An Information Theoretic Approach to Model Selection: A Tutorial with Monte Carlo Confirmation

M. Christopher Newland¹

Published online: 19 June 2019

© Association for Behavior Analysis International 2019

Abstract

A reliance on null hypothesis significance testing (NHST) and misinterpretations of its results are thought to contribute to the replication crisis while impeding the development of a cumulative science. One solution is a data-analytic approach called Information-Theoretic (I-T) Model Selection, which builds upon Maximum Likelihood estimates. In the I-T approach, the scientist examines a set of candidate models and determines for each one the probability that it is the closer to the truth than all others in the set. Although the theoretical development is subtle, the implementation of I-T analysis is straightforward. Models are sorted according to the probability that they are the best in light of the data collected. It encourages the examination of multiple models, something investigators desire and that NHST discourages. This article is structured to address two objectives. The first is to illustrate the application of I-T data analysis to data from a virtual experiment. A noisy delay-discounting data set is generated and seven quantitative models are examined. In the illustration, it is demonstrated that it is not necessary to know the “truth” is to identify the one that is closest to it and that the most likely models conform to the model that generated the data. Second, we examine claims made by advocates of the I-T approach using Monte Carlo simulations in which 10,000 different data sets are generated and analyzed. The simulations showed that 1) the probabilities associated with each model returned by the single virtual experiment approximated those that resulted from the simulations, 2) models that were deemed close to the truth produced the most precise parameter estimates, and 3) adding a single replicate sharpens the ability to identify the most probable model.

Keywords Information-Theoretic · Model comparison · Replication crisis · Akaike Information Criterion · Delay discounting · Monte Carlo simulation

✉ M. Christopher Newland
Chris.Newland@auburn.edu

¹ Department of Psychology, Auburn University, Auburn, AL 36849, USA

The inability to replicate scientific findings has significant implications both for the advancement of our understanding of nature and public confidence in the conclusions of basic and applied research (Branch, 2019). Whether we seek to predict calamities like global warming, determine the impact of pollution, find treatments for serious diseases, or explore theoretical models of behavior, reports that experiments cannot be reproduced slows progress and damages the public acceptance of, and willingness to support, empirically grounded policy recommendations. A degraded science creates opportunities for those who wish to advance a personal or doctrinaire agenda that is disconnected from evidence or that lacks a scientific foundation.

The movement to develop evidence-based treatments (Smith, 2013) or evidence-based policy (Bickel et al., 2016; Higgins, Davis, & Kurti, 2017; Madden, Price, & Sosa, 2016; Newland & Bailey, 2016) relies on the publication of statistically significant effects under the assumption that they will be reproducible. If that foundation is weak, then the treatments or policies will be misguided. Poor replicability also has significant financial implications. For example, pharmaceutical companies report difficulties reproducing both published and in-house preclinical studies conducted to identify the next generation of therapeutic drugs (Prinz, Schlange, & Asadullah, 2011). As a result, drugs that were in the pipeline for clinical applications were withdrawn and limited resources were wasted.

We do not know the extent to which replication failures arise in behavior analysis, but the canonical small-*N* designs commonly employed in behavior analytic research have within- and between-participant replication baked into them (Perone, 2018). A study may, for example, contain a within-participant replication in the form of an ABAB experimental design, in which the A phase is a baseline and the B phase is an intervention, to examine the impact of an intervention. The second AB phase, when it occurs, constitutes a within-participant replication. Repetition across several individuals then represents multiple within-subject as well as between-subject replications (Baer, Wolf, & Risley, 1968; Kratochwill et al., 2013). This powerful approach yields a cumulative science that is less fraught with the difficulties that accompany null hypothesis significance testing (Killeen, 2005; Sanabria & Killeen, 2007) but certainly not immune to many of the pressures that an investigator feels, such as publication bias or the selection of individual datasets to publish (Shadish, Zelinsky, Vevea, & Kratochwill, 2016; Tincani & Travers, 2019). Behavior analysts, however, still must be aware of limitations of this approach, especially when forming conclusions about the generality of their results in clinical populations. Small-*N* experiments are powerful but present challenges to identifying the population-level success rate of a particular intervention (how frequently does it work?) and broad generality of effects (in what disorders or populations is this treatment effective?).

Null Hypothesis Significance Testing (NHST)

The difficulties in replicating research results have been attributed to a number of causes, including the pressures on scientists to publish in high-impact journals, the necessity of showing positive results to receive funding and prestige and, of course, the all-too-human tendency to see data with a parental eye in which patterns not visible to others seem obvious to the experimenter. Some scientists and data analysts view NHST as the fundamental problem and are suggesting that reliance on this practice be greatly reduced or abandoned altogether (Anderson, Burnham, & Thompson, 2000; Branch, 1999; Branch,

2019; Cohen, 1994; Johnston & Pennypacker, 1993; Perone, 1999; Wilkinson, 1999). Others note that, with appropriate precautions, NHST can be informative and can usefully be continued (Greenland, 2019; Hales, Wesselmann, & Hilgard, 2019; Kmetz, 2019). Skepticism about this widespread practice is appearing with increasing frequency. A recent issue of *The American Statistician* contains 43 papers addressing problems that arise from tests of statistical significance (see Wasserstein, Schirm, & Lazar, 2019, for a summary).

One difficulty with NHST is that it does not, and cannot, answer the very question that a scientist wants to answer and what many, probably most, think that it does answer: what is the probability that our hypothesis is true in light of the data collected ($P[H | \text{Data}]$; Anderson, 2008; Cohen, 1994; Nuzzo, 2014)? Instead, NHST yields the probability of obtaining a result as extreme as the one actually obtained if some null hypothesis is true, that is, it provides an estimate of $P(\text{Data} | H_0)$. This twisted formulation raises logical conundrums that confuse even most advanced and thoughtful investigators. Many examples can be offered to illustrate differences in the two formulations. The probability that it is cloudy given that it is raining is not the same as the probability that it is raining given that it is cloudy. The probability that someone is a male given the presence of a nose (50%) is not the same as the probability of a nose given that someone is male (99+%).

In addition to the logical difficulties, the H_0 is often one that no sane person uncontaminated by instruction in NHST would even bother to ask. For example, suppose one wanted to know about the height difference between mature pine trees and saplings. The null hypothesis is the claim that they are both the same height so, after measuring saplings and trees, the probability that 30-foot pine trees could have come from a population whose mean is 6 inches is assessed. Of course, this probability is vanishingly small, but so what? In a strict sense, all that is assessed is the probability that the highly unlikely null hypothesis could have yielded the result. The answer that we need is the probability that the null hypothesis, or a more reasonable one, is true *given the data collected*. Many scientists forget that that $P(\text{Data} | H_0)$ is the question actually being tested so they write and talk as though they have determined $P(H_0 | \text{Data})$, a subtle turn of logic that is a step down the road of replication failures.

What's the Problem?

The reliance on NHST can lead to results that may not be replicated, especially when based on a single study reporting a p value close to 0.05. Such failures are more likely when there is a financial or other stake in the outcome that could bias publication decisions, if a research is on a hot topic, or if the result is surprising or, prior to conducting the research, unlikely (Ioannidis, 2005). In fact, how close $P(\text{Data} | H_1)$ is to $P(H_1 | \text{Data})$ depends greatly on the plausibility of the alternative hypothesis (Nuzzo, 2014; Sellke, Bayarri, & Berger, 2001). An example of this situation is the tree/sapling example above. The conclusion (adult trees are taller than saplings) is highly plausible, so $P(\text{Data} | H_1)$ approximates $P(H_1 | \text{data})$.

If the conclusion is a long shot then the probability that the results will be replicated, are much lower than if the outcome is highly plausible. If the outcome is judged to have only a 5% chance of occurring before the study is conducted, a surprising outcome, then the probability that it will be replicated is only 11% if H_0 is rejected with a p value of 0.05 and only 30% if $p = 0.01$, so solid results and extensive replication must be produced before such a result is accepted (Nuzzo, 2014). This is simply another way of saying that

extraordinary claims require extraordinary evidence. In another calculation, if the true difference between two means is actually equal to the difference obtained in an experiment, with a p value of 0.05 the probability of replication is still only 50% (Goodman, 1992). These values are far different from what many investigators expect.

Long shots are important. Although highly plausible outcomes can confirm the validity of one's methods, they rarely push science forward. For that, we need surprising outcomes: "we often feel compelled to turn our attention from problems which directly confront us to some unexpected new phenomena which introduce fresh problems or which necessitate a revision of old points of view" (Pavlov, 1927/1960). It is the surprising results that challenge well-established theories, open new avenues of inquiries, and lead us to question our intuitions as they become confirmed. It has taken sustained research programs and multiple replications to confirm such unexpected results as the robustness of false memories and unreliability of eyewitness testimony (Kassin, Tubb, Hosch, & Memon, 2001; Loftus, 1993, 2003), flavor aversion (Davidson & Riley, 2015; Freeman & Riley, 2009), conditioning of the immune system (Cohen, Moynihan, & Ader, 1994), conditioned compensatory responses (Siegel, Baptista, Kim, McDonald, & Weise-Kelly, 2000), or hyperbolic discounting (Ainslie & Monterosso, 2003; Chung & Herrnstein, 1967; Madden, Begotka, Raiff, & Kastern, 2003; Rachlin, 2006), to name a few examples. Such replication is necessary to rule out false surprises before they are applied, because many will not be replicated. Unfortunately, it is also in the surprising results that the discrepancy between the two formulations, $P(\text{Data} \mid H_0)$ and $P(H_0 \mid \text{Data})$, is the greatest, so it is in these surprising results that NHST is most likely to lead us astray.

What's the Solution?

The ultimate solution is replication, the foundation on which science grows, though replication is not as simple as it may sound (Laraway, Snyderski, Pradhan, & Huitema, 2019, offer a thoughtful discussion of replication within psychology and behavior analysis). Replication sometimes occurs within a single report, especially within behavior analysis, but this is not common in many areas so requiring replication could slow progress by placing too high a threshold on publishing. Other suggestions include the preregistration of data-analysis plans, to reduce the prevalence of p -hacking, and stating the false-positive rate, to provide some context for the p -value (Hales et al., 2019; Leek et al., 2017). These are all solid suggestions that can only improve the integrity of science. Other suggestions include tactics that are embedded in data analysis and that predict the outcomes of attempts to replicate (Killeen, 2019).

