Psychonomic Bulletin & Review
2004, 11 (5), 791-806

Likelihood ratios: A simple and flexible statistic for empirical psychologists

SCOTT GLOVER
Pennsylvania State University, University Park, Pennsylvania

and

PETER DIXON
University of Alberta, Edmonton, Alberta, Canada

Empirical studies in psychology typically employ null hypothesis significance testing to draw statistical inferences. We propose that likelihood ratios are a more straightforward alternative to this approach. Likelihood ratios provide a measure of the fit of two competing models; the statistic represents a direct comparison of the relative likelihood of the data, given the best fit of the two models. Likelihood ratios offer an intuitive, easily interpretable statistic that allows the researcher great flexibility in framing empirical arguments. In support of this position, we report the results of a survey of empirical articles in psychology, in which the common uses of statistics by empirical psychologists is examined. From the results of this survey, we show that likelihood ratios are able to serve all the important statistical needs of researchers in empirical psychology in a format that is more straightforward and easier to interpret than traditional inferential statistics.

A likelihood ratio statistic reflects the relative likelihood of the data, given two competing models. Likelihood ratios provide an intuitive approach to summarizing the evidence provided by an experiment. Because they describe evidence, rather than embody a decision, they can easily be adapted to the various goals for which inferential statistics might be used. In contrast, the logic of null hypothesis significance testing is often at odds with the goals of the researcher, and that approach, despite its common usage, is generally ill suited for the varied purposes to which it is put.

We develop our thesis in three sections. In the first section, we provide an introduction to the use of likelihood ratios as model comparisons and describe how they relate to more traditional statistics. In the second section, we report the results of a small survey in which we identify some of the common goals of significance testing in empirical psychology. In the third section, we describe how likelihood ratios can be used to achieve each of these goals more directly than traditional significance tests.

The present work was supported by the Natural Sciences and Engineering Research Council of Canada through a fellowship to the first author and a grant to the second author. The authors thank Michael Lee, Geoff Loftus, and an anonymous reviewer for their insightful comments on previous versions of this article. Correspondence concerning this article should be addressed to S. Glover, Department of Psychology, Royal Holloway University of London, Egham, Surrey TW20 0EX, England (e-mail: scott.glover@rhul.ac.uk).

LIKELIHOOD RATIOS

A likelihood ratio can be thought of as a comparison of two statistical models of the observed data. Each model provides a probability density for the observations and a set of unknown parameters that can be estimated from the data. In a broad range of common situations, the density is the multivariate normal distribution, and the parameters are the means in the various conditions, together with the error variance. The two models differ in terms of the constraints on the condition means. For example, a model in which two condition means differ might be compared with one in which the means are identical. The match of each model to the observations can be indexed by calculating the likelihood of the data, given the best estimates of the model parameters: The more likely the data are, the better the match. In this case, the best parameter estimates are those that maximize the likelihood of the data, which are termed *maximum-likelihood estimates*. The ratio of two such likelihoods is the maximum likelihood ratio; it provides an index of the relative match of the two models to the observed data. Formally, the likelihood ratio can be written as

$$\lambda = \frac{f\left(\mathbf{X} \mid \hat{\theta}_2\right)}{f\left(\mathbf{X} \mid \hat{\theta}_1\right)}, \tag{1}$$

where f is the probability density, $\mathbf{X}$ is the vector of observations, and $\hat{\theta}_1$ and $\hat{\theta}_2$ are the vectors of parameter estimates that maximize the likelihood under the two models.

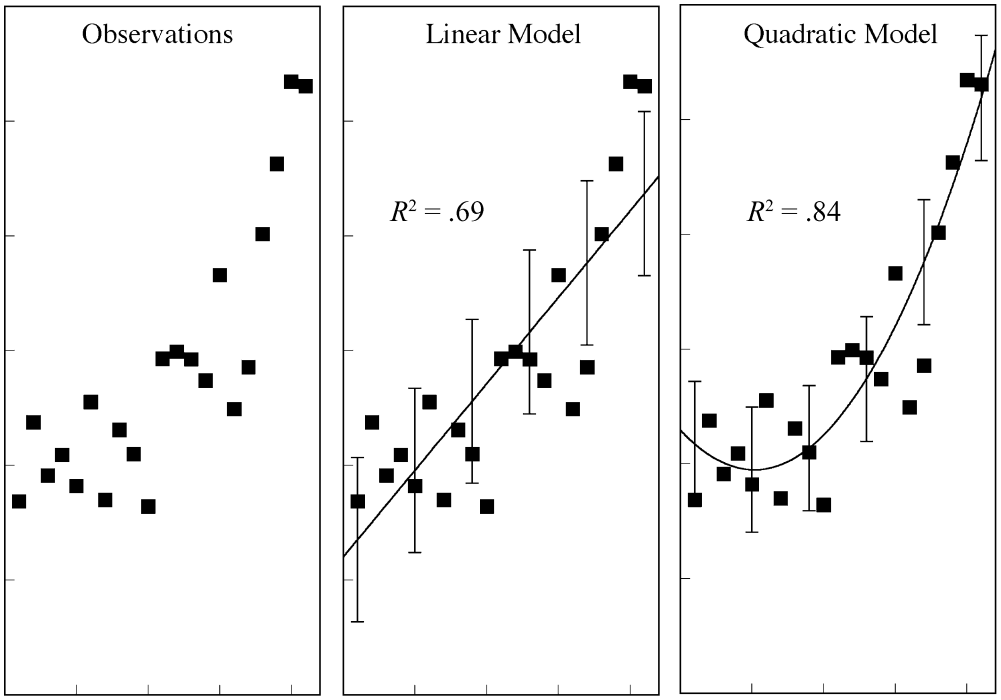

Figure 1. Comparison of linear (middle panel) and quadratic (right panel) model fits of a theoretical data set. Error bars indicate the standard deviations of the observations under the best-fitting model; for clarity, only six are shown.

Viewing statistical inference as model comparison is a well-established perspective. Judd and McClelland (1989), for example, organized an entire introductory textbook around this approach. Furthermore, the use of likelihood ratios in statistical inference is common (e.g., Edwards, 1972; Royall, 1997), and the role of likelihood in model comparison is well established (e.g., Akaike, 1973; Schwartz, 1978). Furthermore, the likelihood ratio plays a pivotal role in most approaches to hypothesis testing. In Bayesian hypothesis testing, the posterior odds of two hypotheses are related to the products of the likelihood ratio and the prior odds (e.g., Sivia, 1996). In the decision-theoretic approach advocated by Neyman and Pearson (1928, 1933), any suitable decision rule can be viewed as a decision based on the value of the likelihood ratio. Fisher (1955) advocated the use of the log likelihood ratio as an index of the evidence against a null hypothesis. In all of these approaches, the likelihood ratio represents the evidence provided by the data with respect to two models. Although the form of the likelihood ratio varies to some extent in these different approaches, they all have a common conceptual basis.

Consider the hypothetical data shown in Figure 1. The data points in each panel represent the effects of an independent variable X on a dependent variable Y. The straight line in the middle panel indicates the best fitting linear model of the 21 observations, whereas the line in the right panel shows the fit of a more complicated model that includes a quadratic component. The fit of the models can be described by the correlation between the predicted and the observed values, and the value of R^2 for each model is indicated in the figure. The standard deviation shown by the error bars is an index of the residual variance that is not explained by the model and is proportional to $\sqrt{(1-R^2)}$. The error bars are shown on the curve to indicate that the estimate depends on which model is being fit (cf. Estes, 1997).

It appears from Figure 1 that the data are more likely given the best-fitting quadratic model on the right than given the best-fitting linear model. This is reflected in the smaller deviations from the predicted values and the larger value of R^2. As a consequence, $1-R^2$ and the estimate of the standard error are smaller for the quadratic model than for the linear model. In fact, the likelihood is related to the inverse of the standard deviation. The ratio of the likelihoods thus indexes the relative quality of the two fits, and with normally distributed data, one can write the likelihood ratio as

$$\lambda = \left(\frac{1/s_2^2}{1/s_1^2}\right)^{\frac{n}{2}} = \left(\frac{s_1^2}{s_2^2}\right)^{\frac{n}{2}} = \left(\frac{1-R_1^2}{1-R_2^2}\right)^{\frac{n}{2}}, \qquad (2)$$

where s_1 and s_2 are the two estimates of the standard deviation, n is the number of observations, and R_1^2 and R_2^2 describe the quality of the fits of the two models. (A proof is provided in Appendix A.) In this example, the R^2 values are .689 and .837, and the value of the likelihood ratio is 862.6. In simple terms, this means that the data are 862.6 times as likely to occur if the second model

(and its best-fitting parameter values) is true than if the first model (and its best-fitting parameter values) is true.

