavior that was certainly grounded in timing theory. Most importantly, we believed that our theory explained the data beyond just curve-fitting, although the editor and reviewers did not agree. After the initial distress at the rejection, we looked at the reviewer's criticisms more objectively. The reviewers had a point. We were describing the dynamics of temporal discrimination using well-known functions (Guilhardi & Church, 2005), and trying to incorporate these functions into an existing theory of asymptotic temporal discrimination (Kirkpatrick, 2002; Kirkpatrick & Church, 2003) to extend its generality to the dynamics of acquisition and extinction. Although we were attempting to generalize our predictions to many dependent measures, we were not convincing the reviewers that we had accomplished more than just curve fitting. And, even if we had accomplished this, why should we use a complex theory to describe data that could also be described by simple well-known functions?

Consider, for example, the response rate gradients as a function of time shown in Figure 1 from rats trained on a 60- and 120-s peak procedure (Yi, 2007) that was similar to the procedure used by others (Catania, 1970; Roberts, 1981). In the Yi (2007) procedure, the first lever-press response after a fixed interval from the onset of the stimulus delivered food. The rats were trained with two stimuli (e.g., noise and light), each associated with a different fixed interval (60 and 120 s). In three-fourths of the stimulus presentations, a lever-press response after the fixed interval resulted on the delivery of food, stimulus termination, and a 20-s period with no stimulus. On one-fourth of the stimulus presentations, no food was delivered and the stimulus remained on for 260 s. The data in all panels of Figure 1 are the response rate gradients averaged over trials in which no food was delivered.

The right panels of Figure 1 show the fits of three theoretical explanations to the response rate gradients. The three explanations are scalar expectancy theory (SET, Gibbon, 1977, 1991; Gibbon & Church, 1990; Gibbon, Church, & Meck, 1984), the learning-to-time theory (LeT, Machado, 1997), and a modular theory (Guilhardi & Church, 2005; Guilhardi et al., 2007; Kirkpatrick, 2002; Kirkpatrick & Church, 2003). Although the theories were slightly different in the number of free parameters (7, 4, and 6 free for SET, LeT, and modular theory, respectively), they all accounted for most of the variance in the 520 data points ($\omega^2 = 0.994$, 0.989, and 0.997 for SET, LeT, and modular theory, respectively). The left panel of Figure 1 shows the fits of an empirical explanation to the data. The empirical explanation was defined as the well-known Gaussian function (with two parameters, mean and standard deviation) and an additional two parameters (a vertical scaling and a horizontal shift parameter). It had the same number of parameters as LeT which had the fewest parameters of the theories, and it accounted for approximately the same amount of the variance ($\omega^2 = 0.997$) as the modular theory which accounted for the highest proportion of the variance ($\omega^2 = 0.998$). The question is whether there is any reason to use a theoretical explanation with intervening variables if an alternative empirical explanation with no intervening variables can also predict the data. The remainder of this article will focus on the question: Are theoretical explanations better than empirical explanations, and if so, why?

To address this question, secondary data analysis from 12 rats trained under a fixed-interval procedure (Guilhardi & Church, 2004a; Guilhardi & Church, 2004b) were used to compare empirical and theoretical explanations. Training on the fixed-interval procedure occurred as follows: a cycle began with the delivery of food; 20 s later, a stimulus (such as noise) began; after a fixed interval from the onset of a stimulus, a head-entry into the food cup terminated the stimulus, and delivered a food pellet. Three groups of four rats were trained for 10 sessions

(each with 60 stimuli presentations) at three different intervals (30, 60, or 120 s). The data shown in Figure 2 (and used for analysis) are from the last five sessions of training.

Description of Empirical and Theoretical Explanations

An Empirical Explanation

The empirical explanation was an ogive function described in Equation 1:

$$f(t) = (d - a) * \frac{1}{\left(1 + e^{-\frac{(t-c)}{b}}\right)} + a$$

Equation 1

The empirical explanation, defined by Equation 1, describes response rate as a function of the time since the onset of the stimulus, $f(t)$. The parameters are the start (a), the slope (b), the center (c), and the end (d). Note that $(d - a)$ is the range of the ogive function; parameters d and a shift the function vertically.

The fits of this empirical explanation (Equation 1) of the response rate gradients of the individual rats trained on a fixed interval 30 s, 60 s, and 120 s are shown in Figure 2 (the left, center, and right columns, respectively). The open circles are data and the smooth lines are the best (least-squares criterion) fit of the ogive function to the data.

The parameter estimates from the empirical explanation fit of the individual rat data are shown in columns 2 to 4 of Table 1. (The values are the means and, in parenthesis, the standard errors of the mean.) An analysis of variance with one between-subjects factor (FI duration) was performed to identify differences across fixed-intervals for each of the parameters. There was a statistically significant effect of fixed-interval condition in the estimate of the end (d), center (c), and slope (b), but not on the start (a), as shown in the last two columns of Table 1. Note that the values of the center and slope parameters increase approximately proportionally to the fixed interval duration. If the relative center and slope parameters are used (c/I and b/I , where $I = FI$ duration), there is no statistically significant effect of fixed-interval condition (relative center, $F_{(2,9)} = 0.11$, $p = 0.899$ and relative slope, $F_{(2,9)} = 1.23$, $p = 0.338$.)

A Theoretical Explanation

The theoretical explanation was the modular theory of learning and performance (Guilhardi & Church, 2005; Guilhardi, Keen, MacInnis, & Church, 2005; Guilhardi et al., 2007; Kirkpatrick, 2002).

