Available online at www.sciencedirect.com
SCIENCE DIRECT®
Behavioural Processes 69 (2005) 45–58
BEHAVIOURAL PROCESSES
www.elsevier.com/locate/behavproc

A Turing test of a timing theory☆

Russell M. Church*, Paulo Guilhardi

Department of Psychology, Box 1853, Brown University, Providence, RI 02912, USA

Received 29 October 2004; received in revised form 3 January 2005; accepted 3 January 2005

Abstract

A quantitative theory of timing or conditioning can be evaluated with a Turing test in which the behavioral results of an experiment can be compared with the predicted results from the theory. An example is described based upon an experiment in which 12 rats were trained on three fixed-interval schedules of reinforcement, and a simulation of the predicted results from a packet theory of timing. An objective classification rule was used to determine whether a sample from the data or a sample from the theory was more similar to another sample from the theory. With an ideal theory, the expected probability of a correct classification would be 0.5. The observed probability of a correct classification was 0.6, which was slightly, but reliably, greater than 0.5. A Turing test provides a graded metric for the evaluation of a quantitative theory.

Keywords: Time discrimination; Turing test; Theory evaluation

Procedures used in the study of animal timing have led to the identification of reliable functional relationships between quantitative features of procedures and results. For example, a fixed-interval schedule of reinforcement leads to a response rate that increases as a function of time since the previous reinforcement (Dews, 1970). The relative response rate (as a proportion of the maximum response) increases as a function of the fixed-interval (as a proportion of the total interval). The same function typically provides a good approximation of the behavior under a wide range of fixed-intervals. This is the superposition result, which has also been called "timescale invariance" (Church, 2002).

Theories of timing have been developed that account for these quantitative results. Examples of these are scalar timing theory (Gibbon, 1977), behavioral theory of timing (Killeen and Fetterman, 1988), learning to time model (Machado, 1997), spectral timing model (Grossberg and Schmajuk, 1991), multiple oscillator model (Church and Broadbent, 1990), the multiple time-scale model (Staddon and Higa, 1999), Packet theory (Kirkpatrick, 2002; Kirkpatrick and Church,

☆ Manuscript based on symposium in honor of Donald S. Blough at Conference on Comparative Cognition (CO3), Melbourne, FL, March, 2004.

* Corresponding author. Tel.: +1 401 863 2328; fax: +1 401 863 1300.
E-mail address: Russell_Church@Brown.edu (R.M. Church).

0376-6357/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.beproc.2005.01.001

2003), but there are many others. (In these quantitative timing theories, the terms "theory" and "model" are used interchangeably.) One of the purposes of these theories is to account for the behavior of the animal in timing procedures based on simple assumptions of a well-specified process model.

One of the bases for the evaluation of a quantitative theory of timing is the extent to which it fits observed data. This is usually done by defining one or more summary measures of behavior, and comparing the predictions of the model with the values of the observed data. The variance of the difference of the observed from the predicted values of the summary measure (unexplained variance) is typically compared to the variance of the difference of the observed from the mean values of the summary measure (total variance). The standard index is the proportion of variance accounted for:

$$\omega^2 = \frac{\sigma_t^2 - \sigma_u^2}{\sigma_t^2} \quad (1)$$

where ω^2 is the proportion of variance accounted for, σ_u^2 the unexplained variance, and σ_t^2 is the total variance.

The main purpose of this article is to show how a Turing test provides an alternative way to evaluate the extent to which the predictions of a timing theory fit observed data. It describes a particular timing procedure, the results that were recorded, summary measures of the behavior, a quantitative process model (Packet theory), and comparison of the predictions of the model and the summary measures of behavior. It concludes with a description of the Turing test, how it can be used to evaluate a quantitative process model. The question is the extent to which a person, or a computer algorithm, can correctly discriminate between data that was generated by an experimental subject and data generated by a quantitative theory.

1. Specification of the procedures

Many timing and conditioning procedures can be described by the specification of a small number of stimuli, reinforcers, and responses, and the contingencies among these events. Such experiments are often conducted with well-known species in simple environments. They may use one or more stimuli (such as houselight, white noise, and clicker), one or more measured responses (such as head-entry into the food cup, lever press, and licking on the tube of a water bottle), and one or more reinforcers (such as food or shock). The standard time-line diagram shown in the top panel of Fig. 1 includes a single stimulus (houselight), a single response (head-entry), and a single reinforcer (a pellet of food).

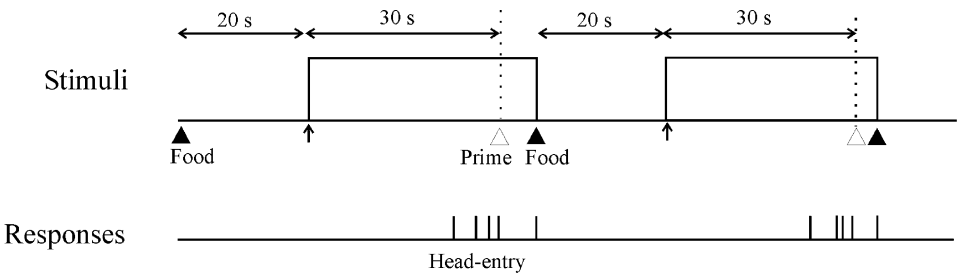
The procedure that is illustrated in the time-line diagram provides information regarding the contingencies of reinforcement. The time of an onset of a stimulus (such as the houselight) is indicated by an arrow; the time after which a response will be reinforced is indicated by an open triangle (labeled "Prime"); the time of a food delivery is indicated by a solid triangle, and the times of head-entry responses is indicated by the vertical marks on the line labeled "Responses."

The contingencies of reinforcement may be described in words, in a diagram, or with formal notation. They are nearly always described in a paragraph, but it is difficult to describe a procedure precisely, completely, and succinctly in words. The contingencies of reinforcement may also be described with a time-line diagram (as in the top panel of Fig. 1), but this requires more space than a formal notation. Unfortunately, the standard formal notation for conditioning procedures is too succinct to be useful for many purposes. For example, this procedure would be described as A+, where A is the symbol for the houselight, and + is the symbol for food. This does not provide information about the duration of the stimulus, the duration between stimuli, the time of delivery of the food during the stimulus (or even whether or not the food was delivered shortly after stimulus termination, as in trace conditioning).

