



## 3 Conditioning from an information processing perspective

4 C.R. Gallistel\*

5 *Rutgers Center for Cognitive Science (RuCCS), Rutgers University at New Brunswick, Busch Campus,*  
6 *152 Frelinghuysen Road, Piscataway, NJ 08854-8020, USA*

### 7 Abstract

8 The framework provided by Claude Shannon's [Bell Syst. Technol. J. 27 (1948) 623] theory of information leads to a  
9 quantitatively oriented reconceptualization of the processes that mediate conditioning. The focus shifts from processes set in  
10 motion by individual events to processes sensitive to the information carried by the flow of events. The conception of what  
11 properties of the conditioned and unconditioned stimuli are important shifts from the tangible properties to the intangible  
12 properties of number, duration, frequency and contingency. In this view, a stimulus becomes a CS if its onset substantially  
13 reduces the subject's uncertainty about the time of occurrence of the next US. One way to represent the subject's knowledge  
14 of that time of occurrence is by the cumulative probability function, which has two limiting forms: (1) The state of maximal  
15 uncertainty (minimal knowledge) is represented by the inverse exponential function for the random rate condition, in which the  
16 US is equally likely at any moment. (2) The limit to the subject's attainable certainty is represented by the cumulative normal  
17 function, whose momentary expectation is the CS–US latency minus the time elapsed since CS onset. Its standard deviation is  
18 the Weber fraction times the CS–US latency.

19 © 2003 Published by Elsevier Science B.V.

20 *Keywords:* Information; Learning; Models; Conditioning; Randomness

### 21 1. Introduction

22 The late 1960s saw the publication of three exper-  
23 imental papers that demanded far reaching revision  
24 in our conception of conditioning. Rescorla (1968)  
25 showed that if one removed the contingency between  
26 a to-be-conditioned stimulus (CS) and the uncondi-  
27 tioned stimulus (US) without altering the temporal  
28 pairing, the subject did not develop a conditioned re-  
29 sponse to the CS (Fig. 1). This result implied that it  
30 was contingency, not temporal pairing, that generated  
31 conditioned responding.

32 Kamin (1969) showed that pairing a CS and a US  
33 did not produce a conditioned response to the CS if it

34 was always presented together with another, already  
35 conditioned CS (Fig. 2). The already conditioned CS  
36 blocked conditioning to the newly introduced CS.  
37 This implied that when a subject "expected" the US,  
38 that is when it was not "surprised" by it, then pairing  
39 it with a CS did not produce a conditioned response.  
40 I have followed Kamin in putting scare quotes around  
41 "surprise." He did so presumably to warn that it was  
42 unclear what the scientific meaning of the word might  
43 be. Also, as a behaviorist, he was chary of the notion  
44 of expectation, without which there can be no surprise.

45 Kamin (1969) further showed that when two CSs  
46 were presented together from the outset of condition-  
47 ing, the conditioned response that developed to one  
48 of them was much stronger than the conditioned re-  
49 sponse that developed to the other. Kamin called this  
50 "overshadowing." A striking example of it had been

\* Tel.: +1-732-445-2973.

E-mail address: [galliste@ruccs.rutgers.edu](mailto:galliste@ruccs.rutgers.edu) (C.R. Gallistel).

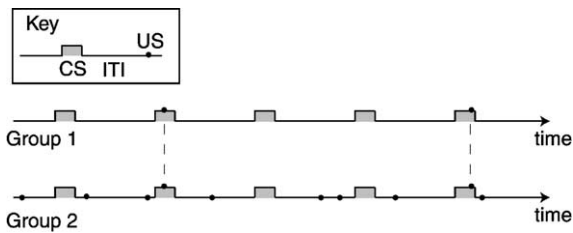


Fig. 1. Rescorla's (1968) truly random control experiment. The temporal pairing of USs with CSs is the same in the two groups (dashed vertical lines). In Group 1, there is a CS–US contingency, because the US occurs only when the CS is on. In the second group, there is no CS–US contingency; the CS is on one-fifth of the time and one-fifth of the USs occur when it is on. Subjects in the first group develop a conditioned response to the CS; subjects in the second group do not. CS: conditioned stimulus; US: unconditioned stimulus; ITI: intertrial interval.

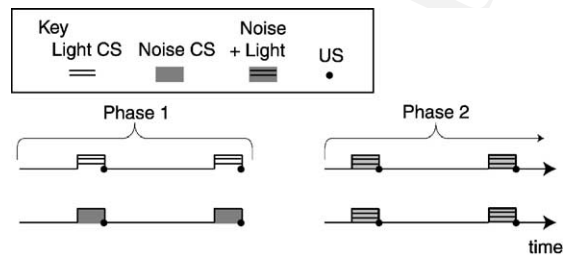


Fig. 2. Kamin's (1969) blocking protocol. Subjects are conditioned with one CS in Phase 1. In Phase 2, a second CS is given in compound with (together with) the already conditioned CS. Subjects do not develop a conditioned response to the second CS, despite its repeated pairing with the US.

published some years before by Reynolds (1961) (see 51  
 Fig. 3). He interpreted it in terms of selective attention, an interpretation that has continued to be popular 52  
 (Mackintosh, 1975; Pearce and Hall, 1980). These experiments showed that when two stimuli are redun- 53  
 54

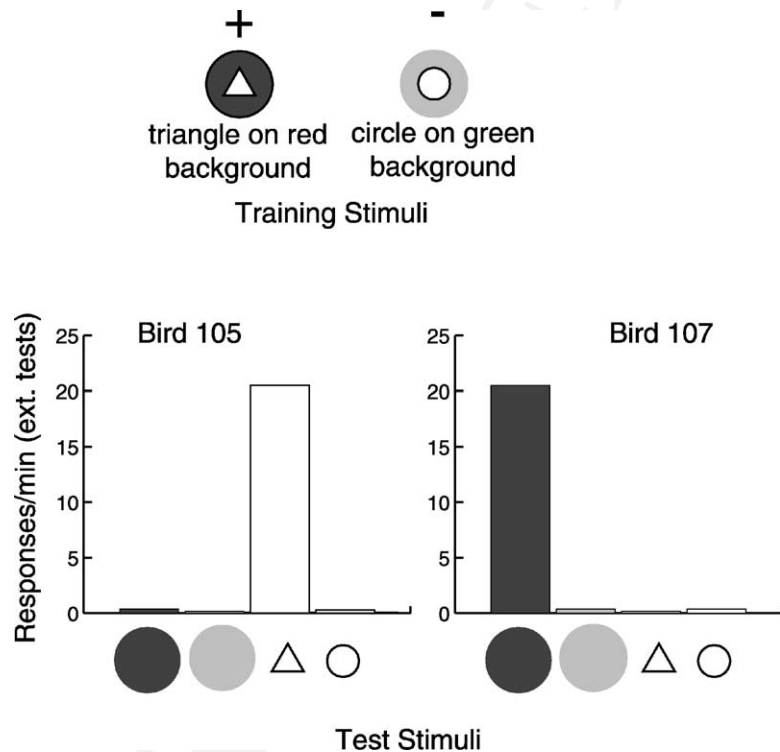


Fig. 3. Overshadowing. Reynolds (1961) trained two pigeons to peck the red key with a white triangle for food reward. Pecking the green key with the white circle did not yield reward. When he tested each of the four stimulus components in isolation, he found that one bird was conditioned to the white triangle, while the other was conditioned to the red background. Each bird responded to one of the two elements composing the positive stimulus to the exclusion of the other element.