The Modular Theory—A flow diagram of the modular theory is shown in Figure 3 (adapted from Guilhardi, in press; Guilhardi et al., 2007). The diagram contains the basic elements of the theory and the inputs and outputs of each element. The *time markers* refer to the stimulus. The onset of the stimulus starts a *clock* and the information provided by the clock is available to the entire system at all times. The delivery of a reinforcer strengthens, and the non-delivery weakens, the association between the time marker and the response in *strength memory*. Strength memory is primarily responsible for the observed overall response rate. The delivery of a reinforcer also affects *pattern memory* which is the representation of the expected time from the onset of the time marker to the next reinforcer based on experience. Pattern memory is primarily responsible for the observed distribution of responses in time, such as the observed shape of the fixed-interval response rate gradients. A decision to respond is based on pattern memory, the current value of strength memory, and an operant response rate. Once the decision to respond is made, a packet of responses is initiated.

A more detailed flow diagram, Figure 4, shows the functional forms of all the elements from the modular theory with procedure as the input, perception, memory, and decision as intervening modules, and packet of responses as the output.

Perception: Subjective time is linearly related to physical time, t (the “Clock” in the perception module shown in Figure 4). During cycles in which reinforcement is delivered, perception is a function that starts at zero (time of stimulus onset) and increases linearly with physical time until it reaches the subjective time of reinforcement $\hat{T}$. The subjective time of reinforcement $\hat{T}$ is defined as the physical time of reinforcement T multiplied by K (a normally distributed variable with mean μ_k and coefficient of variation γ_k). During cycles in which no reinforcement is delivered, perception is a function that starts at zero (time of stimulus onset) and increases linearly with physical time until the time of the stimulus termination.

Pattern memory: After the delivery of reinforcement (on cycles with reinforcement), these perception functions are transformed into expected time to reinforcement functions. The expected time to reinforcement is a function that starts at KT and decreases in time to zero (the “Perceptual Store” in the perception module in Figure 4). After cycles with no reinforcement, the expected time to reinforcement is a function that starts at the mean subjective time of reinforcement, decreases in time to zero, and then increases until the subjective time of stimulus termination. Pattern memory is the weighted average of expected time to reinforcement functions (the “Reference” in the memory module in Figure 4).

Strength Memory: Strength memory changes in time with experience, simultaneously with pattern memory. Response strength weakens during extinction and strengthens with reinforcement (the “Strength Memory” in the memory module in Figure 4). The strengthening and weakening of response strength is defined by a linear operator rule (Bush & Mosteller, 1951, 1955), similar to the rule used in the Rescorla & Wagner (Rescorla & Wagner, 1972) and learning to time theory (Machado, 1997). See Guilhardi, et al. (2007) for a detailed description of strength memory in the modular theory of learning and performance.

Decision: The decision to respond is based on a thresholded pattern memory (the “State” in the memory module in Figure 4), the current strength (the “Strength Memory” in the memory module in Figure 4), and a constant operant rate of responding (the “Operant Rate” in the decision module in Figure 4). The threshold (the “Threshold” in the memory module in Figure 4) is set at a percentile of pattern memory, b (normally distributed with a mean μ_b and coefficient of variation γ_b).

Applications of the Modular Theory for the Fixed-interval and Peak Procedures

—The equations for the perception, memory, and decision modules of the modular theory, as well as the explicit solution for the response rate as a function of time gradients are described in detail in Guilhardi, et al. (2007) for the fixed-interval, and Yi (2007) and Guilhardi (in press) for the peak procedure. The parameters consisted of a rate parameter (A), the mean (μ_b) and standard deviation (γ_b) of the pattern memory threshold (B), the mean (μ_k) and coefficient of variation (γ_k) of the pattern memory K , and an operant response rate parameter (R_0). Because the parameters μ_k was set to one and γ_k set to zero during the fixed-interval procedure, they are reported only on applications of the theory to the peak procedure.

The modular theory predicts on single cycles, a low response rate of R_0 , followed by an abrupt transition to a high rate of responding (parameter A). The abrupt transition occurs at a time determined by the mean and standard deviation of the threshold B . In the present application, the modular theory was fit to each of the rat’s response rate gradients using the explicit solution. This produced an estimate of the five parameters for each rat. The parameters were then used to simulate data. During the simulation, the computer generated stimuli and reinforcers that

were used as inputs to the perception, memory, and decision modules of the theoretical explanation. The theoretical explanation then generated responses as output. The goal of the simulation was to obtain theory-generated data to compare with the data obtained from the rats. The simulated data consisted of times of events (such as stimulus onset, responses, and reinforcers) corresponding to each rat on each cycle of training. These simulated data were used to determine any dependent measure that could also be determined from the rat's data.

The fits of the theoretical explanation to the response rate gradients of the individual rats trained on a fixed interval 30 s, 60 s, and 120 s are shown in Figure 5 (the left, center, and right columns, respectively). The empty symbols are data and the smooth lines the best (least-squares criterion) fit of the modular theory to the data.

The mean parameter estimates from the theoretical explanation fit to the data from individual rats are shown in Table 2. The values in parenthesis indicate the standard error of the mean. Analysis of variance with one between-subjects factor (FI duration) was performed to identify differences across fixed-intervals for each of the parameters. The only statistically significant effect of fixed-interval condition was on the estimate of the rate parameter (A).

An Evaluation of the Empirical and Theoretical Explanations of Temporal Discrimination

The empirical and theoretical explanations used as examples were similar in their adequacy (how well they fit the data) and in some determinants of complexity (the number of free parameters and the number of parameter estimates). These and other criteria for model selection are described by Myung (2000). The focus of the evaluation in this article will be on the ability of the empirical and theoretical explanations to generalize to other samples of the behavior, other dependent measures, and other procedures.