The $1-R^2$ terms correspond to residual variation that is not explained by each model. Thus, an equivalent description of the likelihood is

$$\lambda = \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}} \right)^{\frac{n}{2}}. \quad (3)$$

This conceptual formula provides a straightforward approach to calculating likelihood ratios for many kinds of model comparisons. In particular, when the models in question are linear, the requisite information about unexplained variation can typically be found in a factorial analysis of variance (ANOVA) table.

Correcting for Model Complexity

Although arguments based on this form of the likelihood ratio are sometimes possible, there is an obvious problem apparent in this example: Because the quadratic model includes an extra parameter, it will always fit the data better than the linear model, regardless of what results are obtained. Generally, the likelihood ratio will always favor the more complex of two nested models. This is a well-known issue in model comparison, and a variety of techniques have been devised to address it.

Two common approaches are the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwartz, 1978). In both cases, the assumption is that there is a (sometimes large) collection of possible models, and one wishes to compute an index that allows one to determine the "best" model. The AIC measure derives from an information measure of the expected discrepancy between the true model and the model under consideration. It can be written as

$$\text{AIC} = -2\ln(l) + 2k, \quad (4)$$

where l is the maximum likelihood of the data and k is the number of parameters. If one is considering only two models, then, according to the AIC measure, Model 2 should be selected over Model 1 when $\lambda_A = Q_A\lambda$ is greater than 1, where λ is the maximum likelihood ratio, Q_A is

$$Q_A = \exp(k_1 - k_2), \quad (5)$$

and k_1 and k_2 are the number of parameters in Models 1 and 2. Hurvich and Tsai (1989) discussed a correction of AIC for small samples. Using the corrected AIC criterion, Model 2 should be selected over Model 1 when $\lambda_C = Q_C\lambda$ is greater than 1, where Q_C is

$$Q_C = \exp\left[k_1\left(\frac{n}{n-k_1-1} \right) - k_2\left(\frac{n}{n-k_2-1} \right) \right]. \quad (6)$$

It is common to discuss the model selection problem as one of picking the best model. However, we argue that in experimental psychology, these indices are best used merely to describe the evidence, rather than to make decisions. Thus, for example, a value of $\lambda_C = 1.2$ might indeed indicate that Model 2 is preferable to Model 1 if one of the two models has to be selected and if the researcher is indifferent a priori. However, a likelihood ratio of that magnitude could not be regarded as very strong evidence one way or the other and should not be persuasive if there were other reasons to prefer Model 1.

A comparable but somewhat different index can be derived using principles of Bayesian inference. The alternative index is referred to as the BIC (Schwartz, 1978), and is defined as

$$\text{BIC} = -2\ln(l) + k\ln(n), \quad (7)$$

where n is the sample size. Using this criterion, Model 2 should be preferred over Model 1 when $\lambda_B = Q_B\lambda$ is greater than 1, where Q_B is

$$Q_B = \left[\exp(k_1 - k_2)\right]^{\ln(n)/2}. \quad (8)$$

In this case, λ_B can be viewed as an estimate of the Bayesian posterior odds of the two models, assuming uninformative priors.

Pitt, Myung, and Zhang (2002) have argued convincingly that the number of parameters as captured by these indices is only one aspect of model flexibility and that when one is dealing with nonlinear models, the functional form of the model also needs to be taken into account. For example, they noted that in psychophysics, Stevens's power model is more flexible (i.e., can account for more varied patterns of data) than Fechner's logarithmic model, even though they both have the same number of parameters. As an alternative to such indices as AIC or BIC, they described a measure based on *minimum-description length* (MDL; Rissanen, 1996). With the MDL approach, Model 2 should be preferred over Model 1 when $\lambda_M = Q_M\lambda$ is greater than 1, where Q_M is

$$Q_M = \left[\exp(k_1 - k_2)\right]^{\frac{\ln(n/2\pi)}{2}} C_{1,2}. \quad (9)$$

Here, $C_{1,2}$ indexes the relative complexity of the two models as determined by their functional forms; in many cases, $C_{1,2}$ must be estimated using numerical methods. The MDL approach leads to a correction of the likelihood ratio that is related to that found with AIC and BIC but also includes information about model form. Although this approach is important in some situations, it is our impression that many theoretical distinctions in experimental psychology can be cast in terms of linear models. (The standard factorial ANOVA model is a prime example.) For these approaches, an adjustment made purely on the basis of number of parameters (as in Equations 6 and 8) may be adequate.

All of these approaches to model comparisons incorporate the likelihood ratio because it captures the evidence provided by the data with respect to the two possible models. The likelihood ratio is then adjusted to reflect other aspects of the models (such as the number of model parameters and their possible values). The approaches differ in how these other factors should be weighed in arriving at a final assessment of the relative merits of the two models. Such adjustments are crucial in assessing the advantage a model might accrue on the basis of additional parameters or degrees of freedom.

However, the question of which adjustment one should use is a difficult conceptual problem. Ultimately, the answer may depend on relatively subtle distinctions in the assumptions one makes about the goals of theory development and model comparison. For example, a rough characterization of the difference between the information-theoretic and Bayesian approaches is that in the Bayesian approach, one assumes that either of the models is potentially true, and the goal is to estimate the odds that one model is correct. On the other hand, in the information-theoretic approach, one assumes that neither of the models is likely to be precisely correct and that the reality could easily be different, at least in detail. Consequently, the goal is to estimate the odds that one model provides a better approximation.

The relative merits of these different perspectives are beyond the scope of the present article. Fortunately, in a wide range of realistic situations, the choice of adjustment procedure has little impact on the conclusions one might draw from the evidence. The reason is that empirical investigations are typically designed to provide overwhelming evidence for a particular theoretical interpretation. In other words, one usually attempts to find evidence that would convince all possible observers, regardless of their initial level of skepticism or how strongly committed they might be to alternative perspectives. Such strong evidence will often be compelling regardless of how the likelihood ratio is corrected for number of parameters or model flexibility.

As an illustration, the Q_C and Q_B adjustments can be applied to the model fits shown in Figure 1. In this calculation, the linear model has three parameters (the intercept, the slope, and the error standard deviation), and the quadratic model has an additional parameter for the quadratic component, for a total of four. Thus, using the information-theoretic approach in Equation 6, λ_C for the model fits in Figure 1 would be

$$\begin{aligned}\lambda_C &= Q_C\lambda \\ &= \exp\left[k_1\left(\frac{n}{n-k_1-1}\right)-k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda \\ &= \exp\left[2\left(\frac{21}{20-4-1}\right)-3\left(\frac{21}{20-3-1}\right)\right](862.6) \\ &= 184.2.\end{aligned}$$

On the other hand, using the Bayesian adjustment in Equation 8, λ_B would be

$$\begin{aligned}\lambda_B &= Q_B\lambda \\ &= \left[\exp\left(k_1-k_2\right)\right]^{\ln(n)/2}\lambda \\ &= \left[\exp\left(3-4\right)\right]^{\ln(21)/2}\left(862.6\right) \\ &= 188.2.\end{aligned}$$

Both values strongly favor the quadratic model over the linear model. With large sample sizes, the values of λ_C and λ_B would be expected to differ more substantially. However, an experiment with a large sample size is also likely to be powerful and to produce strong evidence regardless of which adjustment is used. Generally, we believe that the choice of a correction factor will be critical only when the theoretical distinctions hinge on relatively subtle aspects of the data.

Relations to Other Approaches

Because likelihood ratios are central to the development of traditional inferential statistics, they can be readily derived from those statistics. For example, in the regression context illustrated in Figure 1, the likelihood ratio is related to a significance test of the quadratic component, and the F ratio for that test is given by

$$F\left(1, df_{\text{error}}\right) = df_{\text{error}}\left(\lambda^{2/n}-1\right). \tag{10}$$

Equivalently,

$$\lambda = \left[1+\frac{F\left(1, df_{\text{error}}\right)}{df_{\text{error}}}\right]^{n/2}. \tag{11}$$

(A proof is provided in Appendix B.) In this case, the test of the quadratic component would yield $F(1,18) = 16.27$, $p < .001$.