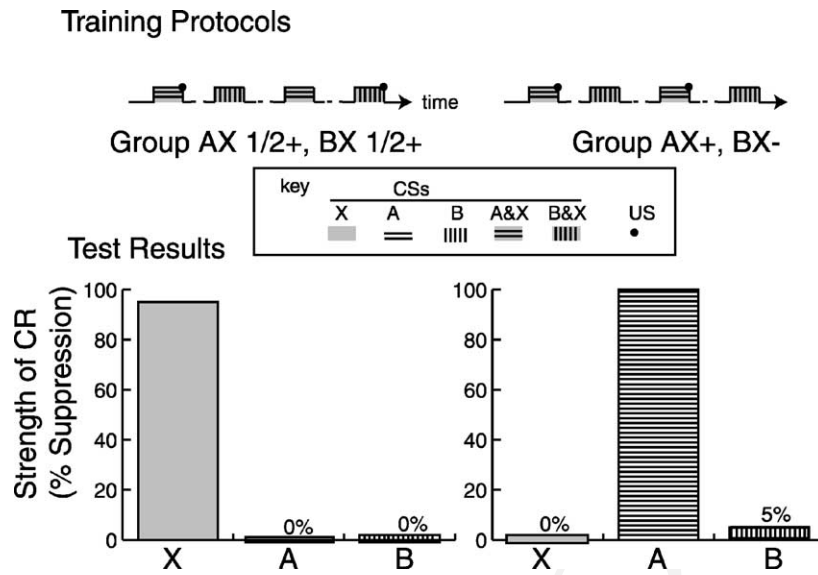


Fig. 4. The effect of relative validity. For both groups of rats, there are an infinity of rates of US occurrence that could be ascribed to the different CSs (X, A, and B) to give rates indistinguishable from the experimentally programmed rates, under the assumption that the rates predicted by the different CSs are additive. However, in both groups, subjects settle on (show a conditioned response to) the only CS that by itself carries all the available information about US timing (X in one case, A in the other). Because it alone can carry all the information and because it is the only CS that can do so, it is called the relatively more valid predictor. Data are from Table 3 on p. 175 of Wagner et al. (1968).

55 dant predictors, the conditioning process eliminates  
 56 one of them from consideration. Also, they underscore  
 57 Rescorla's conclusion that temporal pairing is not sufficient  
 58 for conditioning.

59 Wagner et al. (1968) showed that when one CS  
 60 is more reliably informative of US delivery, it is the  
 61 only one to which the conditioned response develops  
 62 (Fig. 4). It does not develop to the other less reliably in-  
 63 formative CSs, even though they are frequently paired  
 64 with the US. Once again, this showed that informa-  
 65 tional considerations determine conditioning, not pro-  
 66 cesses set in motion by temporal pairing.

67 When the history of this period is written, it will  
 68 be a challenge to explain why psychologists were sur-  
 69 prised to discover that temporally pairing a CS and  
 70 US was not sufficient to produce conditioned behav-  
 71 ior because it had long been known that this was not  
 72 necessary. When a CS predicts the omission of an oth-  
 73 erwise to be expected US, a conditioned response de-  
 74 velops just as readily as it does when a CS predicts  
 75 a not otherwise to be expected US (Fig. 5). This is  
 76 called inhibitory conditioning, because a CS that pre-  
 77 dicts the omission of the US suppresses (inhibits) the  
 conditioned response to a CS that predicts the US. The

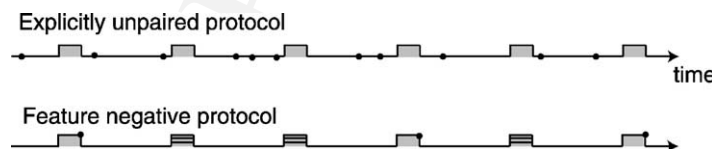


Fig. 5. Two protocols for producing conditioned inhibition. In both, the US is never paired with the CS. Nonetheless, a conditioned response develops to the inhibitory CS (CS-), which is represented by gray in the first protocol and by the horizontal bars in the second protocol. In the second protocol, the inhibitory CS is compounded with (presented together with) the gray CS on some trials. The US occurs only on those trials when this second CS is not present.

78 effects of inhibitory conditioning are not limited to an-  
79 tagonizing the effects of excitatory conditionings. In-  
80 hibitory conditioning produces conditioned responses  
81 of its own (Hearst and Jenkins, 1974; Wasserman et al.,  
82 1974). During inhibitory conditioning, the CS and US  
83 are never paired, so this phenomenon demonstrates  
84 that CS–US pairing is not necessary for conditioning.

## 85 2. Contingency not temporal pairing

86 In summary, by 1970, it was known that temporal  
87 pairing is neither necessary nor sufficient to produce  
88 conditioned responding. What is necessary and, ar-  
89 guably, sufficient is CS–US contingency, the CS must  
90 provide information about the US. The importance of  
91 the experimental findings just reviewed was immedi-  
92 ately recognized; they have been widely replicated;  
93 and they are recounted in most textbooks on condi-  
94 tioning and learning. Nonetheless, the conviction that  
95 temporal pairing is the key requirement in associative  
96 learning persists undiminished (Gluck and Thompson,  
97 1987; Hawkins and Kandel, 1984; Krasne, 2002;  
98 Miller and Escobar, 2002; Usherwood, 1993). Why?

99 Rescorla (1972, p. 10) put his finger on the problem  
100 that has blocked our progress in coming to terms with  
101 the implications of these experiments, when he wrote,  
102 “We provide the animal with individual events, not  
103 correlations or information, and an adequate theory  
104 must detail how these events individually affect the  
105 animal. That is to say that we need a theory based on  
106 individual events.”

107 He could not, on reflection, have meant to say that  
108 we do not provide the animal with correlations and in-  
109 formation, because when we construct our experimen-  
110 tal protocols we manipulate the correlations between  
111 the various CSs and the US so as to vary the infor-  
112 mation that the CSs provide about the US. What he  
113 presumably meant was that we take it for granted that  
114 the conditioning process does not operate at this level  
115 of abstraction, because we take it for granted that it in-  
116 volves the creation or modification of associative con-  
117 nections. Events—and more particularly the temporal  
118 and spatial convergence of event-triggered signals in  
119 the nervous system—are what increment associations,  
120 not information.

121 For most psychologists and neurobiologists, infor-  
122 mation does not seem substantial enough to do the job

of making connections. Information must be extracted 123  
from the flow of events by computational processes 124  
that are driven by the intervals between events and the 125  
numbers of events—rather than by their more tangible 126  
properties, the properties that excite sensory receptors 127  
and are therefore the natural focus of empiricist theo- 128  
ries of mind and brain. 129

## 3. Information drives conditioning 130

From an information processing perspective, how- 131  
ever, it is the information in the protocol that drives the 132  
conditioning process. The assumption that condition- 133  
ing involves the forging of connections is a conceptual 134  
roadblock. It prevents our coming to terms with the 135  
implications of these experiments and the many other 136  
experimental results which imply that conditioned be- 137  
havior is a manifestation of the brain’s information 138  
processing activity. 139

About the time he published his paper on blocking 140  
and overshadowing, Leon Kamin moved from Mac- 141  
master to Princeton. Fifty miles to the north of him 142  
there, Claude Shannon worked in the Bell Laborato- 143  
ries, where, 20 years earlier, he published a famous pa- 144  
per that could have enabled Kamin to remove the scare 145  
quotes from around surprise (Shannon, 1948). In it, he 146  
showed how to define information in such a way as to 147  
make it quantifiable. This made it a respectable scien- 148  
tific concept. It also laid a foundation for the modern 149  
communications industry, for computer science, and 150  
for the modern understanding of thermodynamics. It 151  
should, I argue, also become a foundation for our un- 152  
derstanding of the brain (cf. Rieke et al., 1997) and the 153  
process of conditioning (Gallistel and Gibbon, 2002). 154