Adequacy

The means of the goodness-of-fit for the empirical and theoretical explanations based on the response rate gradients are shown in Tables 1 and 2 (empirical and theoretical explanations, respectively). The percentages of variance accounted for (ω^2) were 0.99 or above for both explanations at all fixed-interval conditions. Thus, both explanations adequately accounted for the data.

Complexity

The empirical and theoretical explanations were similar in some determinants of complexity -- both had four free parameters. With four free parameters, three intervals, and 12 rats, there were a total of 144 parameter estimations for the 2520 data points. (Note that only 15 data points per interval per rat are shown in Figure 5.)

The number of parameter estimates could be reduced without a large decrease in goodness-of-fit. For each of the 12 rats, the empirical explanation could be based on a single parameter estimate of relative center, relative slope, and start; although three estimates would be needed for the "end" because it differed across fixed-interval conditions. Similarly, for each of the 12 rats, the theoretical explanation could be based on a single parameter estimate of the mean threshold (μ_b), the coefficient of variation of the threshold (γ_b), and the operant level (R_0); although three estimates would be needed for the overall response rate because it differed across fixed-interval conditions. This could have reduced the number of parameter estimations from 48 to 24 without much loss of variance accounted for. With the exception of the empirical end parameter (d) and the theoretical rate parameter (A), there was reasonable consistency between parameters across rats (as shown by the low standard error of the mean). If individual

differences across rats were ignored (i.e., the same parameter estimated is used for all rats), the number of parameters estimated would be further reduced to six.

The complexity of a theory refers to all factors affecting its flexibility. The theoretical explanation was potentially more flexible than the empirical explanation because it had more equations and several function forms, but this potential flexibility appears not to have been used because the coefficient of variations of the parameter estimations (standard deviation divided by the mean) of the theoretical explanation were similar to those of the empirical explanation. Because the theoretical and empirical explanations were similar in their number of parameters, there was no need to use a measure of theory evaluation that integrates both complexity and adequacy such as the Akaike Information Criterion (Akaike, 1973; Bozdogan, 2000) and Bayesian methods (Schwarz, 1978; Wasserman, 2000). These methods do not incorporate the flexibility produced by function forms into their evaluation algorithm. However, a model selection criterion called “minimum description length” might be applied (Pitt, Myung & Zhang, 2002).

In principle, because the parameters of a theoretical explanations represent relatively invariant underlying mechanisms of the organism’s behavior that it explains, these explanations are more likely to produce parameter estimations that are similar across conditions. This is also the case when the theoretical explanations incorporate principles that integrate data across conditions, such as, the scalar timing principles (Gibbon, 1977, 1991; Gibbon & Church, 1990; Gibbon et al., 1984).

Generality

To emphasize the importance of three types of generality, we maintained the empirical and theoretical explanations with similar complexity (i.e., number of parameters) and adequacy (i.e., goodness-of-fit). Next we describe how these explanations differ in their generality across samples, dependent measures, and procedures.

Generality across samples—The cross-validation method (Browne, 2000; Collyer, 1986; Stone, 1974; Zucchini, 2000) was used to determine the degree to which the empirical and theoretical explanations can generalize across samples. In the cross-validation method, parameters are estimated using some of the data (referred to as calibration sample), and then the estimates are validated on another sample of the data (referred to as validation sample) not used on the parameter calibration. The difference between the goodness-of-fit based on the validation sample ($\omega^2_{\text{validation}}$) and calibration sample ($\omega^2_{\text{calibration}}$) is a measure of the generality of the theory across samples and indicates problems of overfitting due to increased flexibility by the theory.

To test the generality across samples, the data were divided into a calibration sample (defined as “odd cycles”) and validation sample (defined as “even cycles”). This makes it possible to estimate parameters from one sample, but to evaluate them on a different sample. Parameters from the empirical and theoretical explanations were estimated using the calibration sample and then tested for generality using the validation sample. Table 3 shows the mean proportion of variance accounted for by the calibration and validation samples, with the standard errors in parentheses. The fact that the parameters estimated from the calibration sample accounted for only slightly more of the variance than those estimated from the validation sample suggests that there was no serious problem of fitting random error. The difference ($\omega^2_{\text{calibration}} - \omega^2_{\text{validation}}$), provides a measure of this overfitting. The magnitude of the overfitting was only about 1%, and the difference in the magnitude of overfitting with the empirical vs. theoretical explanations was not statistically significant, $t_{(11)} = -1.14$, $p = 0.28$. Thus, in this case, the amount of overfitting of the empirical and theoretical explanations was similar and relatively small.

Although the ability of an explanation to generalize across samples is necessary, it is certainly not sufficient (Busemeyer & Wang, 2000; Mosier, 1951). The predictive value of an explanation can be increased if the explanation can also generalize across dependent measures and procedures. The generality of the empirical and theoretical explanations across dependent measures and procedure are evaluated below.

Generality Across Dependent Measures (Output Generality)—Eleven dependent measures were used to evaluate the generality of the explanations across dependent measures. Most of the dependent measures were selected from Guilhardi and Church (2004b, 2005) and were based on a search for measures used in publications from the *Journal of the Experimental Analysis of Behavior* that used fixed-interval schedules of reinforcement (See Table 1 from Guilhardi & Church, 2004b). These included overall response rate, response gradients described by the four-parameter ogive function (Equation 1), curvature index, and time of transition from a low to a high response rate, among others. The importance of this method of selection of measures is that it provides a somewhat unbiased selection of measures that were used by other researchers rather than measures that could be chosen to favor a particular explanation. An unbiased selection of dependent measures would strengthen conclusions about a theory's generality.