Although this result is related to the conclusion one would draw on the basis of likelihood ratios, there are important conceptual differences. The significance test is a decision rule that presumably entails behavioral consequences; the implication in this instance is that one should behave as if the quadratic component exists. In contrast, the likelihood ratio merely describes the evidence obtained in the experiment that pertains to the two models under consideration. Although it is possible to identify likelihood ratio magnitudes with heuristic labels, such labels should not be construed as decisions. For example, we would usually regard an adjusted likelihood ratio of 3:1 as "clear" evidence in favor of one model relative to another, and values of this size would be considered *moderate to strong* evidence by Goodman and Royall (1988). However, such labels in themselves do not embody a decision rule, and their rhetorical use must be a function of the broader theoretical and empirical context. Indeed, we would concur with the critique of Loftus (1996), Rozeboom (1960), and others that drawing a fixed arbitrary distinction between *significant* and *nonsignificant* (or even between *clear evidence* and *weak evidence*) is misleading and unproductive. Rather, what one may plausibly argue with respect to those two models in the light of the evidence depends on a variety of considerations beyond the results of a single study, including the a priori plausibility of the interpretations, the consistency of the results with previous findings, and relative parsimony. More generally, the fact that likelihood ratios merely describe the evidence, rather than embody a decision rule, is central in adapting likelihood ratios to a range of interpretational goals.

Other means of presenting evidence are also available. For example, the use of confidence intervals has been frequently proposed as an alternative to hypothesis testing (e.g., Tyron, 2001). Loftus and his colleagues have advocated the use of confidence intervals as a suitable

index of variability in conjunction with a graphical presentation (e.g., Loftus, 1996, 2001; Masson & Loftus, 2003). This approach has the advantage of allowing a quick visual evaluation of the size of an effect, relative to the variability in the data. Consequently, confidence intervals or related indices of variability can be especially useful when graphs are used to present evidence for null effects. Generally, we believe that the graphical display of evidence (incorporating suitable error bars) is an indispensable component of intelligent data analysis and presentation. The likelihood ratio methods discussed here provide a quantitative complement to such visual techniques.

A SURVEY OF STATISTICAL GOALS IN EMPIRICAL PSYCHOLOGY

Although the use of significance testing has remained a controversial topic over the years, a large proportion of the literature on significance testing has focused on the analysis of relatively simple situations. For example, one may be interested in whether a relationship exists between two variables, whether two means differ, and so on (e.g., Lykken, 1968; Rozeboom, 1960). However, real research in psychology is rarely this simple, and our impression is that significance testing is used for a broad range of purposes. In this section, we will present the results of a small survey in which the use of significance tests within empirical psychology is explored.

Method

As our sample, we randomly selected two empirical articles from each of six journals: *Canadian Journal of Experimental Psychology, Experimental Brain Research, Journal of Cognitive Neuroscience, Journal of Experimental Psychology: Human Perception and Performance, Nature Neuroscience,* and *Psychonomic Bulletin & Review* (see Table 1).

For each article, the reported significance tests (e.g., each t, F, or p value) were assigned to one or more of six categories, as follows.

Tests of a single hypothesis. These tests include any significance test that compares the author(s)' hypothesis against an unspecified alternative (i.e., the *null*). In this case, there is a single theoretical interpretation for which the results may provide evidence. Following in the conceptual tradition of Fisher (e.g., 1925), evidence is garnered by rejecting a null hypothesis of no difference. The null hypothesis has no theoretical interpretation that is clearly stated in the article. The following example describes a single hypothesis,

> We hypothesized . . . that perception of the direction of eye gaze would elicit activity in regions associated with spatial perception and spatially directed attention, namely, the intraparietal sulcus (IPS),

and the ensuing significance test:

> . . . attention to gaze elicited a stronger response in the left IPS than did attention to identity (0.99% versus 0.80%, n = 7, p = .0001). (Hoffmann & Haxby, 2000, pp. 80–81)

Since Hoffmann and Haxby offered no theoretical alternative to the stated hypothesis, this was classified as a test of a single hypothesis.

Exploratory tests. Exploratory tests are similar to tests of a single hypothesis, except that the research hypothesis is not justified a priori. Because these significance tests are not explicitly motivated by a theoretical account, when the null hypothesis is rejected, the authors may conclude that the result is surprising. An example of the use of exploratory tests is as follows:

> Unexpectedly, there was a significant three-way interaction of response side, distractor position, and congruency, $F(2,28) = 15.57$, $p < .001$. (Diedrichsen, Ivry, Cohen, & Danziger, 2000, p. 115)

Table 1
Number of Significance Tests in Surveyed Articles by Category

Article	Single Replication	Competing Hypothesis	Replication	Pro Forma	Exploratory	Methodological	Total
Adolphs & Tranel, 2000	8	0	0	8	0	1	17
Arbuthnott & Frank, 2000	0	9	3	23	0	0	35
Chochon, Cohen, van de Moortele, & Dehaene, 1999	109	0	0	60	6	1	176
De Gennaro, Ferrara, Urbani, & Bertini, 2000	3	0	0	6	0	11	20
Diedrichsen, Ivry, Cohen, & Danziger, 2000	8	10	6	10	6	1	40
Fugelsang & Thompson, 2000	21	3	12	3	0	1	46
Hoffmann & Haxby, 2000	13	0	8	7	0	2	30
Kinoshita, 2000	0	8	0	29	0	0	37
Prabhakaran, Narayanan, Zhao, & Gabrieli, 2000	17	0	6	18	0	4	45
Servos, 2000	3	0	7	3	0	4	17
Soto-Faraco, 2000	3	33	49	58	0	0	95
Zheng, Myerson, & Hale, 2000	11	0	0	3	0	0	14

Such outcomes are often pursued in subsequent experiments.

Tests of competing hypotheses. In this case, a significance test is used to compare two alternative hypotheses described by the author(s). Competing-hypotheses tests differ from single-hypothesis tests in that there is a theoretical interpretation both to rejecting the null hypothesis and to accepting it. Consequently, the test result is used as evidence in favor of one of two mutually incompatible theoretical alternatives. In the following example, Diedrichsen et al. (2000) analyzed their data in the context of two competing hypotheses:

> For the different congruent conditions, this difference (in reaction times) was only 4 msec. Although this difference was in the direction predicted by the attentional-shift hypothesis, it was not significant, $t(37) = 1.29, p = .203 \ldots$. Thus, these results are in accord with the perceptual-grouping hypothesis. (p. 119)

Replication tests. These tests are intended to confirm the existence of effects or trends that are expected on the basis of previous research. Generally, researchers indicate that they expect such tests to confirm the hypothesis in question. For example,

> the findings of Experiment 1 were replicated in that a reliable negative correlation emerged for the "original" problem format $r(93) = -.46, p < .001$. (Fugelsang & Thompson, 2000, pp. 22–23)

Methodological tests. These tests are performed to confirm the presence or absence of a confound in the analysis. In many cases, failing to reject the null hypothesis is interpreted as evidence that the confound does not exist. In the following example, the question of interest is whether four classes of stimuli are rated similarly by healthy control participants:

> The two groups of normal controls did not differ in their ratings of the four classes of stimuli, as confirmed by a one-way multivariate analysis of variance (MANOVA) on subject's mean ratings [Wilkes lamda = 0.94; $F(4) = 1.29$; $p = 0.28$]. (Adolphs & Tranel, 1999, p. 611)

Pro forma tests. These tests involve comparisons among conditions that are not expected to differ and/or for which there was no clear interpretation for an effect if one were to be found. For example, after describing the effects of interest from an ANOVA, the authors often go on to say something such as the following: "No other main effect or interaction was significant" (De Gennaro, Ferrara, Urbani, & Bertini, 2000, p. 111).

Results and Discussion

The results of our survey are summarized in Figure 2 and broken down by study in Table 1. Figure 2 shows that the large majority of the statistics used in these articles belong to one of three main categories. First, nearly 40% of the statistics were used for either single- or competing-hypothesis testing. Second, about 15% of the statistics were used to test whether previous findings had been replicated. Third, roughly 40% of the statistics were attributed to the *pro forma* category. Relatively few statistics were assigned to either the *exploratory* (about 2%) or the *methodological* (about 8%) category.