In Shannon’s definition, a signal conveys informa- 155  
tion to the extent that it reduces the receiver’s uncer- 156  
tainty about the state of the world. This leads directly 157  
to a rigorous quantification of information. It also ac- 158  
cords with two everyday intuitions: (1) The more infor- 159  
mation we have, the less uncertain we are. (2) Some- 160  
thing that tells us what we already know does not con- 161  
vey information to us. 162

This definition introduces what, for some, is a dis- 163  
comfiting subjectivity into the foundations of the sci- 164  
ence. The information conveyed by a signal cannot 165  
be measured in the absence of knowledge about the 166  
receiver’s (i.e. the subject’s) state of uncertainty be- 167

168 fore and after the signal. To apply Shannon's insight  
 169 to our understanding of conditioning, we have to con-  
 170 sider what it is about the world that the subject rep-  
 171 represents (knows), what the limits are on the precision  
 172 with which the subject knows it, and how those lim-  
 173 its are captured in the subject's representation, that is,  
 174 how the subject represents its uncertainty about the  
 175 state of the world.

#### 176 4. Some simple principles

177 In a first pass, the application of Shannon's defini-  
 178 tion to conditioning might adopt the following princi-  
 179 ples:

- 180 (1) Only informative CSs elicit conditioned re-  
 181 sponses.
- 182 (2) A CS informs a subject to the extent that it reduces  
 183 the subject's uncertainty about the timing of the  
 184 next US.
- 185 (3) The brain's information processing is efficient; it  
 186 minimizes the number of signals (channels) re-  
 187 quired to carry the available information about the  
 188 timing of the next US.
- 189 (4) The information provided by a CS is not a prop-  
 190 erty of the CS itself but rather of its temporal and  
 191 numerical distribution relative to the US. Thus,  
 192 the brain needs timers and counters in order to  
 193 extract from an experimental protocol the infor-  
 194 mation that the CS carries about the timing of the  
 195 US. This is the level of abstraction at which the  
 196 conditioning process operates.
- 197 (5) The uncertainty in the representation of temporal  
 198 and numerical magnitudes is proportional to their  
 199 magnitude (Weber's law). This establishes an up-  
 200 per limit on attainable certainty.
- 201 (6) Poisson (random rate) processes provide the other  
 202 limit, the upper limit on uncertainty. In a random  
 203 rate process, the US is equally likely at every mo-  
 204 ment.

#### 205 5. A simply constructed function

206 The cumulative probability function is a felicitous,  
 207 computationally simple way for a subject to represent  
 208 what it knows about the timing of the next US (Fig. 6).

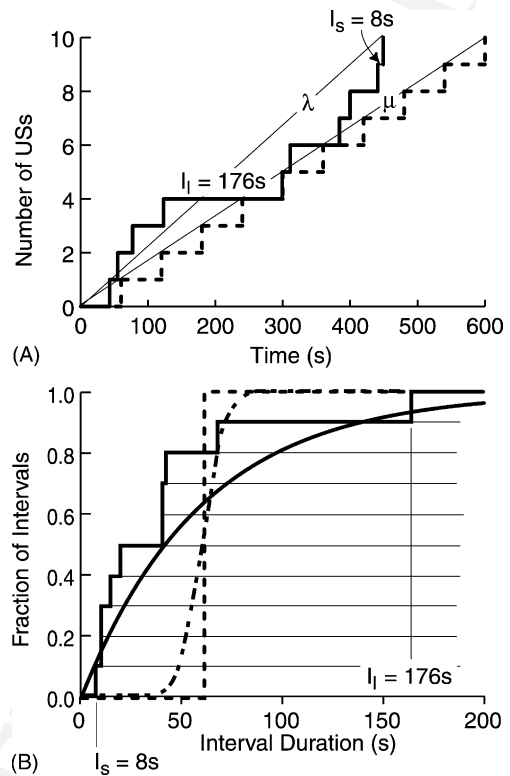


Fig. 6. (A) Two cumulative records. The dashed line is the cumulative record of the first 10 events generated by a fixed time schedule, the schedule that makes the next US maximally predictable. The solid line is from a random rate schedule with the same expectation. This schedule makes the next US maximally unpredictable. (B) Empirical and theoretical cumulative distribution functions. The dashed line with a single large step at 60 s is the (normalized) empirical cumulative distribution for the fixed time schedule. The solid black line with many steps is for the random rate schedule. In both cases, each event produces a unit step in the cumulative event count. The steps differ only in the length of the interval before the step, hence in where they occur when they are stacked. Sorting the steps on the basis of the length of the interval preceding the step and stacking the sorted steps produces the cumulative distribution function. The locations of the shortest ( $I_s$ ) and longest ( $I_l$ ) intervals in the random rate record are indicated. The solid black curve is the inverse exponential function ( $1 - e^{-\lambda t}$ ). The empirical cumulative distribution function for the random rate schedule approaches this curve in the limit, as the number of events becomes arbitrarily large. The dashed curve is the limit on the representation of a step function imposed by the scalar variability (the Weber-law characteristic) in the brain's representation of magnitudes. This representational imprecision, which is proportional to the magnitude represented, turns step distributions into cumulative Gaussian distributions.

209 It is the rescaled (normalized) cumulative distribution  
 210 of experienced interevent intervals. The cumulative  
 211 distribution of the intervals is obtained from sorting  
 212 them by duration, then counting, as one increases the  
 213 given duration, the number of intervals having that du-  
 214 ration or less. Graphically, this is equivalent to stack-  
 215 ing the sorted intervals (Fig. 6B). Dividing this cu-  
 216 mulative distribution function by the total number of  
 217 intervals observed (i.e. normalizing or rescaling it to  
 218 make it asymptotically equal to 1) gives the cumula-  
 219 tive probability function.

220 The simplest case to consider is when the US func-  
 221 tions as its own CS, that is, when the timing of the  
 222 previous US provides all the information that the sub-  
 223 ject has about when to expect the next US. In a fixed  
 224 time schedule, the next US is delivered at a fixed in-  
 225 terval after the previous one. In such a schedule, each  
 226 US is fully informative about when the next US will  
 227 occur. The subject's uncertainty is limited not by vari-  
 228 ability in the world but by the precision with which  
 229 it can represent the world—in this case, by the preci-  
 230 sion with which it can represent the US–US interval.  
 231 In a random rate schedule, the next occurrence of the  
 232 US is equally likely at any moment after the previous  
 233 occurrence. In such a schedule, one US is completely  
 234 uninformative about the timing of the next US. The  
 235 expected interval is, however, the same in both cases.  
 236 What differs is not how long it will be on average un-  
 237 til the next US is experienced, but rather how much  
 238 uncertainty there is about when exactly the next US  
 239 will occur.

240 We know that subjects distinguish between these  
 241 two cases. When the timing of the next US is pre-  
 242 dictable, the conditioned response anticipates it  
 243 (LaBarbera and Church, 1974). The question from  
 244 an information processing perspective is, What rep-  
 245 resentation of its experience and what processing of  
 246 that representation would be required for the subject  
 247 to distinguish these cases?

248 Fig. 6A shows the cumulative record from the case  
 249 in which the US occurs at random (solid line) and the  
 250 case in which it occurs at a fixed interval (dashed line).  
 251 The expectations of the generating processes were 60 s  
 252 in both cases. The slopes ( $\lambda$  and  $\mu$ ) of the cumula-  
 253 tive records estimate those expectations. The slopes of  
 254 the lines drawn from the origins of these cumulative  
 255 records to points somewhere on them differ somewhat  
 256 depending on where in the record a line terminates, be-

257 cause these slopes are estimates of the expectations not  
 258 the expectations themselves. On average, the longer  
 259 the cumulative records, that is, the more events they  
 260 record, the less their slopes differ from one another  
 261 and from the expectations of which they are estimates.  
 262 Put another way, the longer the subject's experience  
 263 with the event-generating process is, the more accu-  
 264 rate its estimate of its expectation will tend to be.