A more complete formal notation is necessary to provide the information that is in a time-line diagram. The procedure in the top panel of Fig. 1 may be written as:

$$/20\,\text{s H}; 30\,\text{s} \rightarrow \text{h}\blacktriangle\blacktriangledown/ \quad (2)$$

with symbols for light onset (H), light termination (h), head-entry response (→), and food delivery (▲▼). This contains information about the duration of the stimulus, the duration between stimuli, and the time of delivery of food during the stimulus. The extension of this notation to many different procedures is in Appendix A.



Time-event list

Time (s)	Event Number	Event Name
0.00	19	Food
20.00	10	Stimulus Onset
44.00	8	Head-entry
45.20	8	Head-entry
46.00	8	Head-entry
49.80	8	Head-entry
50.00	53	Food Prime
55.00	8	Head-entry
55.00	19	Food
55.00	20	Stimulus Termination
75.00	10	Stimulus Onset
97.32	8	Head-entry
101.64	8	Head-entry
102.50	8	Head-entry
103.88	8	Head-entry
105.00	53	Food Prime
107.68	8	Head-entry
107.68	19	Food
107.68	20	Stimulus Termination

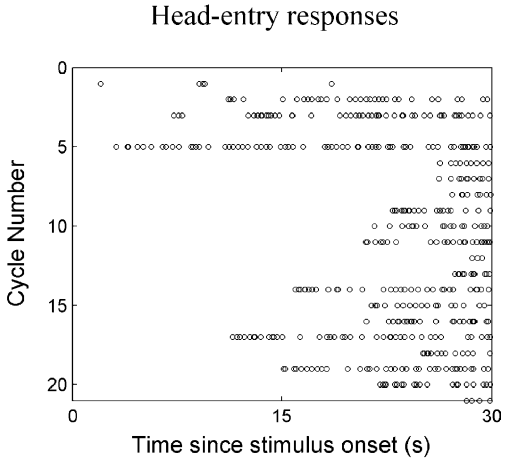


Fig. 1. Procedure and recording of data. Top panel: times of stimulus onset and termination, food availability, and delivery, and the times of head-entry responses are shown for two cycles. Bottom left panel: the time-event list contains three columns for the time in seconds from the beginning of the session, the number of each event, and the name of each event. Bottom right panel: a raster plot of the head-entry responses is shown as a function of time since onset of the light on 30 consecutive cycles with each head-entry response marked by an open circle. Note that most of the responses occurred in the latter portions of the 30-s interval.

2. Recording of the results

The primary data consists of a list of times at which each event occurs (lower left portion of Fig. 1). In this example, the session began with the delivery of food. The houselight went on after 20 s, head-entry responses occurred at 44.00, 45.20, 46.00, and 49.80 s, then food was primed at 50.00 and 55.00 s, the next head-entry response occurred, food was delivered, and the houselight was turned off. The time-event list continues with the data for the second cycle. Records were kept of the times (to the nearest 2 ms) and event numbers; the event numbers and names are redundant. Because they are more efficient for storage and analysis, numbers are typically used, but supplemented with a list of the event name corresponding to each number.

The lower right portion of Fig. 1 shows a raster plot of the head-entry responses. The data come from one rat on a 30-s discriminative fixed-interval schedule. The horizontal axis is the time since stimulus onset, and the vertical axis is the successive cycles in a session (from top to bottom). In a 30-s fixed-interval schedule, responses are clustered toward the end of the interval.

The procedure used to generate the data for this application of a Turing test of a quantitative timing theory was an extension of the simple 30-s discriminative fixed-interval schedule of reinforcement shown in Fig. 1. There were three cycle types (C_1, C_2, and C_3).

$$\begin{aligned} C_1 &= /20\,\text{s H}; 30\,\text{s} \rightarrow \text{h}\blacktriangle\blacktriangledown/ \\ C_2 &= /20\,\text{s N}; 60\,\text{s} \rightarrow \text{n}\blacktriangle\blacktriangledown/ \\ C_3 &= /20\,\text{s C}; 120\,\text{s} \rightarrow \text{c}\blacktriangle\blacktriangledown/ \end{aligned} \quad (3)$$

where H is houselight onset; N is onset of white noise; and C is onset of a clicker. The lower case letters are for the terminations of the stimuli. Thus, in this multiple cued interval procedure, there were three possible intervals (30, 60, and 120 s), each with a different discriminative stimulus. (Of course, the interval and stimulus type was counterbalanced across rats.) The three types of cycles were sampled with replacement, as shown by the following notation:

$$\mathbf{T} = \left[\left(\frac{1}{3}\right) C_1 \left(\frac{1}{3}\right) C_2 \left(\frac{1}{3}\right) C_3\right] \quad (4)$$

Each of 12 rats had 30 sessions of 60 cycles per session with this procedure, and the last 15 of these sessions was used for the present analysis. This was the simultaneous group used in phase 1 of an analysis of the acquisition of temporal discrimination (Guilhardi and Church, submitted for publication).

3. Summary measures of results

Typically, analysis of results consists of the description of summary measures of behavior. Nonetheless, it is useful to record and retain the primary data. From the primary data, it is possible to calculate any summary measure, but the primary data cannot be recreated from the summary measures. The availability of the primary data greatly facilitates comparison of the effects of procedures that were reported with different summary measures; it makes it unnecessary to repeat an experiment in order to analyze a different summary measure. The availability of primary data facilitates secondary data analysis (Guilhardi and Church, 2004; Kurtzman et al., 2002).

A summary measure is often chosen because it is conventional in a subfield. This is a reasonable decision because it facilitates comparison of results, although the comparison is restricted to this single summary measure. In some cases, a particular summary measure is chosen because it is considered to be diagnostic of some important concept or because it is widely used as an operational definition of some important concept. The distinction between a summary measure as diagnostic of a concept or simply an operational definition of a concept is often uncertain.

Multiple summary measures are sometimes used to characterize behavior. In some cases, the summary measures may be independent of each other (i.e., they are not redundant). For the results of a multiple cued interval procedure, the summary measures sensitive to response rate, response pattern, and response bouts may be independent of each other.