These results lead to several interesting observations. First, the tests that conformed most closely to the logic of significance testing were not overwhelmingly common. Tests of a single hypothesis and exploratory tests involve advancing a research hypothesis by rejecting a corresponding null hypothesis, as was suggested by Fisher (1925), and together these make up only about 35% of the tabulated tests. Second, a substantial number of tests involved, at least potentially, advancing a theoretical interpretation by accepting the null hypothesis. This can happen with either tests of competing hypotheses or methodological tests when the null hypotheses are not rejected. The potential pitfalls of this logic are well known (such as when power is low; Cohen, 1977; Loftus, 1996). Finally, over half the tests were conducted even though they had little chance of affecting the beliefs or behavior of the researchers. In particular, the null hypothesis is expected to be false in replication tests, because the effect in question has been previously obtained in similar research; with pro forma tests, researchers have every reason to believe the null hypothesis to be true. In some examples of pro forma tests, researchers explicitly hold to the null hypothesis even if the results are statistically significant (Dixon, 2001).

USING LIKELIHOOD RATIOS

Because likelihood ratios represent the evidence obtained in the study, rather than embodying a rule or a procedure, they can be easily adapted to the various goals with which researchers are concerned. Below, we will describe how likelihood ratios can be used to address each of the common uses of statistics we have identified.

Evaluating a Single Hypothesis

In order to evaluate the evidence for the hypothesis that a given effect is present, the standard logic of significance testing requires that one assess the evidence against the hypothesis of no difference. If the null hypothesis is rejected, one can then embrace the hypothesis of interest. In effect, the goal of the researcher in these situations is to compare two models: one based on the assumption of no difference (the null model) and one based on the existence of some real difference. The likelihood ratio provides precisely the information that would be used in such a comparison.

Consider, as an example, the results of a 2×2 study depicted in Figure 3. In order to make this example concrete, we might imagine that these data represent the number of words recalled in a list-learning experiment and that the factors correspond to the number of learning trials and whether the words were abstract or concrete. Suppose, in this context, that one were interested in the hypothesis that the two factors interact. For example, one might hold the hypothesis that learning trials have a greater effect for abstract words than for concrete words. Using a significance-testing procedure, one would calcu-

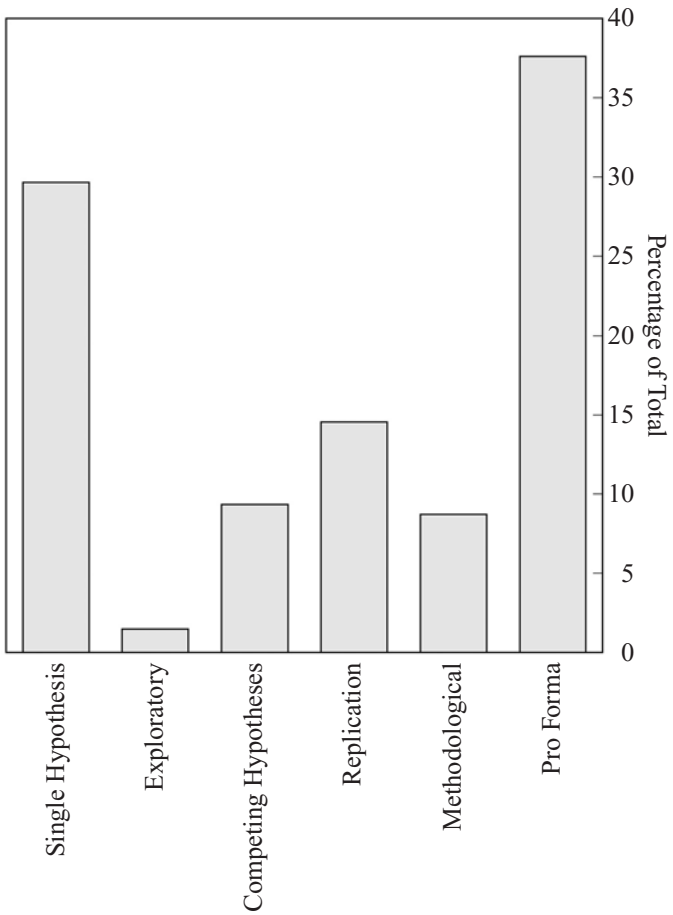

Figure 2. Results of the survey of the use of statistics in a sample of empirical articles taken from experimental psychology and cognitive neuroscience. Bars represent the percentages of statistics falling into each category (see text for explanation of categories).

late the *F* ratio for the ANOVA interaction and compare it with the appropriate critical *F* value. The corresponding approach using likelihood ratios would involve comparing a model that included the interaction with an additive model that included the main effects, but not the interaction. These two models are illustrated in Figure 3: The left panel shows the fit of the additive model (not including the interaction), and the panel on the right shows the fit of a model that includes the interaction. As in Figure 1, the results are graphed with the error bars depicting the standard errors based on the residual variation. Note that because the model on the right is a *full* model that includes all possible degrees of freedom, it matches the condition means exactly. However, the variation among the observations within each condition is error and is not predicted by either model.

The likelihood ratio for comparing these two models can be calculated from the unexplained sources of variation, as described by Equation 3, and these are readily available from an ANOVA table (shown in Table 2). The unexplained variation for the additive model would be the error sum of squares (470.1), as well as the interaction sum of squares (104.1). In contrast, the only source of variation that is not explained by the second model (which includes the interaction) is error. Consequently, the likelihood ratio for comparing the two models would be

$$\lambda = \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}} \right)^{\frac{n}{2}}$$

$$= \left(\frac{SS_{\text{T}\times\text{C}} + SS_{\text{error}}}{SS_{\text{error}}} \right)^{\frac{n}{2}}$$

$$= \left(\frac{104.1 + 470.1}{470.1} \right)^{\frac{40}{2}}$$

$$= 54.7.$$

In order to correct the likelihood ratio for the additional complexity of the second model, one needs to identify the number of parameters in each model. The first model includes parameters for the variance, the overall mean,

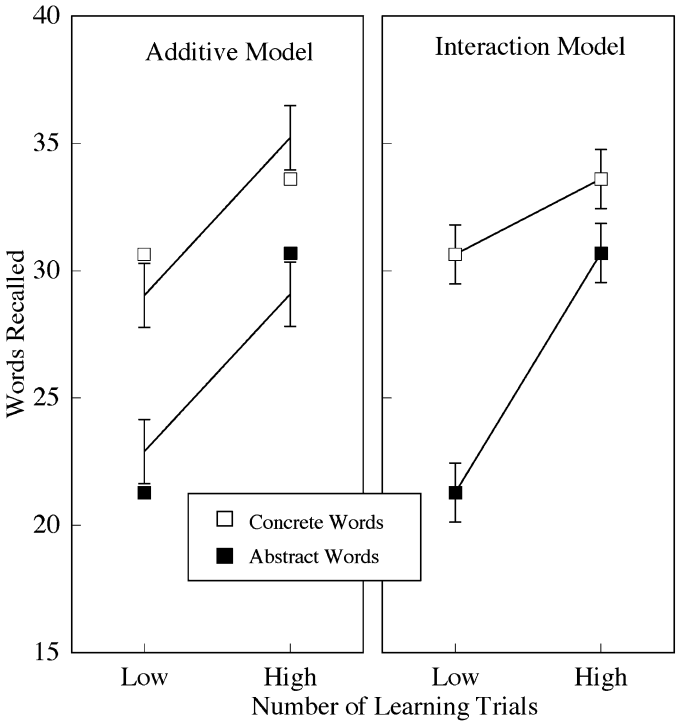

Figure 3. Comparison of the fits of an additive model (left panel) and an interactive model (right panel) in accounting for the effects of word type in a hypothetical data set.

and the two main effects, for a total of four; the second model includes a parameter for the interaction as well, for a total of five. Thus, using Equation 6 to correct for the additional degree of freedom yields

$$\begin{aligned}\lambda_C &= Q_C\lambda \\ &= \exp\left[k_1\left(\frac{n}{n-k_1-1}\right)-k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda \\ &= \exp\left[5\left(\frac{40}{40-4-1}\right)-4\left(\frac{40}{40-3-1}\right)\right](54.7) \\ &= 14.7.\end{aligned}$$

Thus, the data are over 14 times more likely given the interactive model than given the purely additive model, even taking into account the additional degree of freedom. These data clearly favor the hypothesis that learning trials and word type interact.