265 Fig. 6B shows the construction of the cumulative  
 266 distribution functions. The inverse exponential curve  
 267 in Fig. 6B is the limit to which the cumulative distri-  
 268 bution function for the random rate process converges  
 269 as the number of intervals in the stack becomes ar-  
 270 bitrarily large. The curve is entirely specified by the  
 271 expectation of the random rate process, that is, by the  
 272 quantity estimated by  $\lambda$ .

273 The cumulative distribution function for the fixed  
 274 time process would be a step function if the inter-  
 275 vals between events were always exactly the same  
 276 and if the subject could represent them with infinite  
 277 precision—neither of which is true. With the aid of  
 278 modern technology, however, it is easy for us to create  
 279 an event generating process (scheduling mechanism)  
 280 in which the variation in the intervals between USs  
 281 is much smaller than the variation in the subject's rep-  
 282 resentation of those intervals, which is roughly  $\pm 15\%$   
 283 in pigeons, rats and mice (Church et al., 1994; Gibbon  
 284 and Church, 1992; King et al., in press). That the vari-  
 285 ation in the subject's representation of an unvarying  
 286 interval is proportional to the duration of that inter-  
 287 val is a central assumption of Gibbon's (1977) Scalar  
 288 Expectancy Theory. It is justified by extensive exper-  
 289 imental findings in the animal timing literature (see  
 290 Gallistel and Gibbon, 2000, for review). The cumu-  
 291 lative Gaussian curve in Fig. 6B (dashed sigmoid) is  
 292 the representation of a fixed interval schedule that re-  
 293 sults from scalar variability in the representing pro-  
 294 cess, with a Weber fraction (coefficient of variation)  
 295 of 0.15. It is the cumulative Gaussian distribution with  
 296 a standard deviation equal to 15% of its mean.

## 6. Acquisition of a timed CR 297

298 The subject can anticipate the next US only if it has  
 299 decided that it is more likely at some moments than  
 300 at others; in other words, only if it recognizes that  
 301 the timing is not random. The question that must be

302 answered in a theory about the acquisition of a timed  
 303 response is, How does the subject recognize that the  
 304 timing is not random?

305 The decision that it is not random can be based  
 306 on a simple comparison of the observed cumulative  
 307 distribution function to its exponential approximation.  
 308 The comparison need only note the maximum deviation  
 309 between the observed cumulative probability  
 310 function and its exponential approximation. This maximum  
 311 deviation is the Kolmogorov–Smirnov statistic.  
 312 The probability of observing a maximum deviation  
 313 of a given size is an a priori knowable (innately programmable)  
 314 function of the number of events. This gives an acquisition theory  
 315 for timed responding: the subject decides that the timing of the next  
 316 US can be predicted if the probability of observing that big a deviation  
 317 given that many events is less than some decision  
 318 criterion (Gallistel and Gibbon, 2002).  
 319

320 If a US predicts the time of occurrence of the next  
 321 US, then it constitutes a temporal landmark, by reference to which  
 322 the animal can locate itself in time.  
 323 When it has perceived the landmark, it can estimate  
 324 its progress through time toward the next occurrence  
 325 of the US, and it can use the cumulative distribution  
 326 function to set decision criteria (response thresholds)  
 327 that enable it to anticipate the next US with whatever  
 328 degree of certainty it wants. In Fig. 7, it has set its  
 329 start criterion so as to be 90% certain that it has begun  
 330 to respond before the next US occurs. At the graphed  
 331 moment, it has advanced past the point in time corresponding  
 332 to this criterion, and so it has begun to respond.  
 333 Its location in time corresponds approximately to the location  
 334 of the steepest portion of the cumulative distribution function,  
 335 so its expectation of the US is maximal at this location in time.  
 336

337 The past is (alas!) a fallible guide to the future, and  
 338 so the subject's expectation may be disappointed. It  
 339 must therefore have a stop criterion as well as a start  
 340 criterion; otherwise it would continue to respond indefinitely,  
 341 if and when the US failed to materialize. In Fig. 7,  
 342 it has set its stop criterion at a level such that it  
 343 stops responding when there is less than 1 chance in 20  
 344 that the anticipated US is going to happen within the  
 345 limits of its uncertainty regarding the anticipated time.  
 346 That is, it stops only when it becomes highly likely  
 347 that its expectation has been violated. This account  
 348 of the timing of a conditioned response is Gibbon's  
 349 (1977) Scalar Expectancy Theory for the peak proce-

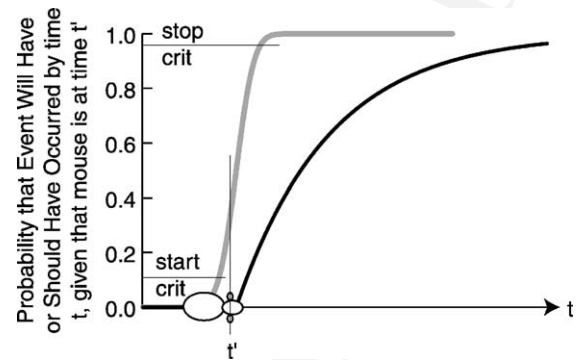


Fig. 7. Cumulative distribution functions as temporal frameworks. When the next US is fixed in time (gray sigmoidal curve), then the subject (the mouse at time  $t'$ ) can use its timer to measure its advance toward and past that temporal location, and it can set decision criteria that govern the timing of its anticipatory behavior (conditioned behavior). In this case, the cumulative probability function gives the probability that the event should have occurred. When the next US has not been temporally localized, the mouse remains always at the origin of the cumulative probability function (black inverse exponential). As the mouse advances through time, it never catches up with it. In this case, the cumulative probability function gives the probability that the event will have occurred (by such and such time in the future).

350 dure. Notice that in the account just rendered there  
 351 are two different decision processes, using two different  
 352 decision variables. One process (first elaborated by  
 353 Gallistel and Gibbon, 2002) explains the acquisition  
 354 of a timed response; the other (Gibbon's, 1977 Scalar  
 355 Expectancy Theory) explains the timing itself (when  
 356 the response starts and when it stops).

357 When, by contrast, the time of occurrence of a US  
 358 does not predict the time of occurrence of the next  
 359 US, then it does not serve as a landmark by reference  
 360 to which the subject may assess its approach to the  
 361 next US. In this case, the subject has only an ego-  
 362 centric temporal framework, a framework whose origin  
 363 is always at the subject's current location in time.  
 364 As it advances through time, it does not come closer  
 365 to the moment when the US may be expected to occur.  
 366 This is a counterintuitive property of random rate  
 367 processes: the expected time to the next event is independent  
 368 of how long it has been since the last event. That is why  
 369 the time between the onset of observation and the first  
 370 occurrence of an event gives an unbiased (but, of course,  
 371 noisy) estimate of the expectation of the process (the  
 372 interval between events). All of these are corollaries of  
 373 the fact that in a random rate pro-

374 cess, the time at which one event occurs provides no  
375 information about the time at which the next event will  
376 occur.