Bayesian approaches to hypothesis testing, in which one's knowledge of a phenomenon is incorporated into the analysis and decision-making, is recommended frequently and its advantages are becoming more widely recognized (Depaoli, Rus, Clifton, van de Schoot, & Tiemensma, 2017; Goodman, 2001; Kruschke, 2013; Young, 2019). Bayesian inference is based on maximum likelihoods and provides the probability of the hypothesis given the dataset by building on Bayes' rule:

$$p(H \mid \text{Data}) = \frac{P(\text{Data} \mid H) P(H)}{P(\text{Data})} \quad (1)$$

That is, the probability of a hypothesis given the data obtained is proportional to the probability of the data given the hypothesis (as in NHST) multiplied by the probability assigned to the hypothesis before the data were collected (the “prior”). For example, if one were calculating allocation of behavior to a source that produced 4 times as many reinforcers as a lean alternative then, all else being equal (e.g., lack of bias), it would be unlikely that 99% of the selections would be on the leaner alternative; it is more likely that the most responses would occur on the richer alternative. This information about prior likelihoods can be incorporated in $P(H)$. The probability of the dataset ($P(\text{Data})$) remains constant throughout these calculations so the result can be viewed as a constant of proportionality. Prior information improves the precision of our estimates and helps with the management of outliers, but the approach is beyond the scope of this article. (Depaoli et al., 2017; Kruschke, 2013; Young (2019) offer a brief introductions with further references). The I-T approach described here is equivalent to a Bayesian analysis with “savvy priors” (Burnham & Anderson, 2002), but much simpler (Killeen, 2019).

A second approach is the statistic called p -Rep (the probability of replication) (Killeen, 2005; Sanabria & Killeen, 2007). By linking the effect size arising from NHST to the theory of signal detection, p -Rep describes the probability of obtaining a new result in the same direction as the original result. For example, if one conducts an experiment showing that males are taller than females and reports a p -value and an effect size, then p -Rep will describe the probability that a second experiment will also show that males are taller than females. Note that this does not report the probability that the *same* p value will be obtained, an unrealistic standard, only that the second outcome is in the same direction as the first. This strategy has been advanced to address some of the problems above. Although p -Rep is not a Bayesian approach because it does not involve speculating about prior probabilities, it was described as being “friendly” to Bayesian analyses (Killeen, 2005).

A third approach is found in information-theoretic (I-T) techniques (Anderson, 2008; Burnham & Anderson, 2002; Wagenmakers & Farrell, 2004). Two primary advantages to this approach can be noted. First, like Bayesian approaches, I-T techniques answer the question that scientists would like from their data: what is the empirical evidence for the hypotheses that we are testing or, likewise, what is the probability of the hypothesis given our data (e.g., $P[H_1 | \text{Data}]$)? An I-T analysis is distinct from Bayesian approaches because it does not require the computation of prior hypotheses for each model parameter (Young, 2019) and therefore is less computationally intensive. Instead, as shown below, the I-T analysis rank-orders all hypotheses considered based on their “distance” from the truth, even though the truth is not known, based on information available from the commonly used statistical packages. Second, the I-T analysis allows, or rather encourages, an investigator to test many hypotheses simultaneously, a reasonable tactic that a scientist wants to do anyway. That is, it allows a researcher to evaluate $P[H_1 | \text{Data}]$, $P[H_2 | \text{Data}]$, $P[H_3 | \text{Data}]$, . . . , $P[H_N | \text{Data}]$, where N represents the number of hypotheses being tested. This contrasts directly with NHST in which an investigator is explicitly forbidden to test multiple hypotheses without taking extra steps, steps that also weaken experimental power. This is because embedded in NHST is the problem that if one tests a large number of hypotheses, then some proportion of them will reach “statistical significance” by chance, just as if one flips a coin enough times, then a sequence of HHHHHHHH will occur on occasion. Of

course, investigators often do test multiple hypotheses without reporting them, a practice called “*p*-hacking” (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Leek et al., 2017).

I-T-based analyses are grounded in the log-likelihood approach to data analyses so first we will briefly examine log-likelihood data analyses. Then we build from log-likelihood analyses to identify the best of a candidate set of regression models. Although the analysis will emphasize modeling regression techniques, a brief comment about extending this to ANOVA approaches will be offered.

Two excellent sources can be recommended to the reader interested in more background and explication. A clearly written and detailed treatment, containing the mathematics behind the approach can be found in (Burnham & Anderson, 2002). A practical and accessible, but still rigorous, introduction to this material is (Anderson, 2008). The present article draws from these two sources.

The I-T approach is emphasized because of its ability to provide quantitative comparisons of behavioral models and the relative simplicity of applying it to regression-based model construction. Quantitative modeling has been important to our understanding of learning, both within and outside of behavior analysis, as evidenced by the explorations of Clark Hull, William Estes, Richard Herrnstein, and their successors, to name just a few examples. It should be emphasized that I-T analysis represents just one possible approach to increasing the probability of replication and it is certainly not a panacea.

I-T Analyses in Behavior Science

Although hypothesis testing remains firmly positioned and, within behavior analysis, may still be on the rise (Hales et al., 2019), I-T analyses grounded in the Akaike Information Criterion (Akaike, 1973), especially as developed by Burnham and Anderson in their book (Burnham & Anderson, 2002), are seeing increasing application in behavior analysis and other behavioral sciences. To explore the emergence of its use, the number of citations to either Hirotugu Akaike or to one of Burnham and Anderson’s papers on this topic or, where full text was available, the number of papers containing the word “Akaike,” was examined for journals in which behavior analysts frequently publish: *Journal of the Experimental Analysis of Behavior*, *Behavioural Processes*, and *Journal of Experimental Psychology: Animal Learning and Cognition* (nee, *Journal of Experimental Psychology: Animal Behavior Processes*). For comparison, two journals that frequently publish on animal behavior but with a heavier emphasis on ecological considerations, where I-T approaches have been used with great frequency, were examined. These are *Animal Behaviour* and *Applied Animal Behavior Science*. To represent applied behavior analysis, the *Journal of Applied Behavior Analysis* was examined but no citations to the key methodologists were found. Limitations of this strategy are acknowledged. It is not comprehensive because citations to other methods papers about I-T analyses would be overlooked. The journals selected are not the sole outlets for behavior analysts nor do they publish only behavior analytic work, but they are representative of behavior science as conducted with experimental animal models so it is likely that this will capture of the extent to which I-T approaches are becoming used and how they are applied.

Using Google Scholar, the above journals were searched for citations to H Akaike or to both Burnham and Anderson through 2018. The first citation located was 1982 (Hunter & Davison, 1982). The search was conducted on January 11, 2019, so it is possible that some papers with 2018 publication dates were missed. A cumulative record of these citations was generated and plotted for each journal (Fig. 1). Citations to the key methods book or papers began in earnest in the 2000s and once it began it showed a steady rise. As expected, the journals closer to ecology (*Animal Behaviour* and *Applied Animal Behaviour Science*) show the highest number of citations (486 and 97, respectively) but for all journals the trend is solid and shows no sign of abating. The manner with which this approach has been used is addressed later.

Likelihood

I-T analyses, like Bayesian analyses, are grounded in likelihood estimates, which can be illustrated with a simple case of estimating the bias of a coin. Suppose a coin is flipped 50 times and it shows “heads” 10 times. With this information, we can examine the likelihood that different estimates of bias describe the results and each estimate of bias can be viewed as a separate model of the coin-flip data. Our goal is to determine which estimate of the coin’s bias, or which statistical model, best describes the results of our experiment. That is, we want to know which the model best fits the data that were collected.

In the NHST approach, we would form a null hypothesis that the probability of a head equals that of a tail. We use the binomial distribution to calculate the probability of obtaining an outcome that extreme or more extreme (10 or fewer heads in this example) given our null hypothesis of no bias. The answer, which is that the probability of obtaining at most 10 heads from an unbiased coin is 0.000012, is relatively uninformative. Wouldn’t it be better to determine just what the bias is given the data collected? Fig. 2 shows the likelihood (i.e., probability) of different biases given the observation of 10 heads. The most likely model is one that says that the coin has a 20% chance of coming up heads, and that probability is close to 0.14 ($P[p(H) = .20|10H] = 0.14$). Because this analysis reveals the probability of models that arise from the data collected, it is more informative than the NHST assertion that the data are unlikely to come from an unbiased coin. It might be noted that this analysis assumes flat priors, or that any probability of bias is equally likely before the coin is tossed. If the prior probability is viewed as 0.50, based on the structure of the coin, then a Bayesian analysis would show posterior probabilities that lie between 0.2 and 0.5.

This example uses a simple binary model to illustrate three important goals of modeling. The first is that we would like to know the probability of a specific parameter or model given the data. When this cannot be determined exactly, at least we can determine which of a set of models is most likely to account for the obtained data: a bias of 0.2 is the most likely in our example but models close to that value, like 0.19 and 0.21, are strong contenders. Second, we would like a metric that can be used to score which models are best. For example, how much better is the model that the $P(\text{Heads}) = 0.2$ than the model that $P(\text{Heads}) = 0.3$, or the model that the probability lies between 0.1 and 0.3? Finally, we would like to acknowledge that it is acceptable, even preferable, to test a large number of models. In fact, an infinite number of models were tested in Fig. 2 because the entire range of probabilities

Fig. 1 Cumulative count of citations to key model-selection methodological books or papers or to Akaike for five journals, beginning in 1982, the earliest citation located, and ending in 2018. Note at the maximum of the vertical axes ranges from 20 to 500

from 0 to 1 was assessed. This is an important distinction from NHST, which discourages the testing of many models.

Fig. 2 Illustration of likelihood that tossed coin that shows heads on 10/50 flips is biased by the value on the horizontal axis

The Data Set

An I-T approach to data analysis will accomplish the three goals described in the previous section. A collection of models is tested against a data set to determine which one best represents the data. To introduce the approach, we will use regression modeling of a delay-discounting data set generated using known parameters. Each of seven candidate models will be fit to the data using least-squares regression. Then, beginning with either the residuals or the Akaike Information Criterion corrected for small data sets (AICc), we will determine for each model the probability that it is the best of all the models examined. Because we know how the data were generated, we can evaluate the ability of this approach to detect the actual model. This single-point determination of accuracy will be supplemented with a Monte Carlo approach in which 10,000 data sets are generated and the candidate models compared.

Consider an experiment in which a mouse selects one of two levers. The “smaller–sooner” (SS) lever provides a drop of milk after a 1.25 sec delay. The “larger–later” (LL) lever other provides 4 drops of milk after a variable delay that varies from 0 s to 91 s. The dependent variable is the proportion of times that the mouse selects the LL lever. Such an experiment might produce the data illustrated in Fig. 3. The solid curve plots the probability of selecting the LL lever at progressively increasing delays as produced by Eq. 2:

$$Y = \frac{0.85}{1 + 0.2D} \quad (2)$$

The delays used were 0, 1.57, 2.46, 3.87, 6.08, 9.54, 15.0, 23.51, 36.91, 58.8 and 91.0 s, values that are logarithmically based. When both reinforcers are delivered

immediately ($D = 0$), the model specifies that the animal chooses the larger reinforcer 85% of the time ($Y = .85$). As the delay increases, the larger reinforcer is selected less frequently. The steepness of the function is determined by the discounting rate ($k = 0.2$): the curve gets half-way to zero at $1/0.2 = 5$ seconds. The data points in Fig. 3 were generated from Eq. 3 for 11 values of D :

$$Y = \frac{0.85}{1 + 0.2D} + \varepsilon_i \quad (3)$$

The error terms (ε_i) were calculated such that a) they have a mean of 0 and a standard deviation of 0.05; b) the resulting Y values remained between 0 and 1, as required of a proportion; and c) the distributions at the extremes are approximately normal, as required by the regression models used. Figure 3 represents a single experiment in which behavior is generated by Eq. 3, with all other factors represented by ε_i . Together, the symbols contain the entire information content and represent the “truth” to be modeled. The truncation of the error function so the data range between 0 and 1 limits the generality somewhat, but because many discounting functions begin at 0.85 to 0.90 or so and are greater than 0 at the longer delays, this limitation is not viewed as severe. Constructing different error functions is possible but doing so would complicate the presentation while offering little added value.