Table 2
ANOVA Table for Evaluating a Single Hypothesis

Source	*df*	Sums of Squares	Mean Square	*F*	*p*
Trials (T)	1	383.1	383.1	29.34	≤.0001
Concreteness (C)	1	376.9	376.9	28.86	≤.0001
T × C	1	104.1	104.1	7.97	.0077
Error	36	470.1	13.1		
Total	39	1,334.2			

Evaluating Competing Hypotheses

Because likelihood ratios can be thought of as comparisons of two statistical models, their use applies directly to situations in which there are two competing theoretical interpretations of the results of an experiment. As an example of this use of likelihood ratios, we will consider a slightly more complex situation, depicted in Figure 4. Again, suppose that the dependent variable is the number of words recalled in a list. In this case, though, suppose that the factors consist of concreteness and familiarity of the words. Here, one might wish to compare a theory that says that there should simply be an effect of familiarity with a theory that says that only unfamiliar abstract words should be difficult to recall. The assumption in the latter account is that familiar concrete, familiar abstract, and unfamiliar concrete words should all have comparable recall, which in turn should be greater than that for unfamiliar abstract words. The fits of these two models is shown in Figure 4: On the left, the prediction line depicts the single main effect of familiarity, and on the right, the prediction is that the upper three points are all equal and larger than the lowest point. An ANOVA table for this design is shown in Table 3.

It is noteworthy here that the usual hypothesis-testing approach to these data does not provide any tests that are directly relevant. Nevertheless, the two fits depicted in Figure 4 can be compared using likelihood ratios. To do so, one would simply find the unexplained variation for each and then apply Equation 3. From the ANOVA table

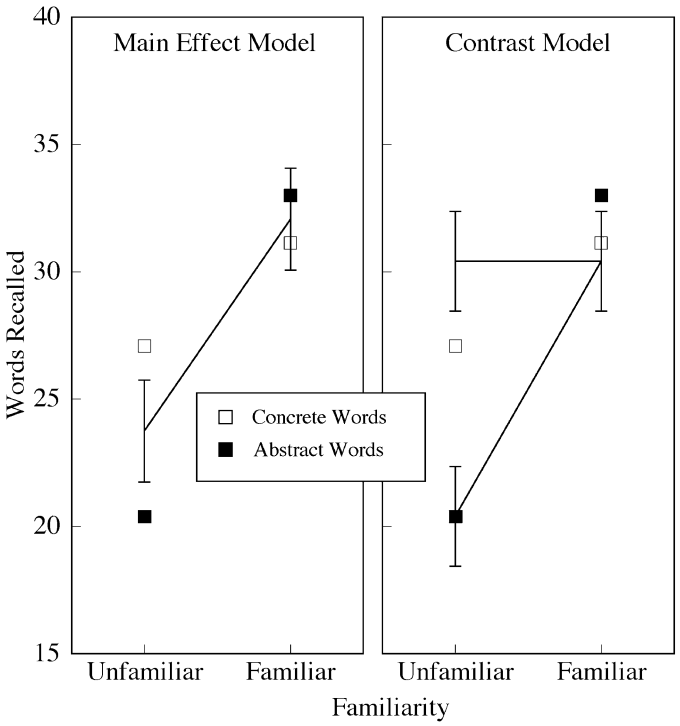

Figure 4. Comparison of the fits of a main effect model (left panel) and a contrast model (right panel) in explaining the data in a hypothetical data set.

shown in Table 3, the unexplained variation in the first model consists of the sums of squares for concreteness, the interaction, and the error term, or 58.4 + 184.0 + 1,271.9 = 1,514.4. To find the unexplained sum of squares for the second model, one would begin by finding an ANOVA contrast that captures the theoretical interpretation. In this case, the interpretation is that the three upper points in Figure 3 are equal and greater than the lower point; thus, the contrast would be −3, 1, 1, 1. The sum of squares predicted by the contrast is given by

$$SS_{\text{contrast}} = \frac{n\left(\sum c_i \bar{X}_i\right)^2}{\sum c_i^2}, \tag{12}$$

where $\bar{X}_i$ is the mean in the ith condition, c_i is the contrast coefficient for that condition, and $n_.$ is the number of observations in each condition. If the sample means shown in Figure 3 are 20.4, 33.0, 27.1, and 31.1, the SS_{contrast} would be 753.9. The unexplained sum of squares would be the total sum of squares minus that which is explained by the contrast—that is, 2,208.6 − 753.9 = 1,454.74. Putting these calculations together with Equation 3 yields the likelihood ratio

$$\lambda = \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}}\right)^{\frac{n}{2}}$$
$$= \left(\frac{1{,}514.4}{1{,}454.7}\right)^{40/2}$$
$$= 2.2.$$

Thus, the data are only 2.2 times as likely given the second model (which predicts lower recall only with unfamiliar abstract words) than given the first model (which predicts only a main effect of familiarity). (In this example, the two models being considered have the same number of parameters, so no adjustment for model complexity is required.) A likelihood ratio of 2.2 would likely be regarded as relatively weak evidence in favor of the second model.

Table 3
ANOVA Table for Evaluating Competing Hypotheses

Source	*df*	Sums of Squares	Mean Square	*F*	*p*
Concreteness (C)	1	58.4	58.4	1.65	.2067
Familiarity (F)	1	694.2	694.2	19.65	<.0001
C × F	1	184.0	184.0	5.21	.0285
Error	36	1,271.9	35.3		
Total	39	2,208.6			

Evidence for Null Effects

Another advantage of using likelihood ratios is that they can easily be adapted to provide evidence for the *lack* of an effect. This allows one to avoid the pitfall of accepting null results with low power. One approach would be to correct the likelihood ratio, using Q_C or Q_B, as in Equation 6 or 8; when so corrected, the likelihood ratio may favor the simpler model. As an example, consider the results presented in Figure 5. Suppose that these data come from a study comparing the recall of a list of

40 words presented in either normal or bold type. The panel on the left shows the fit of the null model, in which typeface is assumed to have no effect; the center panel shows the fit assuming some difference between conditions. Intuitively, the fit in the center does not seem much better than the one on the left, and appropriately, the adjusted likelihood ratio favors the simpler null model.

In particular, the likelihood ratio can be calculated from the ANOVA table in Table 4:

$$\begin{aligned}\lambda &= \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}}\right)^{\frac{n}{2}} \\ &= \left(\frac{SS_{\text{typeface}} + SS_{\text{error}}}{SS_{\text{error}}}\right)^{\frac{n}{2}} \\ &= \left(\frac{14.4 + 507.9}{507.9}\right)^{\frac{40}{2}} \\ &= 1.75.\end{aligned}$$

Correcting for model complexity using Q_B (based on the BIC) yields

$$\begin{aligned}\lambda_B &= Q_B\lambda \\ &= \left[\exp\left(k_1 - k_2\right)\right]^{\ln(n)/2}\lambda \\ &= \left[\exp\left(2-3\right)\right]^{\ln(40)/2}\left(1.75\right) \\ &= 0.28,\end{aligned}$$

or a likelihood ratio of 3.28 in favor of the null model. Similarly, if one corrects for model complexity using Q_C (based on the information-theoretic approach), one obtains

$$\begin{aligned}\lambda_C &= Q_C\lambda \\ &= \exp\left[k_1\left(\frac{n}{n-k_1-1}\right) - k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda \\ &= \exp\left[2\left(\frac{40}{40-2-1}\right) - 3\left(\frac{40}{40-3-1}\right)\right]\left(1.75\right) \\ &= 0.54,\end{aligned}$$

or a likelihood ratio of 1.84 in favor of the null model. Although both of these values favor the null model, the evidence represented by λ_B in this example is not overwhelming, and the value of λ_C is quite equivocal.

Another approach that can sometimes produce more compelling evidence for null effects is to use one's a priori knowledge about effect magnitude. The critical step here is to identify a minimal magnitude for a theoretically interesting effect. Then a likelihood ratio can be constructed in which a null model is compared with a model in which the effect is assumed to be at least as large as that minimal effect magnitude. The null model becomes a plausible alternative when the observed effect is, in fact, smaller than the minimal value. Under such circumstances, the difference between the observed value and the minimal effect magnitude constitutes an overprediction and counts as unexplained variation for the theoretically interesting effect in the likelihood ratio.