### 377 7. Detecting changes in the rate of 378 US occurrence

379 To say that the occurrence of an event in a random  
380 rate process provides no information about when the  
381 next event will occur is not to say that it provides no  
382 information about anything. It provides information  
383 about the value of  $\lambda$ . The objective uncertainty about  
384 the true value of  $\lambda$  is inversely proportional to the  
385 square root of the number of USs so far experienced.  
386 The more events, the less the uncertainty; hence the  
387 more information the subject has acquired from the  
388 stream of events. But this holds true only so long as  
389 the precision with which the subject can represent  $\lambda$ 's  
390 value is not the limiting factor. The decrease in the  
391 inherent or unavoidable uncertainty of a rate estimate  
392 as the square root of the number of events increases  
393 means that the first few events the subject observes  
394 convey substantial information about the value of  $\lambda$ ,  
395 but the amount conveyed by further events diminishes  
396 rapidly.

397 Of course, the rate may change. This violation of  
398 expectation cannot be detected by the just described  
399 process for detecting the failure of the US to occur  
400 at an expected time, because the defining feature  
401 of a random time schedule is that there is no expected  
402 time; the US is equally likely at every moment.  
403 This poses an interesting conceptual challenge  
404 to associative theories of extinction and conditioned  
405 inhibition. The challenge arises because these models  
406 are driven by events rather than by the information  
407 conveyed by a stream of events. In an associative  
408 theory, whether trial based (Rescorla and Wagner,  
409 1972) or real time (see also Brandon et al., 2002, this  
410 volume; Wagner, 1981), the mechanism of extinction  
411 is activated by the failure of an expected event to occur.  
412 This “event”—the failure—must itself have a time  
413 of occurrence. But when there is no expected time of  
414 occurrence for an expected event, how can its failure  
415 to occur have a time of occurrence?

416 In one information processing approach, changes  
417 in the rate of US occurrence are hypothesized to be  
418 detected by a simple ongoing computation (Gallistel

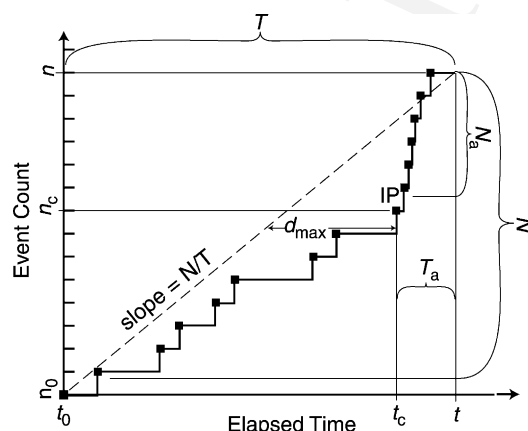


Fig. 8. Changing the rate of a random rate process creates an inflection point (IP) in the cumulative record, which will be at the maximum deviation ( $d_{\max}$ ) from the straight line approximation to the record (dashed line with slope  $N/T$ ). The creditworthiness of the hypothesis that a putative inflection point corresponds to a genuine change in rate is a simple calculation of binomial likelihood ratios. The calculation compares the proportion  $N_a/N$  to the proportion  $T_a/T$  to test whether they differ by more than is to be expected from chance fluctuations. The calculation is spelled out in Gallistel and Gibbon (2002).

and Gibbon, 2002, p. 100 ff). Despite its simplicity, it  
has been shown to be an ideal detector of such changes  
(Gallistel et al., 2001); it uses all the information available  
in the sequence of interevent intervals. What it does is look  
for seeming changes in the slope of the cumulative record  
and calculate how unlikely it is that a given change would  
have arisen through chance fluctuations (Fig. 8). Comparing  
the behavior predicted by this ideal detector to the behavior  
observed when rats adjust to frequent but otherwise random  
changes in the relative rates of reward in a Herrnstein  
matching paradigm show that the rat is an ideal detector of  
changes in random rates (Gallistel et al., 2001). It adjusts  
to them about as fast as is in principle possible. This finding  
poses the question whether it is possible for a model not  
based on information processing assumptions to predict changes  
that are as fast and sudden as the changes rats in fact make.

### 8. Explaining cue competition

The explanations of basic conditioning so far outlined  
all assume that the brain has built into it mech-

440 anisms that embody in their structure the statistics  
 441 of random rate processes and mechanisms that detect  
 442 systematic deviations from randomness. This assump-  
 443 tion constitutes a much stronger commitment to ratio-  
 444 nalism and nativism than empiricist learning theorists  
 445 have traditionally been comfortable with. However, to  
 446 the biologically inclined, at least, it seems no more  
 447 exceptional than saying that the eye embodies in its  
 448 structure the laws of optics. Why should not the struc-  
 449 ture of learning mechanisms reflect the structure of the  
 450 problems they solve just as much as does the structure  
 451 of sensory or metabolic organs (see Gallistel, 1999,  
 452 for an elaboration of this argument)?

453 Everything that has been developed above for the  
 454 univariate case—the case where the US serves as its  
 455 own predictor—applies with minimal modification to  
 456 the multivariate case, where the US is predicted by  
 457 other stimuli (conventionally called CSs). However,  
 458 when we allow other predictors, we confront the es-  
 459 sential problem in multivariate time series analysis—  
 460 determining what predicts what. The classic experi-  
 461 ments on the truly random control (or background con-  
 462 ditioning), blocking, overshadowing, and relative va-  
 463 lidity (reviewed in the introduction) show that subjects  
 464 do a sophisticated job of solving this problem. The as-  
 465 sumption that the information processing mechanisms  
 466 that mediate conditioned behavior are structured by  
 467 the properties of random rate processes again leads to  
 468 a simple model for the mechanism that solves what  
 469 is commonly called the cue competition problem, be-  
 470 cause random rates are additive.

471 Rate Estimation Theory (Gallistel, 1990, Chapter  
 472 13) assumes that in computing which CSs predict  
 473 which US rates, the brain treats all rates as random  
 474 rates. Under that assumption, when CSs are combined  
 475 (compounded), their effects on the rate of US occur-  
 476 rence are additive. Many of the phenomena of what is  
 477 known as cue competition follow in an obvious way  
 478 from this assumption. The effect of background rein-  
 479 forcements (the truly random control, see Fig. 1) is  
 480 readily understood. The rate of US occurrence that  
 481 must be ascribed to the background (the experimen-  
 482 tal chamber) to account for the USs observed dur-  
 483 ing the intertrial interval, when no other predictor is  
 484 present, is the same as the rate observed when the CS  
 485 is present. This latter rate must equal the sum of the  
 486 rate ascribed to the background and the rate ascribed  
 487 to the CS. That can only be true if the rate ascribed to

the CS is 0. That is why subjects in the truly random  
 condition do not develop a conditioned response to  
 the CS. The explanation of the blocking phenomenon  
 (Fig. 2) is essentially the same. No USs occur when  
 there is no CS, so the background must be ascribed  
 a rate of 0. The initially conditioned CS must be as-  
 cribed a rate sufficient to explain the USs that occur  
 when it is present by itself in Phase 1. The rate of US  
 occurrence in Phase 2, when CS 1 is joined by CS 2,  
 is no greater than when CS 1 had the field to itself.  
 Therefore, the rate to be ascribed to CS 2 must be 0.