Two Reasons to Model

We want to model these data, but how the models are used depends upon the goal of the study. We might be interested in characterizing the process that gives rise to the data. Is it a hyperbolic, linear, or exponential process? Perhaps it is a hyperboloid model in

Fig. 3 The true discounting with noise added, representing the truth in the model estimates below

which the discounting rate (here, $k = 0.2$) is modified by raising the denominator, $1 + D$ (Myerson & Green, 1995) or D (Rachlin, 2006) to a power. These all have theoretical implications for how discounting occurs in different species or situations.

As an alternative, we may care little about what form the model takes, but wish to predict what will happen with future implementations of this procedure or to compare the outcome with that obtained under different conditions. This would arise, for example, if we wish to predict what would happen after a delay of 5 s or the delay that corresponds to a 50% decrease in the value of a reinforcer (Franck, Koffarnus, House, & Bickel, 2015). In this case, we want a good prediction. Such an approach might be important if we wanted to determine the effect of a drug, neurotoxicant, genetic background, or environmental stressor on the rate at which discounting occurs if, say, we had models that were difficult to distinguish from one another. By combining competing models, all the information that is available can be harnessed to provide an estimate that is more precise than any single model while being agnostic about what the true model is. As described below, this is accomplished by applying to each candidate model a weighting factor that is based on the probability that a model is close to the true generating function, to each model.

The Models to be Tested

We begin by simulating data from a single subject experiencing 11 delays (Fig. 3). These data could be described by a large number of models but only seven (Table 1) will be evaluated here. To those accustomed to traditional regression modeling, this may seem like too many models to test, but economic or ecological models that take I-T approaches to analysis will sometimes test dozens, if not hundreds, of models. Of course, they would need many more than 11 data points to do this. In fact, seven models push the limit for only 11 data points from one virtual subject; but the limited data set is done here only to keep the example manageable. Later we will examine a situation in which there are two estimates at each delay.

Table 1. Models to be Tested

Model Number & Description	Model
Nested Under a Linear Model	
1. Linear	$Y = a - bD$
2. Mean (Null Model)	$Y = a$ (where $a = \text{Mean } Y$)
Unnested	
3. Exponential	$Y = a e^{bD}$
Nested Under the Hyperboloid	
4. Hyperboloid with fitted intercept	$Y = \frac{Y_0}{(1+kD)^x}$
5. Hyperboloid with constant intercept	$Y = \frac{1}{(1+kD)^x}$
6. Hyperbolic with fitted intercept	$Y = \frac{Y_0}{1+kD}$
7. Hyperbolic with constant intercept	$Y = \frac{1}{1+kD}$

Some models in Table 1 are said to be “nested” because they rely on the same basic equation but include different numbers of parameters. Two are nested under the linear model, one includes a slope and an intercept (Model 1) and a second includes only the intercept (Model 2). The latter, which simply sets the proportion of LL choices to the overall mean, would serve as the null hypothesis in traditional regression modeling. Note that models 1 and 2 could be nested under a polynomial such as $Y = Y_0 + aX + bX^2 + cX^3 + \dots$, so many others could be included in that general class. The models nested under the hyperboloid use a generic hyperbolic equation

$$Y = \frac{Y_0}{(1 + kD)^a} \quad (4)$$

but Y_0 is set to a constant, 1.0, for Models 5 and 7, and a is set to a constant, 1.0, for Models 6 and 7). The exponential model is not nested under any of these.

Whether models are nested does not affect the analysis, but it can be useful to know about nesting when using software applications. Many programs will permit the user to enter a generic equation and request that all nested models be tested. This convenience should be taken with caution because a complex model can have nested models that should not be considered because they have no theoretical or practical support.

Although nested and unnested models can all be tested within a data set, it is necessary that all analyses be confined to a single data set. This is because all comparisons are based on the information contained in the dependent variable. This variable is, in turn, intimately linked to the vertical deviations from a prediction, the sum of squared deviations. Any change of dependent variable fundamentally alters the total information content. Even transforming the dependent variable, such as taking its logarithm or normalizing it, requires a different I-T analysis. (Transforming an independent variable is just another model and is perfectly acceptable). Some approaches to combining information from different datasets are described below.

The Grand Mean

We begin with a simplistic model, the mean, chosen only to illustrate a point and to provide an anchor. The mean (Table 1, Model 2), which commonly serves as the null hypothesis in regression models, predicts that the probability of selecting the LL option is constant at all delays. The top left panel of Fig. 4 shows the obtained data from a single simulation as symbols and the model as a horizontal line. The difference between the truth and the estimated model is illustrated by the vertical lines connecting the model to obtained data. The mean is a good predictor of the proportion of LL choices at a 6-s delays, but grows increasingly inaccurate as the delay diverges from this point.

This example resembles those used to illustrate the quality of a least-squares regression model using the variance accounted. Both the I-T and traditional approaches begin at this point but then they take different paths. Traditional regression modeling begins with the sum of the squared discrepancies between the model and the obtained data and uses this value to calculate both the variance accounted for and the p values for different regression coefficients. A likelihood interpretation of the vertical lines, and that aligns with the I-T approach, is that they represent the distance between the model

Fig. 4 Fits for four models. The symbols in each plot show the data generated from the discounting function plus noise illustrated in Figure 2. The fitted line shows least-squares best estimate using the equation in the upper right corner of each graph. The vertical lines show the deviation between the model and the fit, and can be interpreted as how surprising that data point would be in light of the model or as the loss of information that would occur if that model were adopted

and the true process that generated the data or how far the model would have to move to accommodate the data.

In the I-T framework, the vertical lines' length can be viewed as how surprising, or unlikely, any given data point would be if the model that produced the line were assumed to be the true one (Anderson, 2008). For example, Model 2 (the grand mean) predicts that the LL option will be selected on 34% of the opportunities at a 6-s delay. In fact, the LL option was selected on 41% of the opportunities, a relatively unsurprising result. At a 0-s delay, though, Model 2 still predicts that 34% of the choices will be on the LL option but, in fact, 75% of the choices were made there. It would be shocking if this outcome were generated by a process described by Model 2. The probability of obtaining the data point at 0-s given the model is very small so that datum is a highly informative of a flawed model. This information-based interpretation of the discrepancies between prediction and data leads one down a path very different from the one initiated by a variance-accounted-for interpretation. Discrepancies between observed

and predicted values at lower and higher delays indicate that the model being tested is flawed because $P(\text{Model 2} \mid \text{Data})$ is so small.

The Linear Model and the Kullback-Leibler Information

A linear model (Fig. 4, top right) is used to introduce additional concepts in the I-T approach to model fitting, concepts that will be used to build the model comparisons shown in Table 2. The first is that of Kullback-Leibler Information, also known as the K-L distance, which is derived from the Akaike Information Criterion (AIC) and residual sum of squares. The K-L distance is a fundamental unit that describes the information lost by replacing the true situation (the hyperbolic plus noise) with a model. The K-L distance from the truth provides the basis for comparing models. The original data contain all the information, including the underlying process (the solid line in Fig. 2) plus other influences (the data points) that produce the seemingly random deviations from that process. Each hypothesized model approximates the underlying process, but omits some of the information contained in the original data set. The model with the smallest loss of information is the closest to the truth.

A key discovery by Akaike in his development of the AIC is that the Kullback-Leibler Information can be used to develop an information scale measuring the distance between a model and the truth, hence the designation “K-L distance.” K-L distance has excellent measurement properties, including being measured on a ratio scale, so distances can be compared by subtraction or even by taking ratios. It should be noted, however, that although K-L distance has many characteristics of distance, it lacks an important one. Because it is a conditional probability, the distance from Model A to Model B is not the same as the distance from Model B to Model A. This consideration lies in the background of the developments below and leads to some hesitation in calling it a distance but, because of the designation’s relatively simplicity, because it connotes some important characteristics of model comparisons, and because it works well enough, we will refer to it as a distance.

Each candidate model contains a certain amount of information, which is inevitably less than that contained in the process that generated the data. The questions to be addressed are: how much information is lost when modeling the data, how can we minimize that loss, and how should we proceed in the real situation? We will show that even if we cannot know the truth, we can still determine which model comes closest to it in K-L information units, or K-L distance.

Table 2. Model Comparisons for the Null, Linear, Exponential Models

Model Name	RSS	K	Log-Lik	AICc	ΔAICc	Evidence Ratio	AICc Weight	RSQ	Model Rank
Exponential	0.040	3	15.32	-21.22	0.0	1.0000000	0.9999737	0.941	1
Mean	0.797	2	-1.17	0.34	21.6	0.0000208	0.0000208	0.000	2
Linear	0.359	3	3.21	3.00	24.2	0.0000055	0.0000055	0.549	3

Note: The Evidence Ratio is the probability that this model is better than the highest ranked model. Bolded terms do not depend on the models included in the set of comparisons. The unbolded terms are different from Table 3

The residual sum of squares (RSS) represents a choice point for the modeler. It can lead to estimates of the probability of obtaining data at least as extreme as that actually obtained, as in the conventional NHST framework, or it can lead to estimates of the likelihood ratio and, from that, the probability of the model given the data, as described here. The chain of calculations is long but each stage is simple: RSS → least-squares estimates of variance → Maximum-Likelihood Variance → Log (Likelihood) → AICc → K-L distance → Model probabilities. The K-L distance can be generated from the RSS, the Akaike Information Criterion, or other values provided from most software used to conduct regression modeling. If software provides one of the intermediate steps, then the precursor steps can be eliminated. [Appendix 1](#) summarizes these steps in tabular form.

Log Likelihood and AICc

The log-Likelihood can be determined from the RSS using Eq. 5 (assuming normally distributed residuals derived from least-squares regression):

$$\text{Log(Likelihood)} = -\frac{N}{2} \log \frac{\text{RSS}}{df} = -\frac{N}{2} \log(\sigma^2) \quad (5)$$

where N is the sample size and σ^2 is the residual variance (Akaike, 1973; Burnham & Anderson, 2002).

The AICc can be calculated from Eq. 5 as follows:

$$\begin{aligned} \text{AICc} &= -2\log(\text{Likelihood}) + 2K \frac{N}{N-K-1} \\ &= -2\log(\text{Likelihood}) + 2K + \frac{2K(K+1)}{N-K-1} \\ &= \text{AIC} + \frac{2K(K+1)}{N-K-1} \end{aligned} \quad (6)$$

where K is the number of estimable parameters, N is the number of data points, and AIC is uncorrected for small samples (Akaike, 1973; Burnham & Anderson, 2002). The variance is an estimated parameter, so K is 1 plus the number of regression parameters estimated (Anderson, 2008; Burnham & Anderson, 2002). Thus, for a linear model, $K = 3$ because the intercept, slope, and variance are all estimated, and for the null model, $K = 2$ because the intercept and variance are estimated. Equation 6 is a sum of $-2\log(\text{Likelihood})$ plus penalties, one penalty for the number of parameters ($2K$) and a second for small sample size. Some sources suggest that the AIC can be used when the ratio n/K is greater than 40 and the AICc should be used otherwise, but because the third term approaches 0 as the sample size gets large the AICc can be used for all applications (Burnham & Anderson, 2002). The penalty terms offer some reassurance when building complex models that the benefits of such models can be counteracted by their cost as long as the number of parameters is smaller than $N - 1$.