Two dependent measures were defined directly from the fit of the ogive function to the data. These measures, the center and slope of an ogive, were the only dependent measures of the list that could be predicted by the empirical explanation. Because the empirical explanation of these two dependent measures was directly estimated from the data during the fitting, they do not provide any information about the generality across dependent measures of the empirical explanation. The empirical approach could not generalize across dependent measures unless further processes were assumed and introduced and, therefore, was not further compared to the theoretical explanation.

From the processes and equations of theoretical explanation, however, it was possible to simulate data. The simulation of the theoretical explanation emulated the procedure and amount of training (number of sessions and cycles) used. The parameters used for the simulation were those estimated from the response rate gradients (Figure 5) described in Table 2. Of course, explicit solutions could also be developed and used, but they might require additional assumptions to predict many different dependent measures.

The measures used to evaluate generality across dependent measures of the theoretical explanation are shown in Table 4. The table also shows the means for each of the 11 dependent measures for the three fixed-interval conditions of both the data and the theoretical explanation (labeled "Theory"). Standard errors of the mean for each of the dependent measures and fixed-interval conditions are enclosed in parenthesis. Statistical analyses were used to determine the effects of fixed-interval condition and of the theory accuracy (effects of comparison between data and theory) for each of the dependent measures. The F values, as well as p values, are also shown in Table 4. Note that there was an effect of fixed-interval condition for all of the dependent variables, except for the curvature index which has a scalar property. The last column of Table 4 (Theory Accuracy) shows the lack of difference between the data from the rats and the predictions of the theory.

The theoretical explanation predicted the results of 11 different dependent measures. There were no significant differences between the 11 dependent measures predicted from the theory and obtained from the rats. In fact, the distribution of p values were scattered approximately uniformly between 0 and 1, as would be expected if data from the rats and simulation of data from the theoretical explanation came from the same source. Although many of the dependent measures are correlated with each other, it is especially impressive that the theoretical

explanation predicted dependent measures that bear no specific relationship to a particular free parameter. One interesting example is the post-reinforcement pause (the seventh dependent measure). The post-reinforcement pause is a function of four of the modules and processes of the modular theory (operant rate, pattern memory, strength memory, and response bouts, and therefore, of several parameters). Nonetheless, the post-reinforcement pause of the rats and the predictions of the theory were similar.

Validity of the theory: Although the theory provided adequate fits of the 11 dependent measures of performance, this is not sufficient to demonstrate the validity of the theory. A sufficiently complex theory could fit 11 specific dependent measures of performance, but fail to generalize to others. A valid theory is one in which the theoretical processes for perception, memory, and decision correspond to actual neural and/or cognitive processes. One way to obtain evidence about the cognitive validity of a theory is to introduce an error into a single module, and compare the predictions of the original theory with the erroneous theory on all of the dependent variables. If the original theory were valid, the error in a single module should affect the dependent variables that are assumed to be affected by this module, but not affect the dependent variables that are assumed not to be affected by this module.

To illustrate this approach, the pattern memory module was modified to produce an ogive-like state on individual cycles, rather than a step function. This was an error that would have a profound effect on performance on individual cycles, but would be difficult to detect on performance averaged across many cycles. The same parameters estimates based upon the response rate gradients (Figure 5) were used to simulate the theoretical explanation with error. Because pattern memory determines the pattern of the observed response rate gradients on individual cycles, the theory with error would produce an ogive-like pattern in contrast to a step function. With the assumption that the centers of the two functions occur at the same time, the difference between them on individual cycles would be diagnostic. The theoretical explanation with error would over-predict the low response rate prior to the center, and under-predict the response rate after the center. Therefore, the prediction was that this error would be readily detected by two of the dependent variables--the response rate prior to t_I (Low) and rate after t_I (High).

Table 5 (with the same structure of Table 4) compares the data to the theoretical explanation with a single error. Note that, as predicted, the theoretical explanation with error over-predicted the low response rate prior to the time of transition to the high state, and under-predicted the response rate at high state (italicized dependent measures in Table 5).

A more general way of comparing the two theoretical explanations (without and with error) explanations is with the use of an adaptation (Church & Guilhardi, 2005) of the original Turing test (Turing, 1950). In order to perform the Turing test, first the data from individual cycles from the rats and two simulations of the theoretical explanation (without error) under the fixed-interval procedure were summarized as local rate ($\log_{10}$ responses per minute) as a function of time since the stimulus onset (s). An example is shown in the left panel of Figure 6. The question was whether the gradient from one of the simulations of the theoretical explanation at cycle n ("Sample") was more similar to another gradient taken from cycle n of second simulation of the theoretical explanation ("Theory"), or from the rat ("Rat"). Similarity was defined based on a comparison between the sum of absolute deviations between the sample and theory response gradients ($|Sample-Theory|$) and the sample and rat response rate gradients ($|Sample-Rat|$). That is, the smaller the sum of the absolute deviation, the more similar. This judgment produced for each cycle a binary decision of whether the theoretical explanation was more similar to itself or more similar to the rat. A proportion of correct detection was calculated based on the comparison of cycles 1 to 300 (last 5 sessions with 60 cycles each) for each rat,

and is shown in the top-right panel of Figure 6. The distribution was centered at approximately 0.5 indicating the theoretical explanation was indistinguishable from the rat.

The same Turing test was performed using an additional two simulations of the theoretical explanation with the error used in Table 5. The results for individual rats are shown in the bottom-right panel of Figure 6. The introduction of error in the theoretical explanation shifted the distribution to a mean greater than 0.5, indicating the inaccuracy of the explanation. The advantages of the Turing test are that it can evaluate the accuracy of the theory without the need to compare it to alternative hypotheses (such as the mean, or an alternative theory) and that the test is very general. The disadvantage of the Turing test is that it does not provide information about the characteristics of the data that the theoretical explanation is failing to predict. Such detailed evaluation was provided earlier by the analysis of individual dependent measures.