3.1. Rate

The mean response rate (in responses per minute) during the last 15 sessions during the stimulus of the 30, 60 and 120-s interval is shown in the top panel of Fig. 2. The mean response rates in the three conditions were (48.2, 44.5, and 36.2 responses per minute with the three intervals, respectively), with the standard error of the mean shown by the error bars (3.7, 4.3, and 3.6 responses per minute, respectively). The differences in response rate at different interval durations were significant ($F_{(2,22)} = 21.3$, $P < 0.001$).

3.2. Pattern

The mean response rate can be calculated as a function of time since stimulus onset for each of the interval durations. These response gradients show both the temporal pattern and the overall response rate. The relative response rate was defined as the mean response rate divided by the maximum rate. The relative response rate as a function of time since stimulus onset is shown in the middle panel of Fig. 2. The three functions were

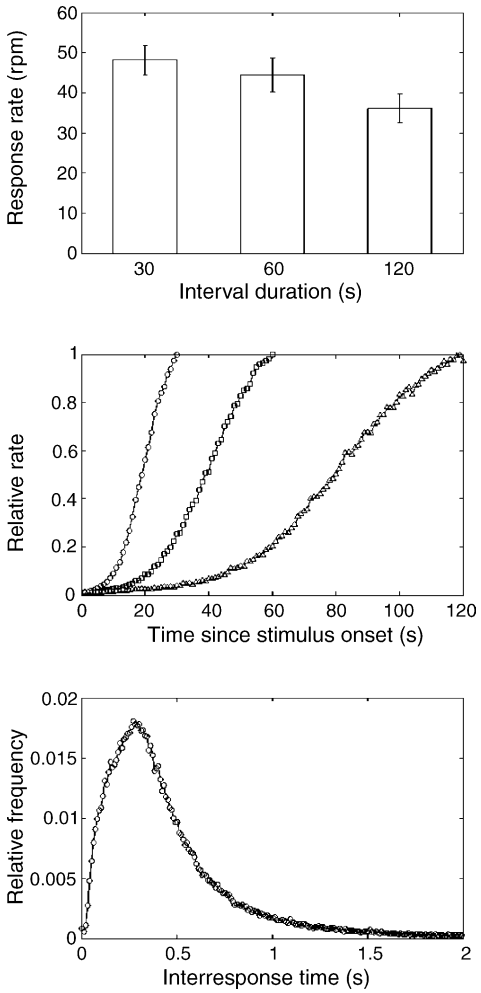


Fig. 2. Three results. Top panel: mean response rate (per minute) as a function of interval duration. Middle panel: relative response rate as a fraction of the maximum rate as a function of interval duration. Bottom panel: relative frequency of interresponse times as a function of all interresponse times in seconds.

obviously very different and related to the interval duration.

3.3. Bout

The relative frequency distribution of interresponse times is defined as the frequency distribution of interresponse times divided by the number of interresponse times. This distribution is shown in the bottom panel of Fig. 2. Most of the interresponse interval were under 2 s, and clustered near a mode at 0.272 s.

4. A process theory

With a specification of the procedures, the recording of the results, and the analysis of summary measures of behavior, some attempts are often made to explain the results. This section describes a quantitative theory that attempts to explain the behavior resulting from timing procedures.

Quantitative theories of timing typically attempt to account for selected summary measures of behavior, and some of them do so for several quite different summary measures such as relative response gradients as a measure of timed performance and choice between two responses as a measure of time perception. A Turing test can be applied to these theories, but the questions must be restricted to the specific procedures to which they apply and the specific summary measures that the procedures have used.

Packet theory is unique in attempting to account for the time of occurrence of individual responses in any procedure in which any stimuli may be regarded as qualitatively different from all others. Predictions from Packet theory can be made for any procedure with any contingency of reinforcement between stimuli, and responses. The wide range of legitimate procedures provides input generality; the prediction of times of responses, which makes it possible to make predictions about any summary measure, provides output generality. This makes it particularly suitable for evaluation with a Turing test.

The essential features of Packet theory were described by Kirkpatrick (2002) and Kirkpatrick and Church (2003). This is referred to as Version 1. It was slightly modified by Guilhardi et al. (in press), which will be referred to as Version 2. For acquisition of a temporal discrimination, it was further modified by Guilhardi and Church (submitted for publication) but because those modifications had inconsequential effects on asymptotic performance, it will be referred to as Version 2a. Version 2a of Packet theory is used in this article.

The input consists of the time of each stimulus onset and each food delivery. Unless otherwise specified, the unit of time will be seconds. This input is transformed by the rules of temporal perception, temporal memory, and temporal decision to produce bouts of responses. The rules are specified in the equations below, and illustrated in the four panels of Fig. 3.

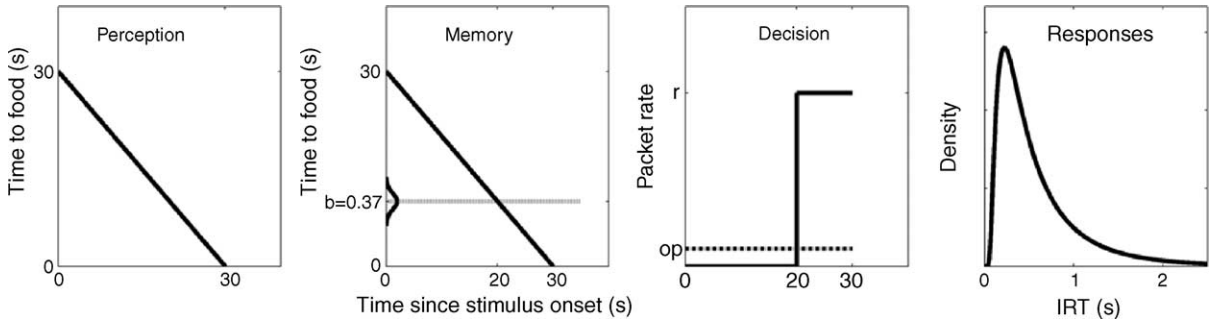


Fig. 3. Packet model of timing. First panel: perception of the time since food. Second panel: memory of the time to food, with a variable threshold. Third panel: decision to initiate a packet of responses based on memory and an operant rate. Fourth panel: distribution of interresponse intervals in a packet. See text for details.