Equations 5 and 6 can be combined to calculate the AICc from the variance, if that is available, as shown in Equation 7. If the log likelihood value is not available, but a maximum-likelihood estimate of variance is then it is still possible to estimate the AICc using Equation 7:

$$AICc = N \log(\hat{\sigma}^2) + 2K + \frac{2K(K+1)}{N-K-1}. \quad (7)$$

Here, $\hat{\sigma}^2$ is the maximum likelihood calculation of the residual variance and is equal to the least-squares estimate multiplied by $(N-2)/N$.

Caution is warranted when using software-provided terms. Software used to conduct least-squares regression does not include the variance in the count of parameters but often packages do not even include the variance when determining the AIC. Sometimes it is not clear whether maximum likelihood or least squares estimates of the variance is provided. A few simple arithmetic checks are warranted.

ΔAICc

Now models can be compared. In Table 2, we examine the results of fitting models 1, 2, and 3 from Table 1 to the data in Fig. 3. Table 2 is an abbreviated version of the final table (Table 3). It is used as an example both because it is simpler to compare three models than seven and because a comparison of Tables 2 and 3 will illustrate the value of testing multiple models. Appendix 1 summarizes the steps used to generate Tables 2 and 3. Additional details can be found in Anderson (2008) and Burnham and Anderson (2002).

The values in Table 2 were produced by the R package AICcmodavg, version 2.1-1, and checked by hand, but they can all be calculated by hand or spreadsheet after retrieving the RSS, log-likelihood, or AIC from a statistical package. The “ΔAICc” column shows the difference between the smallest AICc (that for the exponential in this example) and the other models. For any particular model,

$$\Delta AICc = AICc - AICc_{min} \quad (8)$$

The bolded columns in Table 2 are unaffected by the addition of new models. The ΔAICc column and other columns are not bolded because they change when new models are added. ΔAICc shows the distance between a particular model and the best model in the candidate set (Row 1). Because it is measured on a ratio scale in this example, where the truth is known, a ΔAICc of 8 is twice as far from the truth as a distance of 4. When the truth

Table 3. A Comparison of All Seven Models

Model Name	RSS	K	Log-Lik	AICc	ΔAICc	Evidence Ratio	AICc Weight	RSQ	Model Rank
Hyperbolic+Intercept	0.011	3	22.379	-35.33	0	1.00000	0.848	0.986	1
Hyperboloid+Intercept	0.010	4	22.699	-30.73	4.599	0.10000	0.085	0.987	2
Hyperbolic	0.026	3	17.726	-29.95	5.379	0.06800	0.058	0.968	3
Hyperboloid	0.026	2	17.753	-26.08	9.254	0.00980	0.008	0.968	4
Exponential	0.040	3	15.324	-21.22	14.110	0.00086	0.001	0.950	5
Mean	0.797	2	-1.170	0.34	35.670	0.00000	0.000	0.000	6
Linear	0.359	3	3.213	3.00	38.333	0.00000	0.000	0.549	7

Note: Columns as in Table 2. Bolded terms do not depend on the models included in the set of comparisons

is unknown the scale changes to interval, as described below. Either way, in the context of ΔAICc , small is good because it represents the distance between the model and the truth.

AICc is tethered to the dependent variable that generated it, so its value is only applicable to that variable in a specific data set. In general, the distances cannot be compared across data sets, though some approaches for doing so under certain conditions are described below (“Combining Datasets”). AICcs cannot even be compared across transforms of a dependent variable. Different AICcs must be calculated if, for example, a log transform is applied to the dependent variable because that would produce a different sum of squares. The log of an *independent* variable is simply a different model, however, so if the log of Delay were used as a model it could be added to the models listed in Table 2 because it does not change the dependent variable.

The absolute values of AICcs are immaterial in making these comparisons: only the difference matters. It is similar to a comparison of the height of an adult and a child standing on the seashore. To effect this comparison, one might determine the height above sea level of each person and subtract the two. The base is 0 (sea level), the adult’s head might be 6 feet above sea level and the child’s might be 4 feet. If the two people travel to Denver, Colorado, then the adult’s head might be 5,286 feet above sea level and the child’s 5,284 feet above sea level. The difference in height is still 2 feet regardless of the absolute value of the two numbers.

Evidence Ratio

The “Evidence Ratio” column is the probability that a given model is better than the best model. Any two models can be compared, but normally the comparison is only made against the best model in the candidate set, which is the one with the smallest AICc (or largest LL). Note that the AICc can be negative because, for example, the log of dependent variables between 0 and 1 is negative. The full number line must be kept in mind when comparing AICcs: -4 is smaller than -2.

In Table 2, the exponential model has the smallest AICc so the evidence ratio is 1.0, which is just the probability that the exponential model is as good as itself. The probability that the linear model is as good as the exponential model rounds to 0.00002 and the probability that the mean is the best model rounds to 0.000054. These numbers are given by Eq. 9 (Anderson, 2008; Burnham & Anderson, 2002):

$$\text{Evidence Ratio} = e^{-0.5(\Delta\text{AICc})} \quad (9)$$

This term is called an evidence ratio because it is the ratio of two likelihoods, that of a model under consideration to that of the best model yielding the evidence that the model in the numerator is better than that in the denominator. An important point is that the evidence ratios depend on the model producing the smallest AICc, $\text{AICc}_{\min}$, so they depend on what models are included in the comparison table.

The importance of the K parameter can be seen in the observation that the evidence ratio (and AICc weight) for the null model (mean) is actually better than for the linear model, even though the null model accounts for 0% of the variance. The parameter K exerts a penalty for the slope term in the linear model, making that even worse than no model at all for this dataset.

A rule of thumb has been offered in interpreting the ΔAICc and corresponding evidence ratio (Anderson, 2008; Burnham & Anderson, 2002). Call the best model “Model A.” The comparison models are then pitted against Model A. When the ΔAICc ranges from 0 to 2 (corresponding to an evidence ratio of 1 to 0.37), the comparison model is considered to have substantial support, meaning that it is not distinguishable from Model A. A ΔAICc of 4 to 7 (corresponding evidence ratio of 0.14 to 0.03) provides less support, and a ΔAICc of greater than 10 (corresponding evidence ratio of 0.0067) provides essentially no support for that comparison model because it is highly unlikely to be better than Model A. These values are deliberately provided as ranges separated by gaps to discourage the treatment of these values as bright lines in the way that $\alpha = 0.05$ value has come to be treated.

“Truth Drops Out as a Constant”

Table 2 compares three models against the best one, but what has happened to the true model? Recall that K-L information can be viewed as the information that is lost when using a model to represent reality, so it is a distance from the model under examination to the truth. Consider what happens when we subtract the information lost in Model 2 from that in Model 1.

$$\begin{aligned} \text{Information Lost Using Model 1} - \text{Information Lost Using Model 2} = \\ (\text{Truth} - \text{Estimate}_1) - (\text{Truth} - \text{Estimate}_2) = \\ (\text{Truth} - \text{Truth}) + (\text{Estimate}_2 - \text{Estimate}_1) = \\ \text{Estimate}_2 - \text{Estimate}_1 \end{aligned} \quad (10)$$

The truth does not change from model to model so it drops out when comparing models or, as noted by Burnham and Anderson (2002, p. 58) in a memorable phrase, “Truth . . . drops out as a constant.” This result enables us to conduct multiple model comparisons in the realistic situation in which we do not know what the truth is. We only care which model minimizes the information lost, i.e., which is closest to the truth, even if we cannot know just how much information was actually lost. By the way, the transformation in (9) changes the scale of information from ratio to interval.

AICc Weight

The “AICc Weight” column will form the basis for the model comparisons. Its calculation is quite simple:

$$\text{AICc Weight} = \frac{e^{-0.5(\Delta\text{AICc})_i}}{\sum_{j=1}^n e^{-0.5(\Delta\text{AICc})_j}} = \frac{\text{Model Likelihood for } i^{\text{th}} \text{ model}}{\text{Sum of All Model Likelihoods}} \quad (11)$$

In short, The AICc weight is the evidence ratio for a particular model divided by the sum of all the evidence ratios. Note that the sum of the AICc Weight column is 1.0, which is one of the characteristics that allows this term to be interpreted as a probability.

Its interpretation is that the number is the AICc weight is the probability that the corresponding model is the best in the candidate set.

The AICc weight serves two functions. The first flows from its interpretation as the probability that a model is the best model of those in the candidate set given the data. Thus, in Table 2, the exponential model, with an AICc weight of 0.999, has a 99.9% chance of being the best model of those in Table 2. The probability that the others are the best model is essentially 0. A better model might exist, but it cannot be assessed if it is not included. A corollary is that the truth is not necessarily, and almost certainly is not, among the set of models being compared, if only because the truth contains not just the process generating order in the data but also the processes that gives rise to “noise.”

The second function of the AICc Weight will be examined again below in the section “Multimodel Inference.” In brief, it is a weight that can be applied to each model if one is interested in combining them to predict the outcome for a specific value of the independent variable. For example, all the models can be combined, with each model weighted by the AICc weight, to make a prediction that is better than that from any single model.

The penultimate column of Table 2 is the R^2 , or variance accounted for (here, because most models are nonlinear, a pseudo R^2 is estimated). This column is included only for reference because, as a measure of model comparison, it is decidedly poorer than the AICc Weight, as will become clear below. The last column is just the rank order of the models based on the AICc.

Comparison of Seven Models

We now add four models to the candidate set. Figure 4 (bottom right) shows fits for one of these, the hyperbolic + intercept model. Table 3 contains a quantitative comparison of the seven models listed in Table 1. Note that the RSS, Log Likelihood, K, AICc, and R^2 are the same value as those listed in Table 2. This is because they derive from the residuals, which do not change. The values in the remaining columns are changed by the addition of new candidates because they derive from comparisons against the best model. Take, for example, the exponential model. When compared against two poor models (Table 2), the likelihood that it is the best candidate of the three is 99.9%. However, the addition of the models nested under the hyperboloid reduces this likelihood to 9.2% (Table 3). This 10-fold difference illustrates the way in which the quality of each model is substantially influenced by the models against which it is being compared, and drives home the importance of candidate selection at the beginning of the process.

The best model in Table 3 is the hyperbolic + intercept model in which both the Y intercept and discounting parameter (k) are estimated. This model has a 0.848 chance of being the best of the seven models in Table 3 for this dataset. At 0.085, the hyperboloid + intercept model, in which the intercept is estimated, is second. The interpretation of the outcome of this single experiment would be that a model in which the intercept is estimated model is likely to be the best model for this data set, with an 0.933 probability ($0.848 + 0.085$). As a reminder, we are saying nothing here about delayed discounting in the real world, we are simply seeking to model a known a hyperbolic model with an intercept of 0.85 and a discounting rate of 0.2.