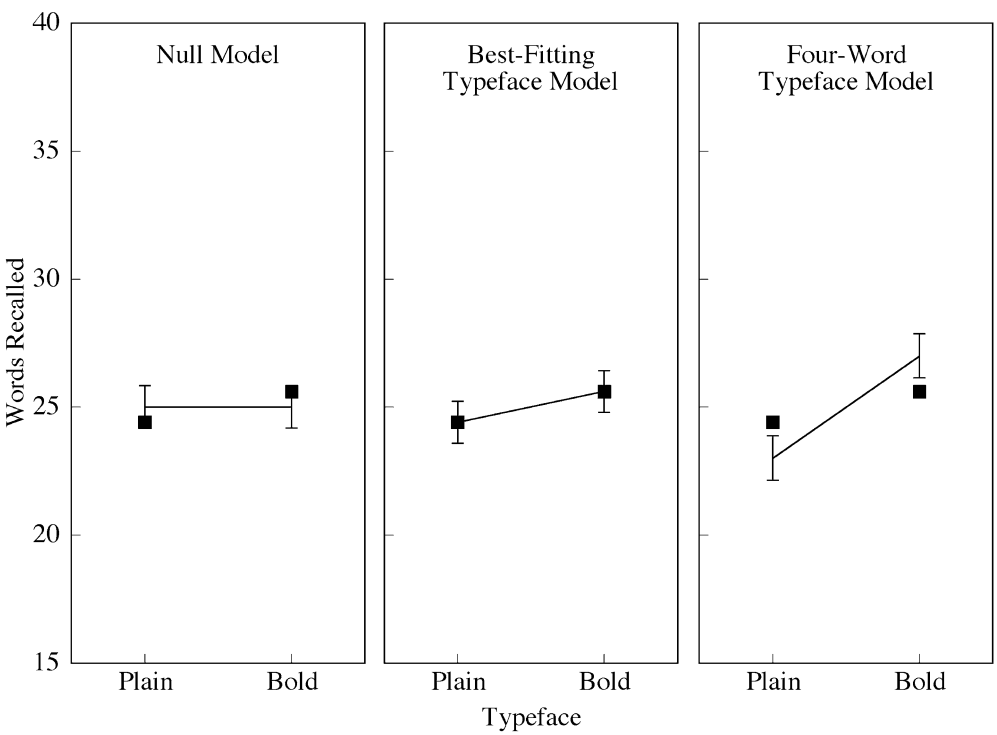

Figure 5. Comparison of the fits of a null model that predicts no effect of typeface (left panel), the best-fitting typeface model that predicts some effect of typeface (middle panel), and a model that predicts a 4-word effect of typeface (right panel) in a hypothetical data set.

Table 4
ANOVA Table for Evaluating a Null Hypothesis

Source	*df*	Sums of Squares	Mean Square	*F*	*p*
Typeface	1	14.4	14.4	1.08	.3058
Error	38	507.9	13.4		
Total	39	522.3			

As an illustration in the typeface example shown in Figure 5, one might be able to argue that the effect of this manipulation has to be at least 10% (i.e., 4 words) to be theoretically interesting. The right-hand panel shows the fit of a model in which such a 4-word effect is predicted. As before, the error bars are derived from the residual variation that is not explained by the model. Since the obtained effect in the figure is 1.2 words, the null model actually fits better. As before, the unexplained variation for the null model includes the sum of squares for error and that for the obtained effect, or 507.9 + 14.4 = 522.3. The unexplained variation for a model that predicts an effect of exactly 4 words includes the sum of squares for error, as well as the sum of squares for the overprediction. When the effect involves the comparison of two conditions, the overprediction sum of squares is

$$\begin{aligned} SS_{\text{OP}} &= \frac{n}{4}\left(\text{predicted} - \text{obtained}\right)^2 \\ &= \frac{40}{4}\left(4.0 - 1.2\right)^2 \\ &= 78.4. \end{aligned}$$

Thus, the total unexplained variation for the typeface model is 507.9 + 78.4 = 586.3, and the likelihood ratio in favor of the null model is

$$\begin{aligned} \lambda &= \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}}\right)^{\frac{n}{2}} \\ &= \left(\frac{SS_{\text{typeface}} + SS_{\text{error}}}{SS_{\text{OP}} + SS_{\text{error}}}\right) \\ &= \left(\frac{522.3}{586.3}\right)^{40/2} \\ &= 0.099, \end{aligned}$$

or a likelihood ratio of 10.1 in favor of the null model. In other words, the study provides clear evidence for the null model, relative to a model that predicts a 4-word effect. (Note that the magnitude of the effect is fixed in both models: In the null model, the effect is assumed to be 0, and in the typeface model, it is assumed to be 4. Thus, the models have the same number of parameters, and the likelihood ratio does not need to be corrected for model complexity.) This use of likelihood ratios represents a straightforward and intuitive method of documenting evidence for a null hypothesis over a theoretically plausible alternative. However, comparable uses of prior knowledge would also be a natural element of Bayesian inference methods.

Replication

Likelihood ratios represent a convenient means by which one may evaluate evidence for replication of previous findings. We argue that a suitable approach to such an evaluation is to combine the likelihood ratio derived from the current results with that derived from the previous findings. The aggregate likelihood ratio then constitutes an overall assessment of the evidence for the effect, in light of both studies together. As will be outlined below, likelihood ratios for independent sets of data can sometimes be combined simply by multiplying them together.

Our view of replication is at variance with what is occasionally suggested in some research reports, including some examples from the present survey. It is sometimes implied that replication involves finding significant effects where significant effects had been found before or finding nonsignificant effects where nonsignificant effects had previously been found. However, thinking of replication in terms of patterns of significance and nonsignificance is fallacious. For example, failing to reject a null hypothesis is often only weak evidence that the effect in question does not exist, and when power is weak to moderate, there is a substantial probability of failing to reject the null hypothesis even if the effect is real. Thus, even if a test is significant in one study, but not in another, the two studies could be quite consistent with one another. As a consequence, it is much better to think of replication in terms of patterns of means, rather than patterns of significance; from this perspective, one would say that a study replicates another if the patterns of means are similar in the two studies.

Because likelihood ratios summarize the evidence, rather than providing a binary decision, they lend themselves much more readily to cross-experiment comparisons of this sort. For example, suppose that one study provided clear evidence for an effect with a (corrected) likelihood ratio of 10 (which also happened to be statistically significant) and that another study failed to find clear evidence for that effect with a corrected likelihood ratio of 2 (which was not statistically significant). If one used patterns of significance to describe the results of the experiments, one might be led to the conclusion that the second study failed to replicate the first. However, an examination of the likelihood ratios would make it clear that the second study merely failed to find strong evidence *for* the effect; it did not find any evidence *against* the effect. Indeed, if it is sensible to assume that effect size might vary across experiments, one could interpret the two experiments as independent tests of whether an effect of some magnitude exists, and one could simply multiply the likelihood ratios to find the combined evidence for that effect. In this case, strong evidence for an effect (a likelihood ratio of 10), combined with weak evidence (a likelihood ratio of 2), leads to even stronger evidence for that effect (a likelihood ratio of 20). In a sense, this amounts to a small-scale form of meta-analysis.

Under other circumstances, one could use likelihood ratios to describe the evidence for failing to replicate, something not easily done with traditional statistics. To

do this using likelihood ratios, one needs to explicitly consider the patterns of means in the two experiments. The data shown in Figure 5 provide an illustration of how this might proceed. In this case, suppose that a prior study had actually shown an effect of 4 words and that one's current study obtained an effect of 1.2 words under comparable conditions. In order to assess the evidence for failing to replicate, one would compare a typeface model that predicts the obtained 1.2-word effect with a typeface model that assumes an effect of 4 words. From Table 4, the unexplained sum of squares for assuming the obtained 1.2-word effect is 507.9. As calculated previously, the unexplained sum of squares assuming a 4-word effect is 586.3. Thus, the likelihood ratio in favor of the model assuming a 1.2-word effect is

$$\lambda = \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}}\right)^{\frac{n}{2}}$$
$$= \left(\frac{586.3}{507.9}\right)^{\frac{40}{2}}$$
$$= 17.6.$$

In this case, the 1.2-word effect model has an additional free parameter: Model 1, which predicts an effect of 4 words a priori, has two parameters (the mean and the variance), whereas Model 2 has three (the mean, the variance, and the effect of typeface). Using Equation 6 to correct for this additional flexibility yields

$$\lambda_C = Q_C\lambda$$
$$= \exp\left[k_1\left(\frac{n}{n-k_1-1}\right) - k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda$$
$$= \exp\left[2\left(\frac{40}{40-2-1}\right) - 3\left(\frac{40}{40-3-1}\right)\right](17.6)$$
$$= 5.5.$$

Thus, one could conclude that there is clear evidence that the study failed to replicate the 4-word effect size that had been obtained previously. Because this claim refers to differences in the patterns of means, it is not prey to the fallacies involved in comparing patterns of significance.