4.1. Perception

The perceived time to food is determined by the duration between the last food delivery and the preceding stimulus onset (perception panel of Fig. 3).

$$e_i(t) = d_i - t \tag{5}$$

where d_i is the duration between the last reinforcement and the preceding stimulus onset, t the time since stimulus onset in seconds, and $e_i(t)$ is the perceived time to reinforcement as a function of time since stimulus onset. In this example, the duration between the last reinforcement and the preceding stimulus onset was 30 s. Thus, the perceived time to reinforcement was 30 s at stimulus onset and it decreased linearly for 30–0 s. This transformation contains no free parameters.

4.2. Memory

The updated remembered time to food is a weighted mean of the perceived time to food and the previous remembered time to food (memory panel of Fig. 3).

$$E_{i+1}(t) = \alpha\, e_i(t) + (1 - \alpha)E_i(t) \tag{6}$$

where $e_i(t)$ is the current perceived time to reinforcement as a function of time since stimulus onset, $E_i(t)$ the previous remembered time to food as a function of time since stimulus onset, α the learning rate (a value between 0 and 1), and $E_{i+1}(t)$ is updated remembered time to food as a function of stimulus onset. In this example, the value of α was 0.0125 (a value previously used for acquisition); the value of α does not have much effect at asymptote. Of course, if the perception and the memory were identical, the remembered time to food would not be changed by Eq. (6). The initial remembered time to food was vector with a length equal to the number of seconds from stimulus onset to food containing random values from a normal distribution with a mean of 400 s, and a standard deviation of 280 s. These initial values also did not have much effect at asymptote.

The horizontal line in the memory panel is a threshold that determines whether the animal will be in the low or high response state. The proportion of time during the stimulus in which the animal will be in a high state is determined as follows: in every cycle, the threshold is a proportion which is a single random sample from a normal distribution with a mean of 0.37 and a coefficient of variation of 0.44; the time of this threshold (b) is the remembered time to food, such that the proportion below the threshold is b. If the remembered time to food is above the threshold, the animal is in the low state; if it is below the threshold, the animal is in the high state. These transformations consists of five free parameters (α, μ, σ/μ, and the mean and standard deviation of the initial remembered time to food). At asymptote, the most influential parameters are μ, and σ/u.

4.3. Decision

If the animal is in the high state, the rate of packet initiation is r packets per second. This is shown by the step function in the Decision panel of Fig. 3 that begins

at 0 and then, at 20 s after stimulus onset, goes to the rate of r packets per second.

$$r = -0.8 \log_{10} E_i(0) + 2.0 \quad (7)$$

where $E_i(0)$ is the expected time to food at onset of a stimulus.

There is also an operant rate of packet initiation that occurs throughout all sessions shown by the horizontal dotted line in the decision panel of Fig. 3. The operant rate of packet initiation was 0.01 per second. During the low state the rate of packet initiation was the operant rate (op), and in the high state it was the sum of r and op. This transformation consists of three free parameters: op, and the slope and intercept of the function relating rate of packet initiation to the mean interfood interval (Eq. (7)).

4.4. Response

If a packet is initiated, responses may occur in a clustered manner. The number of responses in the cluster is a random sample from a Poisson distribution with a mean of five responses; the distribution of these responses is distributed as a Wald distribution (an inverse Gaussian shown in the response panel of Fig. 3) which has two parameters, a mean of 0.54 s and a standard deviation of 0.71 s. This consists of three free parameters (mean of Poisson, and mean and standard deviation of Wald).

5. Comparison of predictions of model and summary measures of results

A comparison of the predictions of a quantitative model, such as Packet theory, with the behavioral results is usually based on summary measures of responding, such as those shown in Fig. 2. The primary question is usually the extent to which the model provides a good fit to the data, and the measure of goodness-of-fit is the proportion of variance accounted for by the model, ω^2, as defined in Eq. (1). Note that this is a comparison of the model under consideration and an alternative model, that all values are at the mean.

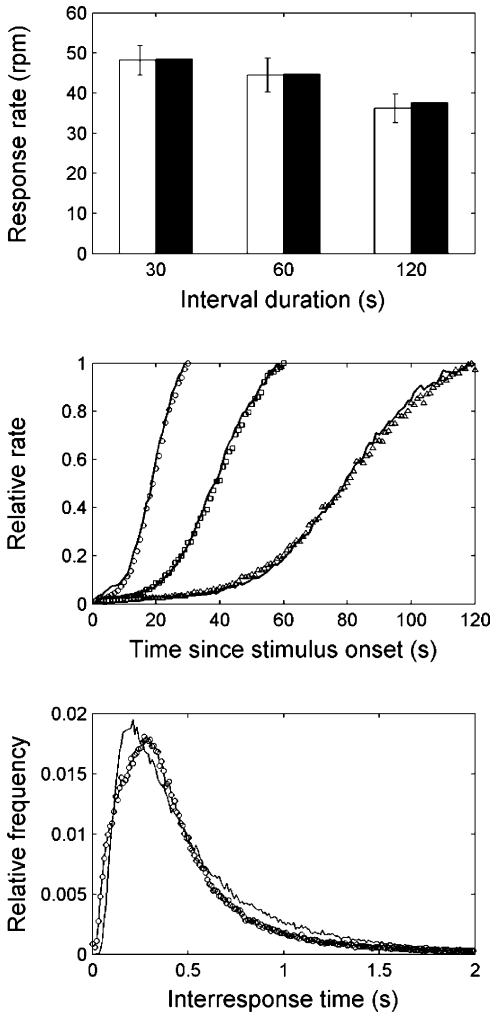


Fig. 4. Comparison of predictions of model and behavior of rats. Top panel: the mean response rate (per minute) as a function of interval duration is shown with solid bars for the simulation and open bars for the data. Middle panel: relative response rate as a function of interval duration is shown with thin lines for the simulation and open symbols for the data. Bottom panel: relative frequency of interresponse times as a function of time since previous response in seconds with thin lines for the simulation and open symbols for the data.

5.1. Rate

A comparison of the observed response rate (open bars) with predictions of the model (solid bars) is shown in the top panel of Fig. 4. The data and the predictions are similar ($\omega^2 = 0.973$).