The last column, R^2 , shows that the proportion of variance accounted for differs very little among the top models. The uniformity of R^2 values across models, even the exponential model, in comparison to the variability in weighted AICc, demonstrates weakness of R^2 as an indicator of model quality.

Confirmation of the Model Probabilities

The premise of the model-comparison approach is that the AICc weight describes the probability that a model is the best of the candidate set. The previous exercise applied seven models to a single noisy data set in order to determine which model is most likely to be closest to the truth, which is Eq. 3. The hyperbolic + intercept model had the highest AICc Weight, meaning that it is probably closest to the truth, i.e., that it suffers the least loss of all the information contained in the true set of influences that generated the data. Two others, the hyperboloid + intercept and the hyperbolic models, were indistinguishable from one another and, although less likely than the top model, were still not unreasonable possibilities.

To examine more rigorously the claim that the AICc Weight approximates the probability that a model is closest to the truth, 10,000 data sets were generated and examined using model-comparison. Each model in Table 3 was fit to each data set, for a total of 70,000 fits. The data sets were generated in the same fashion as described in Eq. 3: random noise with a mean of zero and a standard deviation of 0.05 was applied to Eq. 2 with the stipulation that Y , the choice of the LL option, must be less than or equal to 1 and greater than or equal to 0. The data sets differed only in the process that produced the noise: the same delay-discounting process was always present. Each run mimics a situation in which a natural process (a hyperbolic + intercept discounting function) combines with the multitude of events that produce the noise. On occasion, a data set could be generated by chance that is best fit by some other function so this exercise exemplifies the estimation of the discounting process and the sundry events that influence the real data collected. This exercise also provides evidence about the reproducibility of our single virtual experiment.

For each run of the data, the seven models were rank-ordered as in Table 3 and the number of times that a model was ranked as best was tabulated. If the AICc Weights in Table 3 reflect the probability that a model is the best model then the tabulation should approximate the AICc weights. That is, the hyperbolic + intercept model should be the best on about 0.848 of the runs, or approximately 8,480 of the 10,000 runs. By similar calculations, the hyperboloid + intercept model should be the best on approximately 850 of the trials and the hyperbolic model should outperform all the others on approximately 580 of the data sets examined.

The results of the simulations are summarized in the columns under “One Replicate” in Table 4. The hyperbolic + intercept model was the best model on 0.846 of the simulations, as compared with an AICc Weight of 0.848 (it is a coincidence that the AICc weight of this particular run so closely matches the actual proportion). The hyperbolic model was the best on 0.075 of the runs (AICc Weight of 0.0576) and, on rare occasions, a data set occurred for which the hyperboloid model was the best. The ranking of models obtained from the single example with 11 data points listed in the third column of Table 4 approximated the ranking determined by the 10,000 simulations. The conclusion that the hyperbolic and hyperboloid models are indistinguishable is consistent with the

observation that their order in Table 3 was flipped in the simulation. The linear and mean models were never selected in the simulation. In addition, the nonlinear least-squares estimate did not converge on about 0.05% of the cases (about 5 times in 10,000 runs). Inspection of those failures revealed that the data happened to be structured such that an exponential model fit well.

The best-model information in Table 4 provides only a glimpse of the ability of the I-T approach at identifying the correct model. More complete information can be found in frequency distributions in Appendix 2.

Why wasn't the hyperbolic + intercept identified on 100% of the cases? The reason is that it was not the best model on 100% of the cases. In the simulation, the truth was a hyperbolic + intercept discounting function plus other factors, captured by noise with a known mean and standard deviation. On some cases, the extraneous factors combined in such a way that the raw data were actually best fit by one of the other models, but this happened relatively rarely.

The right columns show what happens when a single replicate is added. Here, two estimates were taken at each delay interval but otherwise the simulation was conducted in the same way as in the left columns. Adding a second replicate greatly reduced the contribution of extraneous factors, made the probability estimates in the AICc weights more accurate, and eliminated the nonconvergence problem—all nonlinear least-squares estimates converged. The issue, then, is not that the model failed to identify the generating function but rather that chance events conspired to produce something different, and the impact of that conspiracy was diminished by adding only a single replicate. Additional detail about this exercise can be found in Fig. 6 in the Appendix.

Coefficient Estimates

A key goal of modeling data is the identification of important characteristics of a phenomenon, including parameter values like the discounting rate or preference at a 0 delay. To see how model-selection can help accomplish this goal, coefficient estimates

Table 4. Number of Times Each Model was the Most Probable out of 10,000 Runs with 1 Data-Point per Run

Model name	One Replicate			Two Replicates		
	Frequency	Proportion	AICc Weight ¹	Frequency	Proportion	AICc Weight ¹
Hyperbolic + Intercept	8459	0.846	0.8480	8821	0.8820	0.8030
Hyperbolic	750	0.075	0.0576	5	0.0005	0.0006
Hyperboloid + Intercept	377	0.038	0.0851	1167	0.1167	0.1950
Hyperboloid	211	0.021	0.0083	5	0.0005	0.0003
Exponential	203	0.020	0.0007	2	0.0002	0.0009
Mean	0	0.000	0.0000	0	0.0000	0.0000
Linear	0	0.000	0.0000	0	0.0000	0.0000

¹ AICc Weights came from a single, arbitrarily selected run and are used to illustrate the correspondence between AICc weight and obtained proportion

from each of the 10,000 runs are presented in histograms (Fig. 5). For simplicity, only models nested under the hyperboloid + intercept equation (Table 1) are examined.

The top row of Fig. 5 shows estimates from the hyperbolic + intercept function, Model 6, and the best model in Table 4. The histograms show the distribution of parameter estimates across the 10,000 simulated datasets. The vertical black lines show the true value of k (0.2) and the intercept (0.85). For an overwhelming majority of the data sets generated, the values of k and Y_0 were very close to that of the generating function. The second row shows the distributions from the hyperbolic function in which the intercept was forced to be 1.0. When the intercept is not included, the parameter estimates are biased toward longer values, indicating a higher discounting rate, a bias toward “impulsivity.” This is seen in both the hyperbolic (Row 2) and hyperboloid (Row 4) models. The bottom two rows show the two hyperboloid models, in which a power term is included (Models 1 and 2). Both models that included an intercept converged on good model parameters, even though the hyperboloid +

Fig. 5 Frequency histograms showing the parameter estimates (across columns) for each of four hyperboloid-based discounting models (down rows) after 10,000 simulations. Elements are empty when a model does not include a parameter. The vertical black line shows the true parameter value from Eq. 2. Note log scaling of the horizontal axis. The horizontal axes are truncated at 0.1 and 10, making rare excessive values are not visible

intercept model had a low model probability (Table 3, Fig. 5). The variability in the estimate of k was quite high for both hyperboloid models.

The data presented in Fig. 5 are informative in at least three ways. First, Fig. 5 confirms that parameter values estimated using the model that is closest to the truth has the least variability and the most precise estimate of parameters. This is seen in the relatively low spread for the hyperbolic + intercept model in the top row and in the fact that they are centered on the parameters that generated the data. Second, the importance of other considerations in model building, such as identifying correlations among covariates, is also seen. Models that omit the intercept produce a heavily biased estimate of k . This is related to a high correlation (0.74 in this exercise) between these two parameters. If the intercept is omitted then k picks up the slack (Young, 2017). Third, and more important for our purposes, the data presented in Table 3 and Fig. 5 go a step further in validating the use of the AICc approach to model selection by showing the accuracy of the best model in estimating important parameters.

In this tutorial, the true model that produced each of the 10,000 data sets is known, allowing us to validate the extent to which the AICc approach is capable of identifying the best model given the data. In reality, we do not know what the truth is, but the foregoing analyses indicate that we can at least locate the models that come closest to it using the techniques described above.

Multimodel Inference

One goal of modeling is to determine the structure of the process that produced the data, as illustrated above. A second goal is to predict what will happen under a certain set of conditions, or to estimate a particular coefficient, such as a matching-law or delay-discounting parameter under different conditions. A common approach is to use the estimate from the best model overall or on an individual-subject level, though such aggregation can result in an inferior prediction (Franck et al., 2015). As the I-T approach has made clear, there is information in the second-best model, as well as in all the other models examined. This information can be used to form a better prediction or to improve the estimate of a particular parameter.

Weighting the predictions of the dependent variable from each model and then combining the weighted predictions increases the precision of the overall estimate (Burnham & Anderson, 2002). Suppose, for example, one examined three linear models on a data set and the model comparisons yielded an AICc Weight of 0.6 for the first model, 0.3 for the second, and 0.1 for the third. A combined model in which the predictions from the three models were weighted by 0.6, 0.3, and 0.1, respectively, will result in a better prediction than from any model alone. In the example above, to predict a discounting function after examining the possible models, the prediction from the hyperbolic + intercept model using one subject would be weighted by 0.848 and that from the hyperboloid + intercept model by 0.085. The hyperbolic, hyperboloid, and exponential models' predictions would be weighted by 0.058, 0.008, and 0.001, respectively. A small contribution from a low-ranking model reflects the fact that it carries little information. By incorporating all of the information available the predictions are more precise (Burnham & Anderson, 2002) as long as the coefficients in the model and not correlated (Cade, 2015). Multicollinearity can present problems in

forming these predictions (Cade, 2015), pointing to the importance of finding the best model. The problem posed by correlated variables is evident in the impact of omitting an important term on the coefficient estimates in Fig. 5.

Discrete Predictors and the Analysis of Variance

The analysis of variance (ANOVA) is an extension of the general linear model so any ANOVA can be rewritten as a regression model, with discrete (0 or 1 as the coefficients) variables, or as a mix of continuous and discrete variables. With this understanding, it is possible to conduct an ANOVA using the I-T approach and the results will have the attractive attributes described earlier (Burnham, Anderson, & Huyvaert, 2011). As with the approach taken here, the point of departure is the residual sum of squares or the AIC, but the latter can be used directly from software only if the parameter count includes the variance. Recall that k for the null model is 2 (not 1 as in ANOVA) because it includes an estimate of both the mean (Y_0) and the variance. From there, add 1 for each additional parameter being estimated. For example, in a study of the impact of adolescent exposure to the environmental contaminant methylmercury on adult behavior there was interest in which exposure level produced the major effects on delay discounting (Boomhower & Newland, 2016).