Generality Across Procedures (Input Generality)—The generality across procedures, as well as the generality across dependent measures, must also be assessed as a way of comparing explanations of behavior. As Busemeyer and Wang (2000) wrote, “After all, accurate *a priori* predictions to new conditions are the hallmark of a good scientific theory.” (p. 171)

The generality across procedures was determined by extending the theoretical explanation from the fixed-interval to a peak procedure (Yi, 2007). The peak procedure is depicted in Figure 7. In Yi’s procedure, after an initial training on a fixed-interval procedure (top, Figure 7), rats were introduced to a procedure in which on 25% of the trials food was not delivered and the stimulus remained on for 260 s. Rats were trained on two intervals (60 and 120 s) that were signaled by two different stimuli (light and noise).

The data, previously shown on the right panels of Figure 1, were once again fit by the theoretical explanation (modular theory), but this time the three of the parameters were estimated from different rats, on a different procedure (fixed-interval). The values of three of the parameters (R_0 , μ_b , and γ_b) were estimated from the mean across rats and across fixed-interval conditions. The parameter A , and the two parameters that had been fixed at 1 and 0 (μ_k and γ_k) were allowed to vary. The application of the theoretical explanation (the modular theory) for the peak procedure is precisely described in Yi (2007) and Guilhardi (in press) and not reproduced here since the goal is to provide an analysis of the theoretical explanation generality across procedures. The data from the peak procedure (circles) and the new fits of the theoretical explanation (smooth line) are shown in Figure 8 (data redrawn with permission from Yi, 2007). The variance accounted for by the theoretical explanation (ω^2) was 0.98 or above, but the small systematic deviations of the observed and predicted results indicate that the theory is not perfect.

The generality across procedure was also tested in a peak procedure for human participants. The procedure is depicted in Figure 9. On regular fixed-interval trials, participants were presented with a target that after a “click” sound that moved at a constant velocity across the computer screen. Participants were asked to shoot at the target to score points. The location of all shots was at a fixed position (in the center) of the screen. Because of the constant velocity of the target, and the fixed position of the shots, the precision of the shots could have been determined by either, or both, time since the click sound and position of the target on the screen. On some trials, however, the target and its trajectory were hidden by a white mask. On those peak trials, participants were also asked to shoot at the target and were able to score points but, on those trials, position of the target on the screen was not a cue that could be used. On these trials, accuracy was determined by time since the click sound. Participants were trained on

targets that reached the center of the screen either at 2, 2.83, or 4 s. These durations were signaled by a different background color (green, black, and blue).

The response rate gradients (circles) and fits of the theoretical explanation (smooth lines) are shown in the top panels of Figure 10. The same parameter restrictions used to fit the peak-interval procedure data from rats were used to fit these data. The values of three of the parameters (R_0 , μ_b , and γ_b) were estimated from the mean across rats and across fixed-interval conditions. The parameter A , and the two parameters that had been fixed at 1 and 0 (μ_k and γ_k) were allowed to vary. The goodness of fit (ω^2) ranged from 0.77 to 0.88 and did not fit the data satisfactorily. We reduced the restriction on parameter flexibility and allowed the threshold mean and coefficient of variation parameters (μ_b and γ_b , respectively) of the theoretical explanation to vary. The new fits are shown in the bottom panel of Figure 10. The goodness of fit (ω^2) was 0.99 or above. An interesting fact is to compare the mean (μ_b) and standard deviation ($\sigma_b = \gamma_b \times \mu_b$) of the threshold from the rats and humans. The mean of the threshold (μ_b) was 0.34 for rats and 0.16 for humans, and the standard deviation of the threshold (σ_b) was 0.21 for rats and 0.16 for humans. These differences suggest that humans were more precise (lower mean of the threshold) and less variable (lower standard deviation of the threshold) than rats. Nonetheless, because the same process theory predicted both humans and rat performance, it suggests differences on species-specific parameter settings, but not underlying processes. The theoretical explanation showed great generality across procedures with very little parameter variability (for data from the same species).

Conclusions

The approach described to compare empirical and theoretical explanations of the data can also be used to compare among theoretical explanations of the data. A somewhat simple recipe is described in Table 6:

The emphasis of the present paper was on the capacity of theoretical explanations to generalize to other dependent measures and procedures. Typically, explanations of behavior originate as empirical explanations that gradually develop to be more theoretical. This development is characterized by an increase in the ability to predict new behavior in new conditions.

Another reviewer once wrote about predictability of behavior while evaluating a theoretical explanation of behavior: "A model that accounts for more features of performance than another model provides a more complete theory of behavior. This seems to me a reasonable way to chart theoretical progress and be a better way to choose among models." The conclusion of this paper is a paraphrased version of this statement: It is our hope that the development of relatively simple theoretical explanations that generalize across samples, dependent variables, and procedures (with minimum parameter estimation flexibility), will provide predictions of new behavior that take into account observed past behavior.

Acknowledgements

This research was supported by National Institute of Mental Health Grant MH44234 to Brown University. The authors would like to acknowledge the comments from Mika MacInnis, Linlin Yi, Marcelo Caetano, David Freestone, and Laura Ortega, and the contributions of many reviewers who provided us with motivation and insights for many of the ideas described.

References

- Akaike, H. Information theory and an extension of the maximum likelihood principle. In: P, BN.; Caski, AF., editors. Second international symposium on information theory. Akademiai Kiado; Budapest: 1973. p. 267-281.