5.2. Pattern

A comparison of the observed response gradients (data points) with predictions of the model (thin lines) is shown in the middle panel of Fig. 4. The data and the measures are similar ($\omega^2 = 0.994$, 0.997, and 0.997 for the 30, 60, and 120-s intervals, respectively).

5.3. Bout

A comparison of the observed bout structure (open circles) with predictions of the model (thin line) is shown in the bottom panel of Fig. 4. Although the discrepancies of the observed and predicted functions are not large ($\omega^2 = 0.974$), the systematic pattern of deviations clearly indicates that the predicted function is just an approximation.

5.4. Evaluation of the model based on summary measures

Based on the summary measures of rate and pattern shown in Fig. 4, it appears that the predictions of the model are nearly identical to the data. The predictions of bout structure indicated that the function forms were not correct, but the discrepancies were small.

The model had five parameters for memory, three for decision, and three for response. A single estimate of each of these 11 parameters was used in fitting the 413 data points in Fig. 4. Some of the estimates have only negligible effects on the summary measures, and some of the estimates may apply to other procedures. Thus, the model is not unduly complex.

On the basis of the analysis of these summary measures, it would appear to be difficult or impossible to discriminate between data that was generated by Packet theory and data generated by a rat in this multiple cued interval procedure. Section 7 will show that such a discrimination can be made. This will indicate that, despite excellent fits based on these three nonredundant summary measures, the Packet theory of multiple cued interval procedure is incomplete.

6. A Turing test

The purpose of this section is to describe the original Turing test, and its application to the evaluation of a timing theory.

6.1. Original Turing test

In 1950, Alan Turing introduced a method to answer the question, “Can machines think?” He proposed a behavioristic approach, a modification of the imitation game, in which an interrogator asks questions and attempts to determine whether the typed answers are coming from a person or a programmed computer (Turing, 1950). This test is now known as “the Turing test.” People are regarded as intelligent, so, if the interrogator is not able to distinguish between the answers from a person and a computer, it may be that the computer is also intelligent, i.e., it can think. Turing’s prediction was:

“I believe that in about 50 years’ time it will be possible to programme computers, with a storage capacity of about 10^9, to make them play the imitation game so well than an *average* interrogator will not have more than *70 percent chance* of making the right identification after *five minutes* of questioning.” (Turing, 1950, p. 442, italics added)

The italics were added to emphasize the graded nature of the criterion that consisted of the skill of the interrogator, the percentage of correct identification, and the length of the questioning.

6.2. A Turing test of a behavioral theory

The Turing test can be readily adapted to evaluate a quantitative theory of behavior. In the original Turing test, an interrogator asks questions to a person or a computer program; in the adaptation of the Turing test, an experimenter administers a procedure to an animal or a computer program. In the original Turing test, either a person or a computer program provides the answers; in the adaptation of the Turing test, either the animal or the computer program responds. In the original Turing test, an interrogator classifies the answer as either coming from a person or a computer program; in the adaptation of the Turing test, a computer algorithm classifies the results as coming either from an animal or from a computer program (Church, 1997, 2001).

The adaptation of the original Turing test to evaluate a quantitative theory of behavior is illustrated in Fig. 5. The procedure delivers stimuli and reinforcers to the animal and receives responses from the animal; the

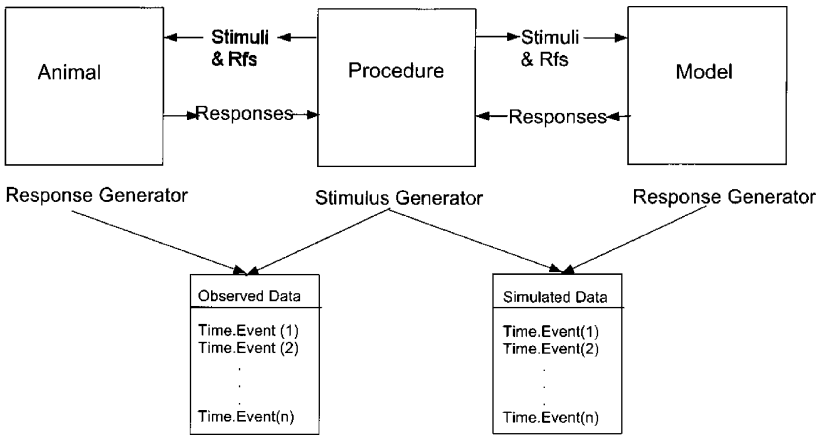


Fig. 5. A Turing test for the evaluation of a quantitative theory of behavior. The same procedure is used for delivering stimuli and reinforcers to the animal and the model, and receiving responses from the animal and the model. The observed data consists of times of the events (stimuli, reinforcements, and responses) from the animal and the procedure; the simulated data consists of times of events from the model and the procedure. (Church, 1997, Copyright 1997 by the American Psychological Association. Adapted with permission).

observed data consists of a list of times at which various events occur, such as stimulus onset and termination, reinforcer delivery, and responses (see also Fig. 1). The procedure also delivers stimuli and reinforcers to the model and receives responses from the model; the simulated data consists of another list of times at which various events occur.

The list of the observed data can be compared with the list of simulated data to determine whether a particular list was produced by an animal or by a computer model. This should be done with an objective evaluation algorithm. Church (1997) wrote:

"Our goal should be to develop models that produce sequences of times of occurrence of events (stimuli and responses) that are indistinguishable from those produced by the animal under many experimental procedures and data analysis techniques (a Turing test)."

The original article included a figure similar to Fig. 5 that included an extension to evaluate models with each other, as well as with the data.

It would now seem reasonable to relax the goal from "indistinguishable" to "less than 1% error." The current error is about 10%, as described in the next section. If the error rate can be reduced by one-half in each of the next 5 years (10, 5, 2.5, 1.25, and 0.625), the goal of an error rate of less than 1% can be reached in 5 years.

7. Comparison of predictions of a model and the primary data

In the present experiment, the events were stimuli (houselight, noise, and clicker), a response (head-entry), and a reinforcer (food), as shown in Eq. (3). One cycle of this experiment consisted of the interval between two successive deliveries of food. This might be a 20-s interval followed by the onset of a houselight, a 30-s interval followed by a head-entry response, termination of the houselight and delivery of food (Eq. (2)).