Use in Behavior Science

The growing trend in the use of the I-T data analysis in selected behavioral journals was shown in Fig. 1. Here, we describe some of the ways that it is implemented. All of the examples in the *Journal of the Experimental Analysis of Behavior*, *Journal of Experimental Psychology: Animal Behavior Processes*, and *Journal of Experimental Psychology: Animal Cognition and Learning* were examined, as were selected papers in the other journals. Some variant of I-T analysis was used to examine models of choice or response strength (Cowie, Davison, & Elliffe, 2014; Hunter & Davison, 1982; Hutsell & Jacobs, 2013; Lau & Glimcher, 2005; MacDonall, 2009; McLean, Grace, & Nevin, 2012; Mitchell, Wilson, & Karalunas, 2015; Navakatikyan, 2007), discounting (Białaszek, Marcowski, & Ostaszewski, 2017; DeHart & Odum, 2015; Jarmolowicz et al., 2018; Rung & Young, 2015; Young, 2017), peak shift (Krägeloh, Elliffe, & Davison, 2006), behavioral momentum (Hall, Smith, & Wynne, 2015; Killeen & Nevin, 2018), associative learning (Cabrera, Sanabria, Shelley, & Killeen, 2009; Hall et al., 2015; Witnauer, Hutchings, & Miller, 2017), the partitioning of behavior into response bouts (Brackney, Cheung, Neisewander, & Sanabria, 2011; Smith, McLean, Shull, Hughes, & Pitts, 2014; Tanno, 2016), navigation strategies (Anselme, Otto, & Güntürkün, 2018), instructional techniques (Warnakulasooriya, Palazzo, & Pritchard, 2007), problem solving by corvids (Cibulski, Wascher, Weiss, & Kotrschal, 2014), models of punishment (Klapes, Riley, & McDowell, 2018), and timing (Beckmann & Young, 2009; Ludvig, Conover, & Shizgal, 2007). This is just a partial, not exhaustive, list to convey the range of applications.

Some investigators used the AIC or the AICc simply to identify the best model to use when examining a particular phenomenon, so this was a relatively minor part of the article. These papers frequently used the AIC or AICc as one might use a p value: in

order identify the best way to describe a data set or to select a specific model for further examination. In some cases, the AIC or AICc was used almost in passing and its description omitted important details, such as the value used for k , how multiple subjects were pooled, or the difference between the top two models. Sometimes the model with the smallest AIC was selected even though the second best differed by only one or a few units and in some examples the actual AIC, much less the AIC differences (ΔAIC), were not provided.

For some papers (Brackney et al., 2011; Klapes et al., 2018; Navakatikyan, 2007; Navakatikyan & Davison, 2010; Smith et al., 2014; Young, 2017), the model-comparison was a core component of a study that was designed to compare different models of a phenomenon using either data collected by the investigator or data shared from authors of other papers. In these, model selection showed the plausibility of different theory-driven quantitative models.

For example, Hunter and Davison, in the first paper identified and published two decades before the Burnham and Anderson book, compared six quantitative models of how response force and reinforcer ratios interact to determine choice (Hunter & Davison, 1982). An assessment of eight models of punishment using a data set provided by another investigator was undertaken by (Klapes et al., 2018). In this, the key values, except AICc weight, which could easily be determined from the information in the paper, were published in a comprehensive table summarized the ability of the models to characterize punishment in concurrent schedules. Brackney et al. (2011) examined 16 different models of response-bout analyses to determine the dynamics of how different interventions (free food, extinction) alter the microstructure of behavior. It is not surprising that the r -squared values provided no firm guidance about model selection. One study was located (Cabrera et al., 2009) that provided the comprehensive set AICc weights as described in Table 3.

Combining Datasets

The I-T approach to analysis was developed to be conducted on a single data-set and only for a single dependent variable (an approach to combining data sets is described at the end of this section.) This is because a new data set or a new dependent variable on the same data set generates a new sum of squares and thus completely novel estimates of AIC. Therefore, the question of how to combine analyses from different subjects arises. The tactics below are restricted to situations in which the same dependent variable (e.g., proportion of “Left” responses) is used for multiple subjects.

The best approach to combining datasets from different individuals is to begin with nonlinear mixed-effects (NLME) models or, if the model can be linearized, linear mixed-effects models. This approach allows the analyst to incorporate the full model of behavior with terms to represent specific factors, such as subject demographics or interventions, and the random selection of subjects. It also uses information from all subjects to estimate model parameters (Young, 2017). For example, 14 models of temporal discrimination were examined in humans in order to determine the role of clock speed in scalar expectancy theory and the behavior theory of timing (Beckmann & Young, 2009). In another example, the concatenated matching law formed the basis for describing the impact of methylmercury exposure on delay discounting in adults

(Boomhower & Newland, 2016). Models were formed from different combinations of terms for magnitude sensitivity, delay sensitivity, dose, and whether individual mice were included as fixed or random effects. The best 7 of 25 models were shown in a comprehensive table to identify the key terms that were affected by exposure and the most sensitive dose.

A second approach is to model each subject individually and present the full set of best models for each individual, often showing that the same model did not fit each individual. For example, to examine what probability distribution to use in analyzing the bout structure of behavior, Tanno (2016) examined 30 different models. A key component of the analysis was the presentation of Δ AICs for each model as applied to each of eight rats. This has the advantage of capturing the different response patterns produced by the animals and it revealed that a single model did not describe all rats. In another, Berg and Grace (2011) show an example with multidimensional scaling and a study investigating models of timing (Ludvig et al., 2007) contains a graphical approach to comparing individual AICs. A somewhat simplified version of this second approach is exemplified in another study of bout structure (Brackney et al., 2011), who conducted an analysis of four models of bout structure and reported the minimum daily Δ AIC in a relatively simple display. This has the advantage of simplifying the message of the paper but at the expense of obscuring individual response patterns.

In another approach to examining individuals, data collected from human participants and simulations were used to compare approaches to using a single parameter, the delay that produced a 50% reduction in value, to characterize individual delay discounting functions (Franck et al., 2015). In this study the same type of function could not be used for all participants. To identify the best model for an individual, the investigators used a Bayesian Information Criterion, which is more conservative than the AICc, but the idea is similar. Like Tanner, they reported heterogeneity: no single model fit all participants. In a simulation, in which the generating model and its parameters were known, Franck et al. generated a population of virtual individuals containing a mix of different discounting strategies, as in the population actually studied. They this simulation to compare different approaches to estimating the delay at which a reinforcer was discounted by 50% (the ED-50 or effective delay for a 50% decline) in a population. In one approach, the best overall model was applied to each individual and in a second approach the prediction from each individual was made by that person's best model. Individualizing the model was superior in estimating the known ED-50 in the heterogeneous population. (As an aside, the multimodel inference strategy described above would be well suited for such a heterogeneous situation.)

Sometimes the model-comparison approach is applied to results taken from groups of individuals. This is frequently seen in the ecology literature, where the impact of environmental parameters on measures of success by a species might be applied. As an example in the behavioral literature, Sanabria and Killeen (2008) examined timing in two rodent strains to test models of timing in a study designed to distinguish impulsivity from motor and motivational influences. To do this, they generated curves representing each strain and fit models to those curves.

Finally, an interesting approach has been referred to in the peer-reviewed literature (Klapes et al., 2018; Navakatikyan, 2007; Navakatikyan & Davison, 2010). It has not itself been published in the peer-reviewed literature but is available on Research Gate (McArdle, Navakatikyan, & Davison, 2019). The McArdle approach can be applied

when the same dependent variable is examined across data sets produced by individuals, and therefore it produces an experiment-wide AICc. In brief, this approach sums the model likelihoods from all subjects and adds a penalty term that incorporates the number of subjects, the number of parameters estimated, and the number of data points that are used in each estimate. For this approach to work best there should be some degree of homogeneity in the ranking of models across individuals. It is possible, however, that one model in the candidate set will be the best for some subjects and another will be the best for other subjects, as described in Tanno (2016) and Franck et al. (2015). The degree to which differences in individual subjects influence the model selection criteria, the ΔAICc s, is not known at present. It seems plausible that if there is widespread disagreement then the ΔAICc s will be too small to select a single model with confidence, but further research with this approach is required.

Precautions

Any statistical approach can be misused so some precautionary comments are warranted (see Anderson & Burnham, 2002, for a comprehensive list). The first issue is scientific. Candidate models must have some empirical or theoretical foundation and should be posited in advance. The examples used here include models that have been explored extensively in the delayed discounting literature and each model, except the null and the linear models, has some theoretical foundation. Additional models could, and often should, be tested with a larger data set. Models that have no rationale behind them—for example, a fourth-order polynomial—could be tested here, but that endeavor would yield nothing of lasting interest. A corollary of this caution is that testing *all* nested subsets of a complex model is often ill-advised, even though software sometimes makes it easy to do this, unless there is a defensible reason for each model.

It must always be remembered that the best model in a set is just that—the best model among those selected for examination. The true model cannot be known, if only because the data are often a function of both an underlying process and variability due to other processes in play at the moment of data collection. The importance of the set of candidate models is evident in comparing Tables 2 and 3. The exponential model was unequivocally the best of the models contained in Table 2, but it barely registered in the larger set of candidate models in Table 3.

Finally, with the I-T approach the investigator can rank order a priori models and estimate the probability that each is the best. To corrupt this approach by turning it into a new statistical test to reject or accept models would undermine the power of this approach as well as the theory behind it. Accordingly, no bright-line criteria have been offered for selecting or rejecting models, and combining this with a NHST approach is discouraged.

Concluding Remarks

The I-T approach builds on likelihood estimates and yields an assessment of a collection of models that might be applied to a data set. It offers numerous advantages. It is simple to apply, requiring only calculations on a spreadsheet. It provides the

probability that a particular model in a set is the best for the data collected, which is what an investigator wishes to know when analyzing data. In addition, unlike NHST, it encourages the testing of multiple models and provides a framework within which to test them. Finally, it helps address the replicability problem as the probabilities that it provides are more closely linked to the likelihood of replication than NHST.

With NHST, one can never accept a hypothesis, only reject it (Killeen, 2005). In contrast, The I-T approach, by identifying plausible models and hypotheses, contributes a *cumulative* science. Theoretical models can be posed, tested, and selected based on criteria that point to the probability of replication. The best models will ultimately be selected by the scientific community.

Acknowledgements Special thanks to Kelly Banna, Alejandro Lazarte, and Dalisa Kendricks for helpful comments on earlier versions.
Supported by ES 024845 from NIH.

Appendix 1

Table 5 Propagation of Calculations for determining the AICc Weight/ Model Probability

Model Name	RSS	K	LL	AICc ¹	Delta AICc	Evidence Ratio (M)	AICc Weight
Model 1	From software or hand-calculated	1 + parameter estimated	$-\frac{n}{2} \log \frac{RSS}{n}$	$-2LL + 2K + \frac{2K(K+1)}{N-K-1}$ OR $\approx N \log \left(\frac{s^2}{\sigma^2} \right) + \frac{2K(K+1)}{N-K-1}$	AICc – MIN (AICc)	$e^{-0.5(\Delta AICc)}$	$\frac{M}{\sum M_i}$
model 2	All entries calculated as in Row 1						
model 3	All entries calculated as in Row 1						

¹Where $=\sigma^2 = \sigma^2 \frac{N-2}{N}$

The calculation of the Model Probabilities (AICc Weights) can be calculated using a simple spreadsheet with formulas as shown below. One can begin with either the residual sum of squares (RSS), the log likelihood (LogLik), the variance ($\sigma^2 = \frac{RSS}{N}$), the AIC or the AICc, depending on what is provided by the software. The term that is used determines how far to the right in the table one begins the calculation. Note that K is 1 + the degrees of freedom because the estimate of the AICc includes an additional parameter for the variance. If using AICc or AIC from a software package, check that it uses the correct value for K. If estimating the AICc from the RSS or variance directly then the estimate in the table’s footnote can be used.