- Bozdogan H. Akaike's Information Criterion and Recent Developments in Information Complexity. *J Math Psychol* 2000;44:62–91. [PubMed: 10733858]
- Browne MW. Cross-Validation Methods. *J Math Psychol* 2000;44:108–132. [PubMed: 10733860]
- Bussemeyer JR, Wang YM. Model Comparisons and Model Selections Based on Generalization Criterion Methodology. *J Math Psychol* 2000;44:171–189. [PubMed: 10733863]
- Bush RR, Mosteller F. A mathematical model for simple learning. *Psychol Rev* 1951;58:313–323. [PubMed: 14883244]
- Bush, RR.; Mosteller, F. Stochastic models for learning. Wiley; New York, NY: 1955. p. 365
- Catania, CA. Reinforcement schedules and psychophysical judgements: A study of some temporal properties of behavior. In: Schoenfeld, WN., editor. *The Theory of Reinforcement Schedules*. Appleton-Century-Crofts; New York, NY: 1970. p. 1-42.
- Church RM, Guilhardi P. A Turing test of a timing theory. *Behav Process* 2005;69:45–58.
- Collyer CE. Goodness-of-fit patterns in a computer cross-validation procedure comparing a linear and a threshold model. *Behav Res Meth Ins C* 1986;18:618–622.
- Gibbon J. Scalar expectancy theory and Weber's law in animal timing. *Psychol Rev* 1977;84:279–325.
- Gibbon J. Origins of scalar timing. *Learn Motiv* 1991;22:3–38.
- Gibbon J, Church RM. Representation of time. *Cognition* 1990;37:23–54. [PubMed: 2269007]
- Gibbon J, Church RM, Meck WH. Scalar timing in memory. *Ann NY Acad Sci* 1984;423:52–77. [PubMed: 6588812]
- Guilhardi P. A Comparison of Empirical and Theoretical Explanations of Temporal Discrimination. *J Exp Psychol Anim B*. in press
- Guilhardi, P.; Church, RM. Guilhardi-BRMIC-2004.zip. 2004a. Retrieved March 20, 2008 from Psychonomic Society Web Archive: <http://www.psychonomic.org/ARCHIVE/>
- Guilhardi P, Church RM. Measures of temporal discrimination in fixed-interval performance: A case study in archiving data. *Behav Res Meth Ins C* 2004b;36:661–669.
- Guilhardi P, Church RM. Dynamics of temporal discrimination. *Learn Behav* 2005;33:399–416. [PubMed: 16573211]
- Guilhardi P, Keen R, MacInnis MLM, Church RM. How rats combine temporal cues. *Behav Process* 2005;69:189–205.
- Guilhardi P, Yi L, Church RM. A modular model of learning and performance. *Psychon B Rev* 2007;14:543–555.
- Kirkpatrick K. Packet theory of conditioning and timing. *Behav Process* 2002;57:89–106.
- Kirkpatrick K, Church RM. Tracking of the expected time to reinforcement in temporal conditioning procedures. *Learn Behav* 2003;31:3–21. [PubMed: 18450066]
- Machado A. Learning the temporal dynamics of behavior. *Psychol Rev* 1997;104:241–265. [PubMed: 9127582]
- Mosier CI. The need and means of cross validation. I. Problems and designs of cross-validation. *Educ Psychol Meas* 1951;11:5–11.
- Myung IJ. The importance of complexity in model selection. *J Math Psychol* 2000;44:190–204. [PubMed: 10733864]
- Pitt MA, Myung IJ, Zhang S. Toward a method of selecting among computational models of cognition. *Psychol Rev* 2002;109:472–491. [PubMed: 12088241]
- Rescorla, RA.; Wagner, AR. A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and nonreinforcement. In: Black, A.; Prokasy, WF., editors. *Classical Conditioning II*. Appleton-Century-Crofts; New York, NY: 1972. p. 64-99.
- Roberts S. Isolation of an internal clock. *J Exp Psychol Anim B* 1981;7:242–268.
- Schwarz G. Estimating the dimension of a model. *Ann Stat* 1978;6:461–464.
- Stone M. Cross-validated choice and assessment of statistical predictions. *J Roy Stat Soc B Met* 1974;36:111–147.
- Turing AM. Computing machinery and intelligence. *Mind* 1950;59:433–460.
- Wasserman L. Bayesian Model Selection and Model Averaging. *J Math Psychol* 2000;44:92–107. [PubMed: 10733859]

- Yi L. Applications of timing theories to a peak procedure. *Behav Process* 2007;75:188–198.
- Zucchini W. An Introduction to Model Selection. *J Math Psychol* 2000;44:41–61. [PubMed: 10733857]

Figure 1.

The data are from rats trained on a 60- and 120-s peak procedure (adapted from Yi, 2007). All four panels show the same data (closed and open circles) fit by different explanations (lines). The right panels show the fits of three theoretical explanations, namely SET, LeT, and modular theory. The left panel shows the same data fitted by an empirical explanation, namely a Gaussian function.

Figure 2. Fits of an empirical explanation (ogive function) to the data of individual rats trained on a fixed-interval 30 s (left column), 60 s (center column), and 120 s (right column).

Figure 3.
A modular theory of learning and performance (Guilhardi et al., 2007)

Figure 4. Functional forms of the perception, memory, and decision modules of the modular theory (Guilhardi et al., 2007)

Figure 5.

Fits of the theoretical explanation (modular theory) to the data of individual rats trained on a fixed-interval 30 s (left column), 60 s (center column), and 120 s (right column).

Figure 6. Turing test procedure (left panel) and results for the theoretical explanation without error (top-right panel) and with error (bottom-right panel).

Peak-Interval Procedure

Figure 7.
A peak-interval procedure used for rats (adapted from Yi, 2007).