The observed rat data (dashed line) came from one cycle from a rat; the simulated model data (dotted line) came from one cycle of Packet theory; the sample model data (thick black line) came from another cycle of Packet theory (Fig. 6). The dependent variable, local rate, was calculated directly from the times of responses as a function of time since stimulus onset. The local rate in responses per minute at a given time is the reciprocal of the interresponse time in minutes. Thus, if two responses are separated by a 1-s interval, the local rate is 60 response per minute (i.e., interresponse time was 1/60 min per response, the reciprocal of which is 60 responses per minute). To reduce the influence of large fluctuations in response rate with very small changes in short interresponse times, the logarithm of the local rate is used in the analysis. It should be noted that this local rate measure contains all the

information available in the times of responses in the primary data–it is possible to reconstruct the original times of the responses based on a sequence of local rate functions for each cycle. (This is also true of the more familiar cumulative record that could also be used.)

The psychophysical procedure that was used is known as a matching-to-sample. The task was to determine whether the observed data or the simulated data more closely resembled (matched) the sample. In this case, it is clear that the sample was more similar to the observed data than the simulated data. Thus, the classification on this cycle would be incorrect.

Of course, there was no need to rely upon human judgments of similarity. The objective index used was:

$$\begin{array}{ll} \text{if median } |r-s| > \text{median } |m-s|, & I=1 \\ \text{else} & I=0 \end{array} \qquad (8)$$

where the logarithm (base 10) of the local response rate in responses per minute of the rat, model, and sample were r, m, and s, respectively.

Since the sample was drawn from the model, a correct identification ($I = 1$) was when the model and the sample were more similar than the rat and the sample. The mean probability of a correct identification (and the standard error of the mean) is shown for the three fixed-interval conditions (30, 60, and 120 s) in the bottom panel of Fig. 6. There were 12 rats, and about 150 cycles for each rat at each of the three intervals.

The mean probability of a correct identification was 0.60, with a standard error of 0.07. This is reliably greater than 0.50 ($t_{11} = 8.6$, $P < 0.001$). The probability of correct identification increased as a function of interval ($F_{2,22} = 4.9$, $P < 0.02$). Although the mean probability of a correct identification of 0.60 demonstrates that Packet theory was substantially better than the worst possible theory in which correct classification would always occur (1.0), it is reliably worse than an ideal theory in which correct classification would occur at chance (0.5).

The essential features of the index used for classification (Eq. (8)) are: (a) the similarity measures are based on data which contain all information necessary to recreate the times of responses, and (b) an objective similarity measure is used. The logarithm of the local response rate on each cycle contains all the information

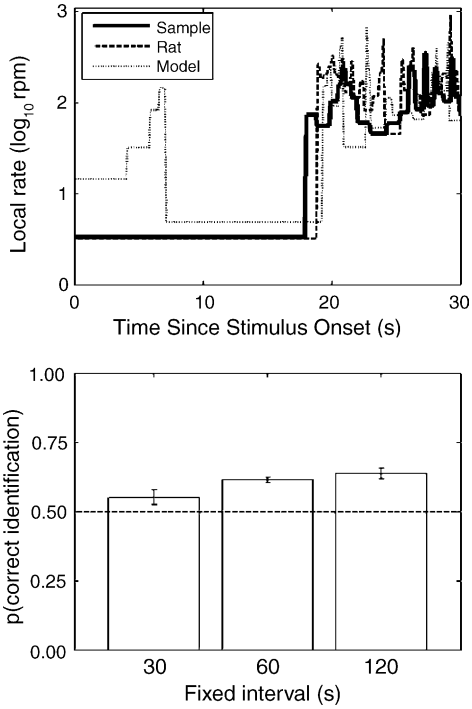


Fig. 6. Application of a Turing test to a time discrimination procedure. Top panel: local response rate as a function of time since stimulus onset is shown for a single cycle for a rat, a model, and an unknown sample. Bottom panel: the probability of correct detection and the standard error is shown for the three fixed-interval conditions.

required to reproduce the times of each response; other measures that are also sufficient to reproduce the times of each response include the cumulative number of responses on each cycle. The median absolute differences provide an objective measure of similarity; the sum of squared differences (least square criterion) provides an alternative measure. Thus, other indices may be used for classification of cycles.

The probability of a correct identification of each cycle, averaged over many cycles, provides a graded metric for the evaluation of a quantitative theory of timing. The value of the difference or ratio between two measures, such as the median absolute differences in Eq. (8), provides an alternative, and potentially more sensitive, graded metric.

A graded metric provides rapid feedback regarding the consequences of modifications of the theory. This can be used for the identification of critical features of the theory, and it can be used for parameter estimation.

The estimates of the parameters used in this article were based on informal searches of the parameter space based on summary statistics, such as those shown in Fig. 5. Standard iterative methods for fitting nonlinear equations may be used to improve the estimates of the parameters. The cycle-by-cycle data can be used, but to avoid overfitting, the same cycles should not be used for estimating parameters and evaluating the goodness of fit (Brown, 2000; Stone, 1974).

Of course, goodness of fit is not the only basis for evaluation of a quantitative model (Myung, 2000; Myung and Pitt, 2002; Pitt et al., 2002). Some authors find goodness of fit not to provide much support for a model (Roberts and Pashler, 2000), while others consider it to be a necessary but not sufficient criterion (Rodgers and Rowe, 2002). Probably most investigators would want a quantitative model that fit the data, but would also seek other criteria such as generality, simplicity, clarity, successful generation of new hypotheses that are subsequently verified, and facilitation of understanding (Myung and Pitt, 2002).

A Turing test provides a way to obtain input generality by the use of the same theory to account for the results of many different procedures; it also provides a way to obtain output generality by the use of the same theory to account for the results of many different summary measures of behavior.

Acknowledgements

National Institute of Mental Health Grant MH44234 to Brown University supported this research.