The unique additive solution is not always so ob-  
 vious as it is in these simple cases. However, there  
 is a computationally simple, universal solution to the  
 problem of estimating the rates of US occurrence to  
 be to ascribed to each CS. The essence of the solution  
 is given by the matrix algebra formula for the vector  
 (list),  $\vec{\lambda}_t$ , of true rate estimates:

$$\vec{\lambda}_t = \mathbf{T}^{-1} \vec{\lambda}_r$$

where

$$\vec{\lambda}_r = \begin{pmatrix} \frac{N_1}{T_1} \\ \frac{N_2}{T_2} \\ \vdots \\ \frac{N_m}{T_m} \end{pmatrix}$$

and

$$\mathbf{T} = \begin{vmatrix} 1 & \frac{T_{1,2}}{T_1} & \cdots & \frac{T_{1,m}}{T_1} \\ \frac{T_{2,1}}{T_2} & 1 & \cdots & \frac{T_{2,m}}{T_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{T_{m,1}}{T_m} & \frac{T_{m,2}}{T_m} & \cdots & 1 \end{vmatrix}$$

with  $N_i$  is the number of reinforcements observed in  
 the presence of CS<sub>*i*</sub>,  $T_i$  the cumulative time that CS<sub>*i*</sub>  
 was (observed to be) present, and  $T_{i,j}$  the cumulative  
 time that CS<sub>*i*</sub> and CS<sub>*j*</sub> were (observed to be) jointly  
 present (without regard to whatever further CSs may  
 also have been concurrently present).

This formula assigns to each CS an estimate of the  
 rate of US occurrence that would be observed if that

519 CS were the sole determinant of  $\lambda$ , the rate of US oc- 567  
 520 currence. The formula fails to make this assignment 568  
 521 only in cases where there are redundant predictors, as 569  
 522 there are in the overshadowing and relative validity 570  
 523 protocols. In those cases, the additive solution to the 571  
 524 rate estimation is not unique. There are an infinity of 572  
 525 solutions consistent with the assumption of additivity. 573  
 526 When that happens, the determinant of the matrix is 574  
 527 zero. 575

528 When the determinant of the full matrix is zero, the 576  
 529 principle of signal parsimony comes into play. The rate 577  
 530 estimating mechanism eliminates possible predictors 578  
 531 so as to find the smallest set of predictors that can ac- 579  
 532 count for the observed rates of US occurrence. If CSs 580  
 533 are viewed as signals that convey information, then 581  
 534 minimizing the number of signals carrying the avail- 582  
 535 able information maximizes the information about the 583  
 536 timing of US occurrence conveyed per signal. In other 584  
 537 words, it maximizes the ratio between the informa- 585  
 538 tion conveyed and the number of signals conveying it. 586  
 539 The information conveyed per conveying signal may 587  
 540 be thought of as the average bandwidth of the signals 588  
 541 (CSs). Thus, this principle might also be called the 589  
 542 principle of bandwidth maximization. 590

543 The principle of signal parsimony comes into play 591  
 544 when the available information about the timing of the 592  
 545 US can be conveyed either by a combination of CSs 593  
 546 or by a single CS. This was true for both groups in 594  
 547 the Wagner, et al. relative validity experiment (Fig. 4), 595  
 548 because the X CS was never presented alone. The fact 596  
 549 that in the AX+, BX– group reinforcements always 597  
 550 occurred on AX trials and never on BX trials could 598  
 551 be explained by the assumption that X predicted the 599  
 552 observed rate of reinforcement on AX trials and that 600  
 553 the rate of reinforcement of B was equal and oppo- 601  
 554 site in sign to the rate of reinforcement of X. In other 602  
 555 words, the positive effect of X on the rate of rein- 603  
 556 forcement, which is seen on AX trials, is offset by 604  
 557 the negative effect of B on BX trials. This explana- 605  
 558 tion is unparsimonious, but it satisfies the additivity 606  
 559 constraint. 607

560 The just mentioned notion of a negative effect on 608  
 561 rate of reinforcement is an odd one on first encounter. 609  
 562 It arises from the fact that the matrix algebra formula 610  
 563 credits CSs that would commonly be called condi- 611  
 564 tioned inhibitors with negative rates of reinforcement. 612  
 565 For what this means in terms of the cumulative prob- 613  
 566 ability function, see below. Intuitively, it means that 614

when inhibitory CSs (CSs predicting negative rates) 567  
 are combined with excitatory CSs (CSs predicting posi- 568  
 tive rates), the predictions cancel. That is, of course, 569  
 one of the behavioral effects by which conditioned in- 570  
 hibitors are recognized—their ability to suppress the 571  
 behavioral effects of conditioned exciters. 572

The interested reader will be able to see that there 573  
 are an infinite number of other unparsimonious so- 574  
 lutions in the Wagner, et al paradigm, all consistent 575  
 with the additivity assumption. They are all of a form 576  
 in which the positive rates of reinforcement attributed 577  
 to A and X sum to the rate observed on AX trials 578  
 and the rate for B is equal and opposite to what- 579  
 ever rate X is credited with, thereby explaining why 580  
 there are no reinforcements on BX trials. The same 581  
 is true for the AX(+1/2), BX(+1/2) group; there are 582  
 an infinite number of ways to “explain” the observed 583  
 rates of reinforcement within the rate-additivity con- 584  
 straint. In each case, however, there is only one so- 585  
 lution that maximizes the ratio between the informa- 586  
 tion conveyed and the number of CSs that convey it. 587  
 That is, there is only one parsimonious solution. The 588  
 unique bandwidth-maximizing solutions are the ones 589  
 the subjects in the Wagner, et al. experiment came to 590  
 (see Fig. 4). 591

Finding the solution(s) that minimize the num- 592  
 ber of predictors will not give a unique solution in 593  
 cases like overshadowing, where more than one CS 594  
 is by itself capable of conveying all the available 595  
 information about the timing of the next US. In 596  
 such cases, considerations other than parsimony must 597  
 determine which CS overshadows which. From an 598  
 information-theoretic perspective, the signal with the 599  
 higher signal to noise ratio is to be preferred, which 600  
 explains the effects of stimulus intensity and discrim- 601  
 inability in determination which CS overshadows 602  
 which (Mackintosh, 1976; Miles and Jenkins, 1973; 603  
 Pavlov, 1927, p. 141 ff). 604

When the characteristics of the signals (CSs) them- 605  
 selves offer no basis for preferment, then the system 606  
 must, in essence, flip a coin to decide which CS is to 607  
 be preferred. Flipping a coin in cases of perfect am- 608  
 biguity is what will generate data like those in Fig. 3. 609  
 Students of perception will recognize the similarity to 610  
 what happens in cases of perfectly ambiguous visual 611  
 stimuli (e.g. the vase-face or old/young lady pictures). 612  
 Some subjects see it one way; some the other; but no 613  
 subject sees both at the same time. 614

## 615 9. Analyticity, simplicity, and intuitiveness

616 Two attractions of the information processing ap-  
 617 proach are that its basic predictions are analytically  
 618 determined and they do not depend on free parameters.  
 619 By traditional scientific criteria, this means that these  
 620 models are simpler than associative models. Associa-  
 621 tive models abound in free parameters (see, for exam-  
 622 ple, Brandon et al., 2002), and their predictions can  
 623 generally be determined only by numerical methods,  
 624 that is, by computer simulation, the details of which  
 625 are rarely available for scrutiny and are hard to com-  
 626 prehend when they are. The matrix algebra formulae  
 627 given above can be implemented with paper and pencil  
 628 by anyone who remembers their linear algebra or by  
 629 anyone who goes to the trouble of reading one of the  
 630 many books that explain how to invert a square matrix,  
 631 and how to multiply a column vector by a matrix.

632 Fortunately, however, it is not necessary to have  
 633 command of the elements of modern algebra to deter-  
 634 mine the basic or first-order predictions of the model.  
 635 Matrix inversion and multiplication are so fundamen-  
 636 tal to many modern uses of mathematics that they are  
 637 built into Excel™. An implementation of the model  
 638 in Excel™, together with instructions for its use and  
 639 experimental protocols for a many basic condition-  
 640 ing experiments, including all those discussed in this  
 641 paper, may be downloaded from the SQAB website  
 642 (<http://sqab.psychology.org/>). Using the spreadsheet,  
 643 someone with no background in modern algebra can  
 644 determine what this model predicts, not only for  
 645 standard experimental protocols, but also for novel  
 646 protocols of their own devising.