Figure 8.

Fits of the theoretical explanation to the response rate of rats trained on a peak-interval procedure with parameters values restricted to previously estimated values from rats trained on a fixed-interval procedure.

Peak-Interval Procedure (Humans)

Figure 9.
A peak-interval procedure used for humans.

Figure 10.

Fits of the theoretical explanation to the response rate of humans trained on a peak-interval procedure with parameters values restricted to previously estimated values from rats trained on a fixed-interval procedure (top panel) or no parameter restrictions (bottom panel).

Mean estimates of parameters of the empirical explanation for individual rats. The values in parenthesis indicate the standard errors of the mean. An analysis of variance with one between-subjects factor (FI duration) was performed to identify differences across fixed-intervals for each of the parameters (last two columns). The proportion of variance accounted for by the empirical equation for each interval is shown in the bottom row.

Table 1

Parameters	FI 30	FI 60	FI 120	F _(2,9)	p
End (<i>d</i>)	116.96 (3.84)	89.35 (10.04)	70.35 (6.79)	10.20	.005
Center (<i>c</i>)	21.07 (0.37)	41.69 (2.03)	81.93 (4.50)	117.35	<.001
Slope (<i>b</i>)	4.58 (0.32)	7.48 (0.62)	16.21 (1.97)	25.17	<.001
Start (<i>a</i>)	1.99 (0.84)	2.78 (0.98)	1.18 (0.75)	0.86	0.456
ω^2	>0.99 (0.002)	> 0.99(<0.001)	0.99 (0.001)		

Mean estimates of parameters of the theoretical explanation for individual rats. The values in parenthesis indicate the standard errors of the mean. An analysis of variance with one between-subjects factor (FI duration) was performed to identify differences across fixed-intervals for each of the parameters (last two columns). The proportion of variance accounted for by the theoretical explanation for each interval is shown in the bottom row.

Table 2

Parameters	FI 30	FI 60	FI 120	F _(2,9)	p
A	122.1 (3.02)	99.57 (10.56)	76.26 (5.39)	10.58	.004
μ_b	0.36 (0.01)	0.32 (0.03)	0.34 (0.03)	0.65	.545
γ_b	0.62 (0.04)	0.66 (0.13)	0.61 (0.03)	0.12	.888
R_0	3.17 (0.61)	3.96 (1.39)	2.03 (0.96)	0.88	.449
ω^2	> 0.99 (0.001)	> 0.99 (0.001)	> 0.99 (0.003)		

Table 3

Generality across samples using the cross-validation method. Parameters of the empirical and theoretical explanations were calculated by fitting the explanations to half of the data (calibration sample) and $\omega^2_{\text{calibration}}$ determined. The generality of the parameter estimations across samples was determined validating the fits based on the other half of the data (validation sample) and $\omega^2_{\text{validation}}$ determined. Overfitting was defined by the difference between $\omega^2_{\text{validation}}$ and $\omega^2_{\text{calibration}}$.

	Empirical	Theoretical
Mean $\omega^2_{\text{calibration}}$	0.987 (0.003)	0.985 (0.003)
Mean $\omega^2_{\text{validation}}$	0.975 (0.004)	0.975 (0.004)
Overfitting	0.012 (0.004)	0.010 (0.004)

Table 4

Eleven dependent measures calculated for the rats and the theoretical explanation ("Theory") for the three fixed-interval conditions. Results from an analysis variance with fixed-interval condition and theoretical explanation accuracy (rat vs. theory) as factors are also shown. The "Post-reinforcement pause" is an example of high agreement between the observed behavior and the estimate from the theory, although this dependent variable was not determined by any single parameter of the theory.

Dependent Variable	Source	FI 30 s	FI 60 s	FI 120 s	Interval Effect	Theory Accuracy
Curvature Index	Rat	0.46 (0.02)	0.49 (0.03)	0.52 (0.05)	$F_{(2,18)} = 1.37, p = .28$	$F_{(1,18)} = 0.02, p = .89$
	Theory	0.47 (0.01)	0.50 (0.03)	0.52 (0.05)		
Center of ogive	Rat	19.98 (0.13)	42.46 (1.92)	78.18 (4.77)	$F_{(2,18)} = 221.25, p < .001$	$F_{(1,18)} = 0.31, p = .59$
	Theory	20.96 (0.32)	42.29 (2.87)	81.24 (3.73)		
Slope of ogive	Rat	4.09 (0.31)	7.92 (0.77)	14.83 (2.39)	$F_{(2,18)} = 31.87, p < .001$	$F_{(1,18)} = 0.11, p = .75$
	Theory	4.26 (0.16)	7.95 (0.65)	15.78 (2.28)		
Point of transition t_1	Rat	19.62 (0.10)	39.33 (1.24)	76.66 (3.77)	$F_{(2,18)} = 378.42, p < .001$	$F_{(1,18)} = 2.03, p = .17$
	Theory	20.33 (0.11)	41.65 (1.54)	81.22 (3.21)		
Rate prior to t_1 (Low)	Rat	6.17 (0.65)	4.96 (0.33)	3.43 (1.19)	$F_{(2,18)} = 3.78, p = .04$	$F_{(1,18)} = 2.05, p = .17$
	Theory	4.65 (0.46)	4.26 (1.43)	2.42 (0.94)		
Rate after t_1 (High)	Rat	131.04 (1.52)	103.61 (8.82)	76.01 (8.21)	$F_{(2,18)} = 30.35, p < .001$	$F_{(1,18)} = 0.08, p = .78$
	Theory	127.44 (4.47)	101.68 (9.44)	76.75 (4.42)		
Post-reinforcement pause	Rat	14.86 (0.51)	25.00 (1.59)	50.32 (7.91)	$F_{(2,18)} = 16.88, p < .001$	$F_{(1,18)} = 0.04, p = .84$
	Theory	16.00 (0.85)	26.88 (4.85)	50.45 (11.79)		
Quarter-life	Rat	21.20 (0.20)	41.88 (0.90)	82.05 (4.21)	$F_{(2,18)} = 330.47, p < .001$	$F_{(1,18)} = 1.04, p = .32$
	Theory	21.55 (0.25)	43.34 (1.67)	86.46 (3.97)		
Time of Median Response	Rat	24.71 (0.17)	49.22 (0.69)	96.83 (2.70)	$F_{(2,18)} = 1117.54, p < .001$	$F_{(1,18)} = 0.65, p = .43$
	Theory	25.08 (0.15)	49.49 (0.98)	99.30 (2.47)		