Pro Forma Evaluations

Our survey of significance tests indicates that it is quite common to report pro forma statistical tests—that is, tests that do not correspond to plausible theoretical interpretations. Our view is that pro forma tests do not correspond to hypothesis testing in the usual sense. Indeed, researchers may fail to embrace the alternative hypothesis even when the results of the test are significant. Instead, we believe these tests are often reported as an indirect measure of the goodness of fit or the adequacy of a model. In particular, if one has a compelling model of the data that accounts for the important effects and interactions, there should be little in the way of systematic deviations among the remaining degrees of freedom. Thus, the reasoning might proceed as follows: Failing to show significant effects among those residual degrees of freedom provides some indication of the adequacy of the model. However, significance tests, being designed with other goals in mind, do not usually provide a suitable measure of goodness of fit, especially in an ANOVA context.

A more useful measure in this regard can be devised using likelihood ratios. The general strategy would be to compare a target model (specified, for example, in terms of a particular constellation of effects and interactions) with, for example, a full model that incorporates all possible degrees of freedom. Of course, because of the additional degrees of freedom, the full model will match the data better than will the target model, and the likelihood ratio will reflect that superior fit. The critical question is whether the additional degrees of freedom capture important, systematic variation or whether the improvement in fit is primarily due to fitting noise. Corrections such as Q_C (Equation 6) provide an index of how much improvement might be expected by chance and can be used to adjust the likelihood ratio. The adjusted likelihood ratio then provides a measure of the adequacy of the target model.

As a concrete example, suppose that one were investigating the effects of intervening activities on the recall of word lists. Subjects first learn a list of words and then perform one of two types of activities (solving anagrams or arithmetic problems) for one of two durations. The interesting results concern the recall of the words after this potentially interfering activity. However, one might also evaluate the subjects' recall immediately, before any intervening activity. These immediate recall data are shown in Figure 6, and the corresponding ANOVA table is shown in Table 5. In the figure, the groups of subjects are labeled by assigned conditions, but these are merely "dummy" labels, since the plotted data were collected prior to these manipulations. Any tests of effects among these conditions constitute pro forma tests, because there is no reason to expect any particular constellation of effects prior to the manipulations, and it is not clear what the interpretation would be if effects were found. However, one might be tempted to use such tests in this instance to provide evidence that the groups are comparable at the outset.

The alternative to significance tests that we propose would involve first calculating the likelihood comparing the target (null) model with a full model. The target model explains none of the effects or interactions, whereas the full model fails to explain only the error sum of square. Thus, from the data in Table 5, the likelihood ratio in favor of the full model would be

$$\lambda = \left(\frac{\text{Model 1 unexplained variation}}{\text{Model 2 unexplained variation}}\right)^{\frac{n}{2}}$$
$$= \left(\frac{SS_{\text{duration}} + SS_{\text{task}} + SS_{D\times T} + SS_{\text{error}}}{SS_{\text{error}}}\right)^{\frac{n}{2}}$$
$$= \left(\frac{65.2 + 119.9 + 29.6 + 1{,}533.5}{1{,}533.5}\right)^{\frac{40}{2}}$$
$$= 13.8.$$

In order to correct the likelihood ratio for the potentially gratuitous additional degrees of freedom, one could use Q_C from Equation 6. In this case, Model 1 has two parameters (the variance and the overall mean), and Model 2 has five (the variance plus the four condition means). Thus,

$$\begin{aligned}\lambda_C &= Q_C\lambda \\ &= \exp\left[k_1\left(\frac{n}{n-k_1-1}\right)-k_2\left(\frac{n}{n-k_2-1}\right)\right]\lambda \\ &= \exp\left(2\frac{40}{40-2-1}-5\frac{40}{40-5-1}\right)(13.8) \\ &= 0.33.\end{aligned}$$

Because this value is less than 1, it suggests that the null model (in which all the conditions are assumed to be the same) provides a reasonable match to the data.

This approach to evaluating the adequacy of a model is appropriate only if there is no specific, theoretically motivated alternative. In contrast, if there is a sensible alternative to the target model, the results could well be different. For example, in the results depicted in Figure 6, if the experimental groups that performed anagrams and arithmetic were treated differently during initial learning, there would be good reason to evaluate the evidence specifically for a main effect of that factor, and the conclusions might well differ.

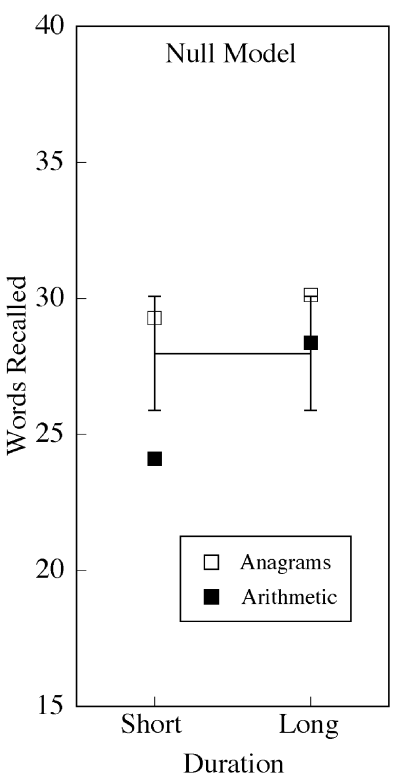

Figure 6. Evaluation of a null model for conditions that are not expected to differ.

Table 5
ANOVA Table for Evaluating Adequacy of a Model

Source	*df*	Sums of Squares	Mean Square	*F*	*p*
Duration (D)	1	65.3	65.3	1.53	.2236
Task (T)	1	119.9	119.9	2.81	.1021
D × T	1	29.6	29.6	0.69	.4100
Error	36	1,533.5	42.6		
Total	39	1,748.4			

CONCLUSIONS

We have shown here how likelihood ratios are derived, how they may be computed and interpreted, and how they can be used to fulfill the most common purposes of reporting statistics in empirical psychology. The critical ingredient in this approach is the incorporation of theoretical and conceptual knowledge in the reporting of evidence. We emphasize that the appropriate analysis of data cannot be described in the abstract and that there is no mechanical or "cook-book" method for dealing with empirical results. Rather, the analysis must depend on the theoretical context that motivated the experiment and its interpretation. In sum, the analysis and reporting of results is not a mechanical or purely objective process; it depends on the goals and arguments of the researcher. Likelihood ratios provide a useful tool from this perspective, because they merely summarize the evidence and can be readily adapted to various goals and arguments as the need arises.

REFERENCES

Adolphs, R., & Tranel, D. (1999). Preferences for visual stimuli following amygdala damage. *Journal of Cognitive Neuroscience*, **11**, 610-616.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory* (pp. 267-281). Budapest: Académiai Kiadó.

Arbuthnott, K., & Frank, J. (2000). Executive control in set switching: Residual switch cost and task-set inhibition. *Canadian Journal of Experimental Psychology*, **54**, 33-41.

Chochon, F., Cohen, L., van de Moortele, P., & Dehaene, S. (1999). Differential contributions of the left and right inferior parietal lobules to number processing. *Journal of Cognitive Neuroscience*, **11**, 617-630.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic Press.

De Gennaro, L., Ferrara, M., Urbani, L., & Bertini, M. (2000). A complementary relationship between wake and REM sleep in the auditory system: A pre-sleep increase of middle-ear muscle activity (MEMA) causes a decrease of MEMA during sleep. *Experimental Brain Research*, **130**, 105-112.

Diedrichsen, J., Ivry, R., Cohen, A., & Danziger, S. (2000). Asymmetries in a unilateral flanker task depend on the direction of the response: The role of attentional shift and perceptual grouping. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 113-126.

Dixon, P. (2001, June). *The logic of pro forma statistics*. Poster presented at the meeting of the Canadian Society for Brain, Behaviour, and Cognitive Science, Quebec.

Edwards, A. W. F. (1972). *Likelihood*. London: Cambridge University Press.

Estes, W. K. (1997). On the communication of information by displays of standard errors and confidence intervals. *Psychonomic Bulletin & Review*, **4**, 330-341.