Although his name is not mentioned in the body of this article, Don Blough's influence on my research, including the research described in this article, has been profound. We have been colleagues in the Psychology Department at Brown University for 48 years, and I have considered him to be an academic brother. For several years, we co-taught a graduate seminar in quantitative models in Psychology in which I learned a great deal from him about psychophysical methods, signal detection theory, multidimensional scaling, and models of reaction time. (I taught sections on conditioning and timing models.) With an original classic LINC computer that he obtained in 1964, he introduced me to the modern world of online computers which he has continued to use skillfully and creatively. The research problems on which we have worked are quite different, but I have been heavily influenced by methods he has used. I have been impressed by his emphasis on the behavior of individual animals, the psychophysical approach to research, within-subject experimental designs, clear specifications of input and output variables with beautiful functional relationships between them that warrant careful study, and his elegant and simple quantitative models of data that direct attention to the underlying processes. Many of the important influences are typically uncited, and frequently not conscious. I am as impressed now as I was when I first learned about his identification of response bouts of pigeons with different functional relations (Blough, 1963), his original demonstration of selection of the least-frequent reinforcement times (Blough, 1966), and his creative development of a quantitative model of operant generalization and discrimination (Blough, 1975). (RMC)

Appendix A

A.1. A notation for timing and conditioning procedures

This procedural notation, referred to as the Brown Notation System, provides an hierarchical structure in which the primary elements are times of events. These elements are combined into cycles, which are combined into treatments, sessions, phases, and experiments.

A.2. Elements

A.2.1. Stimulus

A stimulus is determined by the procedure, or by the procedure and the behavior of the animal. Capital letters are used for onset; lower case letters are used for termination; both are used for a stimulus pulse. For example,

Stimulus	On	Off	Pulse
Houselight	H	h	Hh
Noise	N	n	Nn
Clicker	C	c	Cc
Left lever	V	v	Vv
Right lever	W	w	Ww

A.2.2. Response

A response is a measured behavior of the animal. For example,

Response	On	Off	Pulse	Wingding3 font
Head	→	←	↔	g, f, and n
Left lever	↓	↑	↕	I, h, and o
Right lever	↓↓	↑↑	↓↑	K, J, and E
Lick				R

Note that special characters are available in Wingding3 font. For example, a right-pointing arrow is produced by the lower-case g.

A.2.3. Reinforcer

A reinforcer is a stimulus with certain conventionally recognized special properties.

Reinforcer	On	Off	Pulse	
Food	▲	▼	▲▼	*p*, *q*, and *pq*
Shock				*y*, *x*, and *yx*

A.2.4. Event

An event is the onset and/or termination of a stimulus, response, or reinforcer.

A.2.5. State

A state is the pattern of presence or absence of stimuli, responses, and reinforcers.

A.2.6. Procedural change

A procedural change is a change in the contingencies of reinforcement that are not accompanied by any external event (stimulus, response, or reinforcer).

Procedural change	On	Off	Pulse	
Prime	△	▽	△▽	*r*, *s*, and *r s*

A.2.7. Time

A time is a number and unit relative to the onset of a cycle,

20 s	Example of fixed time of 20 s
2 m	For 2 min
2 h	For 2 h
∼20 s	Example of random time (random sample from exponential distribution with a mean of 20 s).
u (1:1:20 s)	Example of uniformly distributed time (random sample from uniform distribution with a minimum of 1 s, a maximum of 20 s at 1-s intervals.

A.3. Combinations of elements

A.3.1. Cycle

A cycle is composed of one or more states, with a specification of the times and events that occur within each state. A cycle is begun and ended with a backslash; a semicolon is used to separate different states within a cycle; times and events within parentheses are in effect when the state is in effect.

C_1 = /20 s ↓▲▼/	%An example of a cycle of a fixed-interval schedule of reinforcement in which food is delivered following the first left lever response after a 20-s interval.
C_2 = /120 s N; 30 s ▲▼ n/	%An example of a cycle of a cycle consisting of a 120-s interval followed by the onset of noise. After 30-s, food is delivered and the noise is terminated. Note that a cycle may also be specified by several times and events, separated by a semicolon (;).
C_3 = /120 s N; 30 s (∼60 s ▲▼) n/	%This differs from C_2 only by delivering food at random times (with a mean of 60 s) during the 30 s that the light is off. There may be 0, 1, or more than 1 delivery of food during this 30-s interval.

A.3.2. Treatment

A treatment is one or more cycles that may be repeated. Five types are:

Simple. $\mathbf{T} = [C_1]$	%The same cycle type may be repeated.
Sequential. $\mathbf{T} = [C_1 C_2 \cdots C_3]$	%Several different types of cycles may be repeated in the same order.
Probabilistic. $\mathbf{T} = [p_1 C_1 p_2 C_2 \cdots p_n C_n]$	% A random sample of a single type of cycle may be repeated. (This is random sampling with replacement.)
Permutation. $\mathbf{T} = p[C_1 C_2 \cdots C_n]$ A random order of several types of cycles may be repeated. (This is random sampling without replacement).	%Several types of cycles may be executed simultaneously and independently.
Concurrent. $\mathbf{T} = [C_1 \& C_2 \cdots \& C_n]$	

A.3.3. Session

A session is a specification of the treatment between the time the animal enters and leaves the box. It may be specified either in terms of time, number of cycles, or number of treatments or a criterion of performance. For example,

S = 120 m **T**	%A sessions consists of 120 min of a treatment
S = 30 **C**	%A session consists of 30 repetition of a cycle
S = 10 **T**	%A session consists of 10 repetitions of a treatment

A.3.4. Phase

A phase refers the series of sessions with the same treatment. It may be defined either in terms of the number of sessions or some criterion of performance. For example,

P = 20 **S**	%A phase consists of 20 sessions

A.3.5. Experiment

An experiment is a specification of the number of phases.