Dependent Variable	Source	FI 30 s	FI 60 s	FI 120 s	Interval Effect	Theory Accuracy
Rate during stimulus (rpm)	Rat	46.00 (0.82)	36.00 (4.88)	28.03 (2.82)	$F_{(2,18)} = 12.05, p < .001$	$F_{(1,18)} = 2.72, p = .12$
	Theory	39.37 (1.10)	32.00 (4.81)	25.56 (2.50)		
Temporal Discrimination Ratio	Rat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	-	-
	Theory	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)		

Table 5

Eleven dependent measures calculated for the rats and the theoretical explanation with error (“Theory”) for the three fixed-interval conditions. Results from an analysis of variance with fixed-interval condition and theoretical explanation accuracy (rat vs. theory) as factors are also shown. The italicized dependent variables show failure of the theoretical explanation to predict these measures due to the error introduced in the theoretical explanation.

Dependent Variable	Source	FI 30 s	FI 60 s	FI 120 s	Interval Effect	Theory Accuracy
Curvature Index	Rat	0.46 (0.02)	0.49 (0.03)	0.52 (0.05)	$F_{(2,18)} = 0.03, p = .33$	$F_{(1,18)} = 0.03, p = .86$
	Theory	0.47 (0.01)	0.49 (0.02)	0.51 (0.05)		
Center of ogive	Rat	19.98 (0.13)	42.46 (1.92)	78.18 (4.77)	$F_{(2,18)} = 231.79, p < .001$	$F_{(1,18)} = 0.01, p = .93$
	Theory	20.75 (0.45)	40.53 (2.04)	78.78 (3.74)		
Slope of ogive	Rat	4.09 (0.31)	7.92 (0.77)	14.83 (2.39)	$F_{(2,18)} = 24.09, p < .001$	$F_{(1,18)} = 0.23, p = .63$
	Theory	4.41 (0.39)	8.09 (0.57)	16.32 (3.12)		
Point of transition t_1	Rat	19.62 (0.10)	39.33 (1.24)	76.66 (3.77)	$F_{(2,18)} = 437.51, p < .001$	$F_{(1,18)} = 0.92, p = .35$
	Theory	19.08 (0.25)	37.83 (0.76)	74.17 (2.43)		
<i>Rate prior to t_1 (Low)</i>	Rat	6.17 (0.65)	4.96 (0.33)	3.43 (1.19)	$F_{(2,18)} = 4.84, p = .02$	$F_{(1,18)} = 14.88, p = .001$
	Theory	10.50 (0.71)	8.16 (1.19)	6.35 (1.86)		
<i>Rate after t_1 (High)</i>	Rat	131.04 (1.52)	103.61 (8.82)	76.01 (8.21)	$F_{(2,18)} = 28.38, p < .001$	$F_{(1,18)} = 15.34, p < .001$
	Theory	104.89 (2.29)	81.60 (9.39)	61.52 (3.86)		
Post-reinforcement pause	Rat	14.86 (0.51)	25.00 (1.59)	50.32 (7.91)	$F_{(2,18)} = 21.51, p < .001$	$F_{(1,18)} = 2.56, p = .13$
	Theory	12.27 (0.51)	19.45 (2.84)	39.22 (8.43)		
Quarter-life	Rat	21.20 (0.20)	41.88 (0.90)	82.05 (4.21)	$F_{(2,18)} = 281.67, p < .001$	$F_{(1,18)} = 1.28, p = .27$
	Theory	19.71 (0.49)	39.45 (1.12)	78.84 (4.43)		
Time of Median Response	Rat	24.71 (0.17)	49.22 (0.69)	96.83 (2.70)	$F_{(2,18)} = 1134.45, p < .001$	$F_{(1,18)} = 0.64, p = .43$
	Theory	23.98 (0.19)	48.03 (0.75)	95.72 (2.40)		

Dependent Variable	Source	FI 30 s	FI 60 s	FI 120 s	Interval Effect	Theory Accuracy
Rate during stimulus (rpm)	Rat	46.00 (0.82)	36.00 (4.88)	28.03 (2.82)	$F_{(2,18)} = 15.26, p < .001$	$F_{(1,18)} = 0.43, p = .52$
	Theory	43.50 (0.96)	34.75 (4.33)	26.75 (2.70)		
Temporal Discrimination Ratio	Rat	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	-	-
	Theory	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)		

Table 6
Recipe for comparison among theoretical explanations of data.

Approach for comparison among theoretical explanations of data	
1	Given some data (e.g., times of responses);
2	Calculate a summary measure (e.g., response rate gradient)
3	Choose the theories for comparison
4	Estimate the parameters
5	Simulate the theories
6	Evaluate:
	a. Adequacy
	b. Simplicity
	c. Generality
	i. Across samples
	ii. Across dependent measures
	iii. Across procedure