Fisher, R. A. (1925). *Statistical methods for research workers*. New York: Hafner.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society*: Series B, **17**, 69-78.

Fugelsang, J. A., & Thompson, V. (2000). Strategy selection in causal reasoning: When beliefs and covariation collide. *Canadian Journal of Experimental Psychology*, **54**, 15-32.

Goodman, S. N., & Royall, R. (1988). Evidence and scientific research. *American Journal of Public Health*, **78**, 1568-1574.

Hoffmann, E. A., & Haxby, J. (2000). Distinct representations of eye gaze and identity in the distributed human neural system for face perception. *Nature Neuroscience*, **3**, 80-84.

Hurvich, C. M., & Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, **76**, 297-307.

Judd, C. M., & McClelland, G. H. (1989). *Data analysis: A model-comparison approach*. San Diego: Harcourt Brace Jovanovich.

Kinoshita, S. (2000). The left-to-right nature of the masked onset priming effect in naming. *Psychonomic Bulletin & Review*, **7**, 133-141.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, **5**, 161-171.

Loftus, G. R. (2001). Analysis, interpretation, and visual presentation of data. In H. Pashler & J. Wixted (Eds.), *Stevens' Handbook of experimental psychology* (3rd ed., pp. 339-390). New York: Wiley.

Lykken, D. E. (1968). Statistical significance in psychological research. *Psychological Bulletin*, **70**, 151-159.

Masson, M. E. J., & Loftus, G. R. (2003). Using confidence intervals for graphically based data interpretation. *Canadian Journal of Experimental Psychology*, **57**, 203-220.

Neyman, J., & Pearson, E. S. (1928). On the use and interpretation of certain test criteria for purposes of statistical inference. *Biometrika*, **20**, 175-240, 263-294.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London*: *Series A*, **231**, 289-337.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.

Prabhakaran, V., Narayanan, K., Zhao, Z., & Gabrieli, J. (2000). Integration of diverse information in working memory within the frontal lobe. *Nature Neuroscience*, **3**, 85-90.

Rissanen, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40-47.

Royall, R. M. (1997). *Statistical evidence: A likelihood paradigm*. London: Chapman & Hall.

Rozeboom, W. W. (1960). The fallacy of the null-hypothesis significance test. *Psychological Bulletin*, **57**, 416-428.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.

Servos, P. (2000). Distance estimation in the visual and visuomotor systems. *Experimental Brain Research*, **130**, 35-47.

Sivia, D. S. (1996). *Data analysis: A Bayesian tutorial*. Oxford: Oxford University Press.

Soto-Faraco, S. (2000). An auditory repetition deficit under low memory-load. *Journal of Experimental Psychology: Human Perception & Performance*, **26**, 264-278.

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals: An integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, **6**, 371-386.

Zheng, Y., Myerson, J., & Hale, S. (2000). Age and individual differences in visuospatial processing speed: Testing the magnification hypothesis. *Psychonomic Bulletin & Review*, **7**, 113-120.

APPENDIX A
Proof of Equation 2

The likelihood function for a normally distributed random variable is

$$f\left(x\middle|\mu,\sigma^2\right)=\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{\left(x-\mu\right)^2}{\sigma^2}\right].$$

For n independent random variables, the likelihood is the product of the likelihood functions for each variable:

$$\begin{aligned}f\left(x_1,x_2,\ldots,x_n\middle|\mu_1,\mu_2,\ldots,\mu_n,\sigma\right)&=\prod_i\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[-\frac{1}{2}\frac{\left(x_i-\mu_i\right)^2}{\sigma^2}\right]\\&=\left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}}\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\exp\left[-\frac{1}{2}\frac{\sum_i\left(x_i-\mu_i\right)^2}{\sigma^2}\right].\end{aligned}$$

The maximum likelihood for a set of observations can be found by using the maximum likelihood estimates for $\mu_1, \mu_2, \ldots, \mu_n, \sigma$. The maximum likelihood estimates $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_n$ will be constrained by the model under consideration. For example, if two conditions are not expected to differ, the estimated means for the observations in those conditions would be constrained to be the same. Furthermore, the maximum likelihood estimate of σ^2 is

$$\hat{\sigma}^2=\frac{\sum_i\left(x_i-\hat{\mu}_i\right)^2}{n}.$$

Substituting this value for $\hat{\sigma}^2$ in the expression for the maximum likelihood yields

$$\begin{aligned}f\left(x_1,x_2,\ldots,x_n\middle|\hat{\mu}_1,\hat{\mu}_2,\ldots,\hat{\mu}_n,\hat{\sigma}\right)&=\left[\frac{1}{\sum_i\left(x_i-\hat{\mu}_i\right)^2\Big/n}\right]^{\frac{n}{2}}\left(\frac{1}{2\pi}\right)^{\frac{n}{2}}\exp\left[-\frac{1}{2}\frac{\sum_i\left(x_i-\hat{\mu}_i\right)^2}{\sum_i\left(x_i-\hat{\mu}_i\right)^2\Big/n}\right]\\&=\left[\frac{1}{\sum_i\left(x_i-\hat{\mu}_i\right)^2}\right]^{\frac{n}{2}}\left(\frac{n}{2\pi}\right)^{\frac{n}{2}}\exp\left(-\frac{n}{2}\right)\\&=\left(\frac{1}{SSE}\right)^{\frac{n}{2}}B_n,\end{aligned}$$

where B_n depends only on n. Notice that SSE is the sum of the squared deviations from the means predicted by the model. The ratio of two likelihood functions, constrained by different models, is thus

$$\begin{aligned}\lambda&=\frac{\left(\frac{1}{SSE_2}\right)^{\frac{n}{2}}B_n}{\left(\frac{1}{SSE_1}\right)^{\frac{n}{2}}B_n}\\&=\left(\frac{SSE_1}{SSE_2}\right)^{\frac{n}{2}}.\end{aligned}$$

Because the sample variance is $s_2 = SSE/n - 1$, this is equivalent to

$$\begin{aligned}\lambda&=\left(\frac{SSE_1/n-1}{SSE_2/n-1}\right)^{\frac{n}{2}}\\&=\left(\frac{s_1^2}{s_2^2}\right)^{\frac{n}{2}}.\end{aligned}$$

APPENDIX A (Continued)

Also, because $1 - R^2 = SSE/SS_{\text{total}}$, this is the same as

$$\lambda = \left(\frac{SSE_1 / SS_{\text{total}}}{SSE_2 / SS_{\text{total}}} \right)^{\frac{n}{2}}$$

$$= \left(\frac{1 - R_1^2}{1 - R_2^2} \right)^{\frac{n}{2}}.$$

APPENDIX B
Proof of Equations 10 and 11

The total sum of squares for the regression problem is

$$SS_{\text{total}} = SS_{\text{linear}} + SS_{\text{quadratic}} + SS_{\text{error}}.$$

For the linear model, $1 - R_1^2$ is

$$1 - R_1^2 = \frac{SS_{\text{quadratic}} + SS_{\text{error}}}{SS_{\text{total}}}.$$

For the quadratic model, $1 - R_2^2$ is

$$1 - R_2^2 = \frac{SS_{\text{error}}}{SS_{\text{total}}}.$$

So, from Equation 2,

$$\lambda = \left(\frac{1 - R_1^2}{1 - R_2^2} \right)^{\frac{n}{2}} = \left(\frac{SS_{\text{quadratic}} + SS_{\text{error}}}{SS_{\text{error}}} \right)^{\frac{n}{2}}.$$

This can be transformed into an F ratio for the quadratic trend as follows:

$$df_{\text{error}} \left(\lambda^{\frac{2}{n}} - 1 \right) = df_{\text{error}} \left\{ \left[\left(\frac{SS_{\text{quadratic}} + SS_{\text{error}}}{SS_{\text{error}}} \right)^{\frac{n}{2}} \right]^{\frac{2}{n}} - 1 \right\}$$

$$= df_{\text{error}} \left(\frac{SS_{\text{quadratic}} + SS_{\text{error}}}{SS_{\text{error}}} - 1 \right)$$

$$= \frac{SS_{\text{quadratic}}}{SS_{\text{error}} / df_{\text{error}}},$$

which is distributed as $F(1, df_{\text{error}})$ when there is no quadratic trend. Solving for λ yields Equation 11.

(Manuscript received September 16, 2002;
revision accepted for publication October 20, 2003.)