E = 30 **S**	An experiment consist of 30 sessions

A.4. Examples of notation for timing and conditioning procedures

A.4.1. The present experiment

C_1 = /20 s H; 30 s → h▲▼/	%Cycle with houselight, 30-s FI
C_2 = /20 s N; 60 s → n▲▼/	%Cycle with noise, 60-s FI
C_3 = /20 s C; 120 s → c▲▼/	%Cycle with clicker, 120-s FI
T = [(1/3)C_1 (1/3)C_2 (1/3)C_3]	%Treatment consists of random selection of cycle
S = 60 **T**	%60 cycles per session
P = 30 **S**	%30 session in phase
E = **P**	%One phase in experiment

A.4.2. A temporal discrimination procedure

This was the procedure used for pretraining and training by Church and Deluty (1977)

C_1 = /60 s ▲▼/	%Food delivered every 60 s
C_2 = /V 10(↓▲▼)v; W 10(↓▲▼)/	%Left lever inserted, each of 10 left lever responses are followed by food; left lever is withdrawn; same for right lever
C_3 = /30 s l; 2 s L; (VW ↓▲▼vw or ↓↓vw)/	%If light is off for 2 s, left lever followed by food
C_4 = /30 l = 8 s L; (VW ↓vw or ↓↓▲▼vw)/	%If light is off for 8 s, right lever followed by food
$\mathbf{T}_1 = C_1$	%Treatment was magazine training
$\mathbf{S}_1$ = 60 $\mathbf{T}_1$	%Session was 60 treatments
$\mathbf{P}_1 = \mathbf{S}_1$	%Phase 1 was one session
$\mathbf{T}_2 = \mathbf{C}_2$	%Treatment was lever training
$\mathbf{S}_2$ = 2 $\mathbf{T}_2$	%Session was 2 treatments
$\mathbf{P}_2$ = 2 $\mathbf{S}_2$	%Phase 2 was 2 sessions
$\mathbf{T}_3$ = [.5$\mathbf{C}_3$ .5$\mathbf{C}_4$]	%Treatment is probabilistic
$\mathbf{S}_3$ = 50 m $\mathbf{T}_3$	%Sessions were 50 min
$\mathbf{P}_3$ = 20 $\mathbf{S}_3$	%Phase 3 was 20 sessions
E = [$\mathbf{P}_1$ $\mathbf{P}_2$ $\mathbf{P}_3$]	%Experiment was phases 1, 2, and 3

References

Blough, D.S., 1963. Interresponse time as a function of continuous variables: a new method and some data. J. Exp. Anal. Behav. 6, 237–246.

Blough, D.S., 1975. Steady state data and a quantitative model of operant generalization and discrimination. J. Exp. Psychol. Anim. Behav. Process. 1, 3–21.

Blough, D.S., 1966. The reinforcement of least-frequent interresponse times. J. Exp. Anal. Behav. 9, 581–591.

Brown, M.W., 2000. Cross-validation methods. J. Math. Psychol. 44, 108–132.

Church, R.M., 2001. A Turing test for computational and associative theories of learning. Curr. Dir. Psychol. Sci. 10, 132–136.

Church, R.M., 1997. Quantitative models of animal learning and cognition. J. Exp. Psychol. Anim. Behav. Process. 23, 379–389.

Church, R.M., 2002. Temporal learning. In: Gallistel, R., Pashler, H. (Eds.), Stevens' Handbook of Experimental Psychology: Learning, Motivation, and Emotion, 3. Wiley, New York, pp. 365–393.

Church, R.M., Broadbent, H.A., 1990. Alternative representations of time, number, and rate. Cognition 37, 55–81.

Church, R.M., Deluty, M.Z., 1977. Bisection of temporal intervals. J. Exp. Psychol. Anim. Behav. Process. 3, 216–228.

Dews, P.B., 1970. The theory of fixed-interval responding. In: Schoenfeld, W.N. (Ed.), The Theory of Reinforcement Schedules. Appleton-Century-Crofts, New York, pp. 43–61.

Gibbon, J., 1977. Scalar expectancy theory and Webers law in animal timing. Psychol. Rev. 84, 279–325.

Grossberg, S., Schmajuk, N.A., 1991. Neural dynamics of adaptive timing and temporal discrimination during associative learning. In: Grossberg, S., Carpenter, G.A. (Eds.), Pattern Recognition by Self-Organizing Neural Networks. The MIT Press, Cambridge, MA, pp. 637–674.

Guilhardi, P., Church, R.M. Dynamics of temporal discrimination, submitted for publication.

Guilhardi, P., Church, R.M., 2004. Measures of temporal discrimination in fixed-interval performance: a case study in archiving data. Behavior RMIC 36, 661–669.

Guilhardi, P., Keen, R., MacInnis, M.L.M., Church, R.M. The combination rule for multiple intervals. Behav. Process., in press.

Killeen, P.R., Fetterman, G., 1988. A behavioral theory of timing. Psychol. Rev. 95, 274–295.

Kirkpatrick, K., 2002. Packet theory of conditioning and timing. Behav. Process. 57, 89–106.

Kirkpatrick, K., Church, R.M., 2003. Tracking of the expected time to reinforcement in temporal conditioning procedures. Learn. Behav. 31, 3–21.

Kurtzman, H.S., Church, R.M., Crystal, J.D., 2002. Data archiving for animal cognition research: report of an NIMH workshop. Anim. Learn. Behav. 30, 405–412.

Machado, A., 1997. Learning the temporal dynamics of behavior. Psychol. Rev. 104, 241–265.

Myung, I.J., 2000. The importance of complexity in model selection. J. Math. Psychol. 44, 190–204.

Myung, I.J., Pitt, M.A., 2002. Mathematical modelling. In: Pashler, H., Velicer, W. (Eds.), Stevens Handbook of Experimental Psychology: Methodology in Experimental psychology, 4. Wiley, New York, pp. 429–460.

Pitt, M.A., Myung, I.J., Zhang, S., 2002. Toward a method of selecting among computational models of cognition. Psychol. Rev. 109, 472–491.

Roberts, S., Pashler, H., 2000. How persuasive is a good fit? A comment on theory testing. Psychol. Rev. 107, 358–367.

Rodgers, J.L., Rowe, D.C., 2000. Theory development should begin (but not end) with good empirical fits: a comment on Roberts and Pashler (2000). Psychol. Rev. 109, 472–491.

Staddon, J.E.R., Higa, J.J., 1999. Time and memory: toward a pacemaker-free theory of interval timing. J. Exp. Anal. Behav. 71, 215–251.

Stone, M., 1974. Cross-validatory choice and assessment of statistical predictions. J. R. Stat. Soc. 36, 111–147, Series B (Methodological).

Turing, A.M., 1950. Computing machinery and intelligence. Mind 59, 433–460.