The following steps can be used when conducting an I-T analysis:

1. Collect data.
2. Analyze the data using an analysis appropriate to the data set such as linear or nonlinear regression, mixed-effects modeling, or analyses of variance.
3. Extract from that analysis one of the following parameters: residual sum of squares (RSS), likelihood, log likelihood, AIC, or AICc.

- a. If the RSS is used then use Table A1 to estimate the AICc.
- b. If the likelihood is used then take the log-likelihood. Any base can be used as long as it is consistent, but typically the base is the natural logarithm, e . Then proceed from the fourth column of Table A1
- c. If the AIC is extracted then add the correction factor $2K(K+1)/N-K-1$ and proceed from the fifth column of Table A1.
4. Verify that the appropriate K is used. This will usually be the number of regression parameters estimated plus 1 for the variance and 1 for an intercept (if that is not already included).
5. Construct a table similar to Table A1.
6. Locate the smallest AICc and it in Row 1 of the table (for convenience).
7. From the AICc column, construct the Delta AICc column by subtracting each AICc from the smallest one.
8. Calculate the evidence ratios by taking $\exp(-0.5 (AICc - AICc_{MIN}))$. This is the probability that the model is at least as good as the one with the smallest AICc. Note that the evidence ratio of the best model will be 1.0 because the probability that it is at least as good is 1.0.
9. Calculate the AICc Weights by dividing the evidence ratio (Column 7) by the sum of all the evidence ratios.
10. Test to be sure that the sum of the AICc Weights equals 1.0. Each AICc weight is the probability that each model is the best of the candidate set.

Appendix 2

Detailed Example of AICc Performance from Simulations

The data in Table 4 are single-point estimates of the percentage of times that each model was ranked as best but this ordinal measure is inadequate in providing a full representation of the I-T approach because an AICc can be the best by a small or a large amount. The histograms in Fig. 6 illustrate the distribution of actual model probabilities (AICc weights) from these 10,000 data sets and how adding a second replicate dramatically changes the ability of the AICc to converge on good models (Row 2). With only one replicate per fit (top row), the AICc Weights for the hyperbolic + intercept model was heavily left-skewed, with most weights greater than 0.75, meaning that this model was selected as having at least 75% chance of being the best model on most runs. The hyperboloid + intercept and the hyperbolic models overlapped considerably, consistent with their switching places from Table 4. The hyperboloid and exponential models fared more poorly. The exponential model rarely had an AICc greater than 0.25.

Row 2 shows that merely adding a second estimate to a single discounting function sharpened the discriminability of the AICc weight. The hyperbolic + intercept model was showed a sharper right peak but the three poorest models' left peak was also much sharper, indicating that their AICc weights were usually less than 0.05.

The proportions (third and sixth columns) and model probabilities (fourth and seventh columns) in Table 4 lie are consistent with the histograms in Fig. 6. The

Fig. 6 Distribution of AICc-Weight (model probabilities) for five model types. Fits were constructed using one (top row) or two (bottom) replicates per delay interval. Each horizontal axis can be interpreted as the probability that a model is the best of the seven models examined in Table 3. Note that the hyperboloid + intercept model was usually the most probable model of the data set but that for some data sets other models were ranked as the most probable.

hyperbolic + intercept model was most frequently selected as the most probable one across the 10,000 different data sets but the outliers indicate that on rare occasions a data set was generated for which one of the other models, exclusive of the linear one, was deemed probable. Thus, a single experiment stands a good chance of predicting model probabilities and a replication that produces similar values will provide even greater confidence in the conclusions. The linear (and mean, not shown) models were never selected as a good model. These analyses demonstrate the strength of the AICc approach because in most cases, the highest ranked model is the model that was actually used to generate the data sets.

References

- Ainslie, G., & Monterosso, J. R. (2003). Building blocks of self-control: increased tolerance for delay with bundled rewards. *Journal of the Experimental Analysis of Behavior*, 79(1), 37–48.
- Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Caski (Eds.), *Second international symposium on information theory* (pp. 267–281). Budapest: Akademiai Kiado.
- Anderson, D. R. (2008). *Model based inference in the life sciences: A primer on evidence*. New York: Springer.

- Anderson, D. R., & Burnham, K. P. (2002). Avoiding pitfalls when using information: Theoretic methods. *Journal of Wildlife Management*, 66(3), 912–918. <https://doi.org/10.2307/3803155>.
- Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000). Null hypothesis testing: Problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64(4), 912–923.
- Anselme, P., Otto, T., & Güntürkün, O. (2018). Foraging motivation favors the occurrence of Lévy walks. *Behavioural Processes*, 147, 48–60. <https://doi.org/10.1016/j.beproc.2017.12.014>.
- Baer, D. M., Wolf, M. M., & Risley, T. R. (1968). Some current dimensions of applied behavior analysis 1. *Journal of Applied Behavior Analysis*, 1(1), 91–97.
- Beckmann, J. S., & Young, M. E. (2009). Stimulus dynamics and temporal discrimination: Implications for pacemakers. *Journal of Experimental Psychology: Animal Behavior Processes*, 35(4), 525–537. <https://doi.org/10.1037/a0015891>.
- Berg, M. E., & Grace, R. C. (2011). Categorization of multidimensional stimuli by pigeons. *Journal of the Experimental Analysis of Behavior*, 95(3), 305–326. <https://doi.org/10.1901/jeab.2010.94.305>.
- Białaszek, W., Marcowski, P., & Ostaszewski, P. (2017). Physical and cognitive effort discounting across different reward magnitudes: Tests of discounting models. *PLoS One*, 12(7), e0182353–e0182353. <https://doi.org/10.1371/journal.pone.0182353>.
- Bickel, W. K., Pope, D. A., Moody, L. N., Snider, S. E., Athamneh, L. N., Stein, J. S., & Mellis, A. M. (2016). Decision-based disorders: The challenge of dysfunctional health behavior and the need for a science of behavior change. *Policy Insights from the Behavioral and Brain Sciences*, 4(1), 49–56. <https://doi.org/10.1177/2372732216686085>.
- Boomhower, S. R., & Newland, M. C. (2016). Adolescent methylmercury exposure affects choice and delay discounting in mice. *Neurotoxicology*, 57, 136–144. <https://doi.org/10.1016/j.neuro.2016.09.016>.
- Brackney, R. J., Cheung, T. H., Neisewander, J. L., & Sanabria, F. (2011). The isolation of motivational, motoric, and schedule effects on operant performance: A modeling approach. *Journal of the Experimental Analysis of Behavior*, 96(1), 17–38. <https://doi.org/10.1901/jeab.2011.96-17>.
- Branch, M. N. (1999). Statistical inference in behavior analysis: Some things significance testing does and does not do. *The Behavior Analyst*, 22(2), 87–92.
- Branch, M. N. (2019). The “reproducibility crisis”: Might the methods used frequently in behavior-analysis research help? *Perspectives on Behavior Science*, 42(1), 77–89. <https://doi.org/10.1007/s40614-018-0158-5>.
- Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference* (2nd ed.). New York: Springer.
- Burnham, K. P., Anderson, D. R., & Huyvaert, K. P. (2011). AIC model selection and multimodel inference in behavioral ecology: Some background, observations, and comparisons. *Behavioral Ecology & Sociobiology*, 65(1), 23–35. <https://doi.org/10.1007/s00265-010-1029-6>.
- Cabrera, F., Sanabria, F., Shelley, D., & Killeen, P. R. (2009). The “lunching” effect: Pigeons track motion towards food more than motion away from it. *Behavioural Processes*, 82(3), 229–235. <https://doi.org/10.1016/j.beproc.2009.06.010>.
- Cade, B. S. (2015). Model averaging and muddled multimodel inferences. *Ecology*, 96(9), 2370–2382.
- Chung, S. H., & Herrnstein, R. J. (1967). Choice and delay of reinforcement. *Journal of the Experimental Analysis of Behavior*, 10(1), 67–74.
- Cibulski, L., Wascher, C. A. F., Weiss, B. M., & Kotschal, K. (2014). Familiarity with the experimenter influences the performance of Common ravens (*Corvus corax*) and Carrion crows (*Corvus corone corone*) in cognitive tasks. *Behavioural Processes*, 103(100), 129–137. <https://doi.org/10.1016/j.beproc.2013.11.013>.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49, 997–1003.
- Cohen, N., Moynihan, J. A., & Ader, R. (1994). Pavlovian conditioning of the immune system. *International Archives of Allergy and Immunology*, 105(2), 101–106. <https://doi.org/10.1159/000236811>.
- Cowie, S., Davison, M., & Elliffe, D. (2014). A model for food and stimulus changes that signal time-based contingency changes. *Journal of the Experimental Analysis of Behavior*, 102(3), 289–310. <https://doi.org/10.1002/jeab.105>.
- Davidson, T. L., & Riley, A. L. (2015). Taste, sickness, and learning: understanding how we form aversions to particular flavors has led to new ideas about learning—and could have implications for treating obesity and drug use. *American Scientist*, 103(3), 204–212.
- DeHart, W. B., & Odum, A. L. (2015). The effects of the framing of time on delay discounting. *Journal of the Experimental Analysis of Behavior*, 103(1), 10–21. <https://doi.org/10.1002/jeab.125>.
- Depaoli, S., Rus, H. M., Clifton, J. P., van de Schoot, R., & Tiemensma, J. (2017). An introduction to Bayesian statistics in health psychology. *Health Psychology Review*, 11(3), 248–264. <https://doi.org/10.1080/17437199.2017.1343676>.