647 So far as I know, no other contemporary model of  
 648 basic conditioning provides users with a tool that al-  
 649 lows them to calculate analytically the predicted re-  
 650 sults for user-specified real-time protocols. A *real-time*  
 651 protocol is one in which the times of event onset and  
 652 offset are specified. In most programs that simulate as-  
 653 sociative models of conditioning, the user must deliver  
 654 to the theoretical model a parsing of the event stream  
 655 into trials. Absent a principled specification of what  
 656 constitutes a trial, such a model is not actually well  
 657 enough specified to be physically (or computationally)  
 658 realized, because a theoretically crucial aspect of the  
 659 input, the arrangement of events into trials, depends  
 660 on arbitrary experiment-specific assumptions made by  
 661 the user of the model prior to asking the model for

a prediction. (For an analysis of the trial problem 662  
 in associative theories of learning, see Gallistel and 663  
 Gibbon, 2000.) 664

A further attraction is that predictions of the in- 665  
 formation processing models follow from information 666  
 processing principles in an intuitively obvious way. It 667  
 requires neither mathematics nor faith in inscrutable 668  
 computer simulations to see why the theory predicts 669  
 what it predicts. A basic principle is to ask, what does 670  
 the cumulative probability function look like in the 671  
 presence and in the absence of the CS in question? 672  
 (Fig. 6 and accompanying text explains how to con- 673  
 struct this function from an experimental protocol.) If 674  
 the presence of the CS in question does not change 675  
 that function, then there should not be a conditioned 676  
 response to the CS. Exceptions to this principle can 677  
 arise when one CS predicts another CS, which is cred- 678  
 ited with an influence on the expected time to rein- 679  
 forcement. In such cases, the animal may respond to 680  
 the one CS in anticipation of the appearance of the 681  
 second CS. This consideration, like the effects of stim- 682  
 ulus intensity on overshadowing, is not captured by 683  
 the analytic (spreadsheet) version of the model. 684

Let us see how basic information processing princi- 685  
 ples apply to the experimental results we began with. 686

- *Background conditioning (truly random control).* 687  
 The CS does not affect expected time to the next 688  
 US. Therefore, the information conveyed by the CS 689  
 is 0. This follows directly from the application of 690  
 Shannon's definition of information to conditioning 691  
 and the principle that only informative CSs elicit 692  
 conditioned responses. 693
- *Blocking.* Again, the addition of the second CS, in 694  
 compound with the already conditioned CS, does 695  
 not add any information. The US occurs at the time 696  
 already predicted by the first CS. 697
- *Overshadowing.* The CSs are redundant; either CS 698  
 alone can reduce the subject's uncertainty about 699  
 the time of US occurrence by as much as the two 700  
 together. Bandwidth maximization (aka parsimony) 701  
 applies: letting one of them do all the work max- 702  
 imizes the amount of information conveyed per 703  
 conveying signal. 704
- *Relative validity.* Again, the CSs are redundant. One 705  
 could partition the information conveyed among 706  
 two or more of the CSs, but letting one do all the 707  
 work is the parsimonious solution. 708

- 709 • *Conditioned inhibition.* When the matrix algebra  
710 formula is applied to inhibitory protocols, such  
711 as those shown in Fig. 5, it yields negative rates  
712 of reinforcement for the CS–’s (the conditioned  
713 inhibitors). That it must do so is obvious. The ob-  
714 served rates of reinforcement when the CS– is not  
715 present are positive; when the CS– is present, they  
716 are 0. The formula presupposes that rate estimates  
717 combine additively to determine the observed rates  
718 of reinforcement. Therefore, the rate estimate for  
719 the CS–, when added to the rate estimate for the  
720 CS+ (the excitatory CS) must equal 0. You do  
721 not need to do matrix algebra to see that the rate  
722 estimate for the CS– must therefore be equal to  
723 the rate estimate for the CS+ but opposite in sign.  
724 Fig. 9 shows what this implies about the effect of  
725 the presence of the inhibitory CS on the cumula-  
726 tive probability function in the simplest case—the  
727 explicitly unpaired protocol with a random rate of  
728 background reinforcements (upper protocol in  
729 Fig. 5). In the absence of the CS–, the cumulative  
730 probability of experiencing a reinforcement within  
731 some future interval rises toward 1 as the length of  
732 that future interval increases. However, so long as  
733 the CS– is present, the cumulative probability of  
734 experiencing a reinforcement within some future  
735 interval is 0, no matter how long the future interval.
- 736 • *The importance of temporal pairing.* Temporal  
737 pairing is neither necessary nor sufficient for con-  
738 ditioning, but that does not mean that it is unimpor-

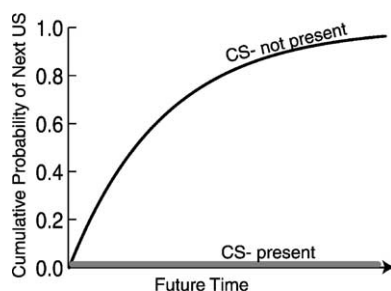


Fig. 9. The effect of an inhibitory CS (a CS–) on the cumulative probability function for the next US. In the absence of the CS–, the function rises exponentially (black curve), with an expectation equal to the expected interval between USs in that chamber (the background expectation). When the CS– is present, the expected interval between reinforcements is infinite (because the expected rate is zero). In other words, the cumulative probability of the next CS is 0 for the indefinite future (gray line lying on the temporal axis).

tant. It is important because the degree to which a temporal landmark reduces a subject’s uncertainty about the timing of the next US is proportional to its proximity to the US. The closer it is, the more it reduces the subject’s uncertainty. This follows from the fact that the degree of imprecision in a subject’s representation of an interval is proportional to the duration of the interval—Gibbon’s (1977) principle of scalar variability in the representation of temporal intervals.

- 739  
740  
741  
742  
743  
744  
745  
746  
747  
748
- *Time scale invariance.* Gallistel and Gibbon (2000) review a variety of findings that suggest that the conditioning process is, within limits, time scale invariant. Changing the intervals in a conditioning protocol has no effect on the results, provided they are all changed by the same factor (the scaling factor). Information processing models are naturally time scale invariant because the information in the flow of events is carried by the proportions among the intervals not by their absolute durations. Scaling the flow up or down changes the amount of time it takes to deliver the information, but it does not change the amount of information delivered by a given number of events.

749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762

Associative models, by contrast, are not time scale invariant. If they are trial-based models, then they must assume a trial duration. Rescaling the protocol will then move events (e.g. CS and US onsets) that were within the same trial into different trials, or vice versa, with strong consequences for the predictions of the model. Real-time associative models make complex assumptions about the dynamics of stimulus traces (see, for example, Brandon et al., 2002). Rescaling the protocol changes where events fall relative to these trace dynamics, again with strong consequences for the predictions of the model.