- Franck, C. T., Koffarnus, M. N., House, L. L., & Bickel, W. K. (2015). Accurate characterization of delay discounting: a multiple model approach using approximate Bayesian model selection and a unified discounting measure. *Journal of the Experimental Analysis of Behavior*, 103(1), 218–233. <https://doi.org/10.1002/jeab.128>.
- Freeman, K. B., & Riley, A. L. (2009). The origins of conditioned taste aversion learning: A historical analysis. In S. Reilly & T. R. Schachtman (Eds.), *Conditioned taste aversion: Behavioral and neural processes* (pp. 9–36). Oxford: Oxford University Press.
- Goodman, S. N. (1992). A comment on replication, P-values and evidence. *Statistics in Medicine*, 11(7), 875–879. <https://doi.org/10.1002/sim.4780110705>.
- Goodman, S. N. (2001). Of P-values and Bayes: A modest proposal. *Epidemiology*, 12(3), 295–297.
- Greenland, S. (2019). Valid P-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73(suppl), 106–114. <https://doi.org/10.1080/00031305.2018.1529625>.
- Hales, A. H., Wessellmann, E. D., & Hilgard, J. (2019). Improving psychological science through transparency and openness: An overview. *Perspectives on Behavior Science*, 42(1), 13–31. <https://doi.org/10.1007/s40614-018-00186-8>.
- Hall, N. J., Smith, D. W., & Wynne, C. D. L. (2015). Pavlovian conditioning enhances resistance to disruption of dogs performing an odor discrimination. *Journal of the Experimental Analysis of Behavior*, 103(3), 484–497. <https://doi.org/10.1002/jeab.151>.
- Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, 13(3), e1002106. <https://doi.org/10.1371/journal.pbio.1002106>.
- Higgins, S. T., Davis, D. R., & Kurti, A. N. (2017). Financial incentives for reducing smoking and promoting other health-related behavior change in vulnerable populations. *Policy Insights from the Behavioral & Brain Sciences*, 4(1), 33–40. <https://doi.org/10.1177/2372732216683518>.
- Hunter, I., & Davison, M. (1982). Independence of response force and reinforcement rate on concurrent variable-interval schedule performance. *Journal of the Experimental Analysis of Behavior*, 37(2), 183–197.
- Hutsell, B. A., & Jacobs, E. A. (2013). Attention and psychophysics in the development of stimulus control. *Journal of the Experimental Analysis of Behavior*, 100(3), 282–300. <https://doi.org/10.1002/jeab.54>.
- Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, 2(8), e124. <https://doi.org/10.1371/journal.pmed.0020124>.
- Jarmolowicz, D. P., Reed, D. D., Francisco, A. J., Bruce, J. M., Lemley, S. M., & Bruce, A. S. (2018). Modeling effects of risk and social distance on vaccination choice. *Journal of the Experimental Analysis of Behavior*, 110(1), 39–53. <https://doi.org/10.1002/jeab.438>.
- Johnston, J. M., & Pennypacker, H. S. (1993). *Strategies and tactics of behavioral research*. Hillsdale: Lawrence Erlbaum Associates.
- Kassin, S., Tubb, V. A., Hosch, M. H., & Memon, A. (2001). On the general acceptance of eyewitness testimony research: A new survey of the experts. *American Psychologist*, 56(5), 405–416.
- Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, 16(5), 345–353. <https://doi.org/10.1111/j.0956-7976.2005.01538.x>.
- Killeen, P. R. (2019). Predict, control, and replicate to understand: How statistics can foster the fundamental goals of science. *Perspectives on Behavior Science*, 42(1), 109–132. <https://doi.org/10.1007/s40614-018-0171-8>.
- Killeen, P. R., & Nevin, J. A. (2018). The basis of behavioral momentum in the nonlinearity of strength. *Journal of the Experimental Analysis of Behavior*, 109(1), 4–32. <https://doi.org/10.1002/jeab.304>.
- Klapes, B., Riley, S., & McDowell, J. J. (2018). Toward a contemporary quantitative model of punishment. *Journal of the Experimental Analysis of Behavior*, 109(2), 336–348. <https://doi.org/10.1002/jeab.317>.
- Kmetz, J. L. (2019). Correcting corrupt research: Recommendations for the profession to stop misuse of p-values. *The American Statistician*, 73(suppl), 36–45. <https://doi.org/10.1080/00031305.2018.1518271>.
- Krägeloh, C. U., Elliffe, D. M., & Davison, M. (2006). Contingency discriminability and peak shift in concurrent schedules. *Journal of the Experimental Analysis of Behavior*, 86(1), 11–30. <https://doi.org/10.1901/jeab.2006.11-05>.
- Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial & Special Education*, 34(1), 26–38.
- Kruschke, J. K. (2013). Bayesian estimation supersedes the t test. *Journal of Experimental Psychology: General*, 142(2), 573–603. <https://doi.org/10.1037/a0029146>.
- Laraway, S., Snyckerski, S., Pradhan, S., & Huitema, B. E. (2019). An overview of scientific reproducibility: Consideration of relevant issues for behavior science/analysis. *Perspectives on Behavior Science*, 42(1), 33–57. <https://doi.org/10.1007/s40614-019-00193-3>.

- Lau, B., & Glimcher, P. W. (2005). Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the Experimental Analysis of Behavior*, 84(3), 555–579. <https://doi.org/10.1901/jeab.2005.110-04>.
- Leek, J., McShane, B. B., Gelman, A., Colquhoun, D., Nuijten, M. B., & Goodman, S. N. (2017). Five ways to fix statistics. *Nature*, 551(7682), 557–559. <https://doi.org/10.1038/d41586-017-07522-z>.
- Loftus, E. F. (1993). The reality of repressed memories. *American Psychologist*, 48(5), 518–537.
- Loftus, E. F. (2003). Make-believe memories. *American Psychologist*, 58(11), 867–873. <https://doi.org/10.1037/0003-066X.58.11.867>.
- Ludvig, E. A., Conover, K., & Shizgal, P. (2007). The effects of reinforcer magnitude on timing in rats. *Journal of the Experimental Analysis of Behavior*, 87(2), 201–218. <https://doi.org/10.1901/jeab.2007.38-06>.
- MacDonald, J. S. (2009). The stay/switch model of concurrent choice. *Journal of the Experimental Analysis of Behavior*, 91(1), 21–39. <https://doi.org/10.1901/jeab.2009.91-21>.
- Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Experimental & Clinical Psychopharmacology*, 11(2), 139–145.
- Madden, G. J., Price, J., & Sosa, F. A. (2016). Behavioral economic approaches to influencing children's dietary decision making at school. *Policy Insights from the Behavioral & Brain Sciences*, 4(1), 41–48. <https://doi.org/10.1177/2372732216683517>.
- McArdle, B., Navakatikyan, M. A., & Davison, M. (2019). *Application of information criteria to behavioral studies*. Retrieved from https://www.researchgate.net/publication/330337138_McArdle_et_al_Application_of_Information_Criteria_Application_of_information_criteria_to_behavioral_studies.
- McLean, A. P., Grace, R. C., & Nevin, J. A. (2012). Response strength in extreme multiple schedules. *Journal of the Experimental Analysis of Behavior*, 97(1), 51–70. <https://doi.org/10.1901/jeab.2012.97-51>.
- Mitchell, S. H., Wilson, V. B., & Karalunas, S. L. (2015). Comparing hyperbolic, delay-amount sensitivity and present-bias models of delay discounting. *Behavioural Processes*, 114, 52–62. <https://doi.org/10.1016/j.beproc.2015.03.006>.
- Myerson, J., & Green, L. (1995). Discounting of delayed rewards: Models of individual choice. *Journal of the Experimental Analysis of Behavior*, 64(3), 263–276.
- Navakatikyan, M. A. (2007). A model for residence time in concurrent variable interval performance. *Journal of the Experimental Analysis of Behavior*, 87(1), 121–141. <https://doi.org/10.1901/jeab.2007.01-06>.
- Navakatikyan, M. A., & Davison, M. (2010). The dynamics of the law of effect: A comparison of models. *Journal of the Experimental Analysis of Behavior*, 93(1), 91–127. <https://doi.org/10.1901/jeab.2010.93-91>.
- Newland, M. C., & Bailey, J. M. (2016). Behavior science and environmental health policy: Methylmercury as an exemplar. *Policy Insights from the Behavioral & Brain Sciences*, 4(1), 96–103. <https://doi.org/10.1177/2372732216686084>.
- Nuzzo, R. (2014). Scientific method: statistical errors. *Nature*, 506(7487), 150–152. <https://doi.org/10.1038/506150a>.
- Pavlov, I. P. (1960). *Conditioned reflexes*. New York: Dover (Original work published 1927).
- Perone, M. (1999). Statistical inference in behavior analysis: Experimental control is better. *The Behavior Analyst*, 22(2), 109–116.
- Perone, M. (2018). How I learned to stop worrying and love replication failures. *Perspectives on Behavior Science*, 42(1), 91–108. <https://doi.org/10.1007/s40614-018-0153-x>.
- Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10, 712. <https://doi.org/10.1038/nrd3439-c1>.
- Rachlin, H. (2006). Notes on discounting. *Journal of the Experimental Analysis of Behavior*, 85(3), 425–435.
- Rung, J. M., & Young, M. E. (2015). Learning to wait for more likely or just more: Greater tolerance to delays of reward with increasingly longer delays. *Journal of the Experimental Analysis of Behavior*, 103(1), 108–124. <https://doi.org/10.1002/jeab.132>.
- Sanabria, F., & Killeen, P. R. (2007). Better statistics for better decisions: Rejecting null hypotheses statistical tests in favor of replication statistics. *Psychology in the Schools*, 44(5), 471–481. <https://doi.org/10.1002/pits.20239>.
- Sanabria, F., & Killeen, P. R. (2008). Evidence for impulsivity in the Spontaneously Hypertensive Rat drawn from complementary response-withholding tasks. *Behavioral & Brain Functions*, 4(1), 7.
- Sellke, T., Bayarri, M. J., & Berger, J. O. (2001). Calibration of p values for testing precise null hypotheses. *American Statistician*, 55(1), 62–71.
- Shadish, W. R., Zelinsky, N. A., Vevea, J. L., & Kratochwill, T. R. (2016). A survey of publication practices of single-case design researchers when treatments have small or large effects. *Journal of Applied Behavior Analysis*, 49(3), 656–673.
- Siegel, S., Baptista, M. A., Kim, J. A., McDonald, R. V., & Weise-Kelly, L. (2000). Pavlovian psychopharmacology: The associative basis of tolerance. *Experimental & Clinical Psychopharmacology*, 8(3), 276–293.
- Smith, T. (2013). What is evidence-based behavior analysis? *The Behavior Analyst*, 36(1), 7–33.

- Smith, T. T., McLean, A. P., Shull, R. L., Hughes, C. E., & Pitts, R. C. (2014). Concurrent performance as bouts of behavior. *Journal of the Experimental Analysis of Behavior*, 102(1), 102–125. <https://doi.org/10.1002/jeab.90>.
- Tanno, T. (2016). Response-bout analysis of interresponse times in variable-ratio and variable-interval schedules. *Behavioural Processes*, 132, 12–21. <https://doi.org/10.1016/j.beproc.2016.09.001>.
- Tincani, M., & Travers, J. (2019). Replication research, publication bias, and applied behavior analysis. *Perspectives on Behavior Science*, 42(1), 59–75. <https://doi.org/10.1007/s40614-019-00191-5>.
- Wagenmakers, E. J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, 11(1), 192–196.
- Warnakulasooriya, R., Palazzo, D. J., & Pritchard, D. E. (2007). Time to completion of web-based physics problems with tutoring. *Journal of the Experimental Analysis of Behavior*, 88(1), 103–113. <https://doi.org/10.1901/jeab.2007.70-06>.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$.”. *American Statistician*, 73(suppl), 1–19. <https://doi.org/10.1080/00031305.2019.1583913>.
- Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594.
- Witnauer, J. E., Hutchings, R., & Miller, R. R. (2017). Methods of comparing associative models and an application to retrospective revaluation. *Behavioural Processes*, 144, 20–32. <https://doi.org/10.1016/j.beproc.2017.08.004>.
- Young, M. E. (2017). Discounting: A practical guide to multilevel analysis of indifference data. *Journal of the Experimental Analysis of Behavior*, 108(1), 97–112. <https://doi.org/10.1002/jeab.265>.
- Young, M. E. (2019). Bayesian data analysis as a tool for behavior analysts. *Journal of the Experimental Analysis of Behavior*, 111(2), 225–238. <https://doi.org/10.1002/jeab.512>.