## 10. Summary

763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774

Shannon’s definition of the information conveyed by a signal in terms of the reduction in the subject’s uncertainty about the state of the world applies naturally to conditioning paradigms, where CSs (and/or the subject’s responses) reduce its uncertainty about when the next US (next reinforcer) will occur. Simple principles rooted in conventional information theoretic considerations, such as bandwidth maximization,

739  
740  
741  
742  
743  
744  
745  
746  
747  
748  
749  
750  
751  
752  
753  
754  
755  
756  
757  
758  
759  
760  
761  
762  
763  
764  
765  
766  
767  
768  
769  
770  
771  
772  
773  
774  
775  
776  
777  
778  
779  
780  
781  
782  
783

784 give intuitively obvious predictions for phenomena  
785 that are generally agreed to be central to our un-  
786 derstanding of the conditioning process—the effects  
787 of background conditioning, inhibitory conditioning,  
788 blocking, overshadowing, and relative validity. They  
789 also explain aspects of conditioned behavior that as-  
790 sociative theories rarely attempt to explain, for exam-  
791 ple, the rarely emphasized fact that the latency of the  
792 conditioned response is proportional to the CS–US  
793 interval (Gallistel and Gibbon, 2000, for review). Fi-  
794 nally, they explain something that it appears deeply  
795 difficult for associative models to explain, the time  
796 scale invariance of the conditioning process (Gallistel  
797 and Gibbon, 2000, for review).

## 798 References

- 799 Brandon, S., Vogel, E., Wagner, A., 2002. Stimulus representation  
800 in SOP: I. Theoretical rationalization and some implications.  
801 Behav. Proc. [SQAB '02 special issue].  
802 Church, R.M., Meck, W.H., Gibbon, J., 1994. Application of scalar  
803 timing theory to individual trials. *J. Exp. Psychol. Anim. Behav.*  
804 *Proc.* 20, 135–155.  
805 Gallistel, C.R., 1990. *The Organization of Learning*. Bradford  
806 Books/MIT Press, Cambridge, MA, 648 pp.  
807 Gallistel, C.R., 1999. The replacement of general-purpose learning  
808 models with adaptively specialized learning modules. In:  
809 Gazzaniga, M.S. (Ed.), *The Cognitive Neurosciences*, 2nd ed.  
810 MIT Press, Cambridge, MA, pp. 1179–1191.  
811 Gallistel, C.R., Gibbon, J., 2000. Time, rate and conditioning.  
812 *Psychol. Rev.* 107, 289–344.  
813 Gallistel, C.R., Gibbon, J., 2002. *The Symbolic Foundations of*  
814 *Conditioned Behavior*. Lawrence Erlbaum Associates, Mahwah,  
815 NJ, 196 pp.  
816 Gallistel, C.R., Mark, T.A., King, A.P., Latham, P.E., 2001. The rat  
817 approximates an ideal detector of changes in rates of reward:  
818 implications for the law of effect. *J. Exp. Psychol. Anim. Behav.*  
819 *Proc.* 27, 354–372.  
820 Gibbon, J., 1977. Scalar Expectancy Theory and Weber's law in  
821 animal timing. *Psychol. Rev.* 84, 279–335.  
822 Gibbon, J., Church, R.M., 1992. Comparison of variance and  
823 covariance patterns in parallel and serial theories of timing. *J.*  
824 *Exp. Anal. Behav.* 57, 393–406.  
825 Gluck, M.A., Thompson, R.F., 1987. Modeling the neural  
826 substrates of associative learning and memory: a computational  
827 approach. *Psychol. Rev.* 94, 176–191.  
828 Hawkins, R.D., Kandel, E.R., 1984. Is there a cell-biological  
829 alphabet for simple forms of learning? *Psych. Rev.* 91, 375–391.  
830 Hearst, E., Jenkins, H.M., 1974. Signtracking: The Stimulus–  
831 Reinforcer Relation and Directed Action. Psychonomic Society,  
832 Austin, TX.  
833 Kamin, L.J., 1969. Predictability, surprise, attention, and condi-  
834 tioning. In: Campbell, B.A., Church, R.M. (Eds.), *Punishment*  
835 *and Aversive Behavior*. Appleton-Century-Crofts, New York,  
836 pp. 276–296.
- King, A.S., McDonald, R., Gallistel, C.R., in press. Screening for  
837 mice that remember incorrectly. *Int. J. Comp. Psychol.* 838  
Krasne, F., 2002. Neural analysis of learning in simple systems. 839  
In: Gallistel, C.R. (Ed.), *Stevens Handbook of Experimental*  
840 *Psychology*, 3rd ed. Vol. 3. Learning and Motivation. Wiley,  
841 New York, pp. 131–200. 842  
LaBarbera, J.D., Church, R.M., 1974. Magnitude of fear as a  
843 function of the expected time to an aversive event. *Anim. Learn.*  
844 *Behav.* 2, 199–202. 845  
Mackintosh, N.J., 1975. A theory of attention: variations in the  
846 associability of stimuli with reinforcement. *Psychol. Rev.* 82,  
847 276–298. 848  
Mackintosh, N.J., 1976. Overshadowing and stimulus intensity.  
849 *Anim. Learn. Behav.* 4, 186–192. 850  
Miles, C.G., Jenkins, H.M., 1973. Overshadowing in operant  
851 conditioning as a function of discriminability. *Learn. Motiv.* 4,  
852 11–27. 853  
Miller, R.R., Escobar, M., 2002. Laws and models of basic  
854 conditioning. In: Gallistel, C.R. (Ed.), *Stevens Handbook*  
855 *of Experimental Psychology*, 3rd ed. Vol. 3. Learning and  
856 Motivation. Wiley, New York, pp. 47–102. 857  
Pavlov, I., 1927. *Conditioned Reflexes* (G.V. Anrep, Trans.). Dover,  
858 New York. 859  
Pearce, J.M., Hall, G., 1980. A model for Pavlovian learning:  
860 variation in the effectiveness of conditioned but not of  
861 unconditioned stimuli. *Psychol. Rev.* 87, 532–552. 862  
Rescorla, R.A., 1968. Probability of shock in the presence and  
863 absence of CS in fear conditioning. *J. Comp. Physiol. Psychol.*  
864 66, 1–5. 865  
Rescorla, R.A., 1972. Informational variables in Pavlovian  
866 conditioning. In: Bower, G.H. (Ed.), *The Psychology of*  
867 *Learning and Motivation*, vol. 6. Academic Press, New York,  
868 pp. 1–46. 869  
Rescorla, R.A., Wagner, A.R., 1972. A theory of Pavlovian  
870 conditioning: variations in the effectiveness of reinforcement  
871 and nonreinforcement. In: Black, A.H., Prokasy, W.F. (Eds.),  
872 *Classical Conditioning II*. Appleton-Century-Crofts, New York,  
873 pp. 64–99. 874  
Reynolds, G.S., 1961. Attention in the pigeon. *J. Exp. Anal. Behav.*  
875 4, 203–208. 876  
Rieke, F., Warland, D., de Ruyter van Steveninck, R., Bialek,  
877 W., 1997. *Spikes: Exploring the Neural Code*. MIT Press,  
878 Cambridge, MA. 879  
Shannon, C.E., 1948. A mathematical theory of communication. 880  
*Bell Syst. Technol. J.* 27 (379–423), 623–656. 881  
Usherwood, P.N.R., 1993. Memories are made of this. *Trends*  
882 *Neurosci.* 16, 427–429. 883  
Wagner, A.R., 1981. SOP: a model of automatic memory  
884 processing in animal behavior. In: Spear, N.E., Miller,  
885 R.R. (Eds.), *Information Processing in Animals: Memory*  
886 *Mechanisms*. Lawrence Erlbaum, Hillsdale, NJ, pp. 5–47. 887  
Wagner, A.R., Logan, F.A., Haberlandt, K., Price, T., 1968. 888  
Stimulus selection in animal discrimination learning. *J.*  
889 *Exp. Psychol.* 76, 171–180. 890  
Wasserman, E.A., Franklin, S.R., Hearst, E., 1974. Pavlovian  
891 appetitive contingencies and approach versus withdrawal to  
892 conditioned stimuli in pigeons. *J. Comp. Physiol. Psychol.* 86,  
893 616–627. 894