The P_{rep} statistic as a measure of confidence in model fitting

F. GREGORY ASHBY AND JEFFREY B. O'BRIEN
University of California, Santa Barbara, California

In traditional statistical methodology (e.g., the ANOVA), confidence in the observed results is often assessed by computing the p value or the power of the test. In most cases, adding more participants to a study will improve these measures more than will increasing the amount of data collected from each participant. Thus, traditional statistical methods are biased in favor of experiments with large numbers of participants. This article proposes a method for computing confidence in the results of experiments in which data are collected from a few participants over many trials. In such experiments, it is common to fit a series of mathematical models to the resulting data and to conclude that the best-fitting model is superior. The probability of replicating this result (i.e., P_{rep}) is derived for any two nested models. Simulations and empirical applications of this new statistic confirm its utility in studies in which data are collected from a few participants over many trials.

During its first half century, research in experimental psychology was dominated by studies in which data were collected from just a few participants over many trials. Typically, the data from each participant were analyzed separately, and generalizations and conclusions were drawn on the basis of these analyses. A well-known and highly influential example of this approach can be found in Ebbinghaus's (1885/1913) classic book *Memory: A Contribution to Experimental Psychology*, which mostly describes a series of experiments with a single participant—namely, Ebbinghaus himself. During the last 50 years, there has been a dramatic shift away from such studies in favor of experiments in which many fewer data are collected from many more participants. This type of experiment has so come to dominate many areas of psychology that some journals virtually refuse to publish other kinds of studies.

Clearly, more data are almost always better than less data, so the optimal experiment might collect data from many participants over many trials. However, collecting data typically requires time, money, and effort. As a result, researchers must decide how to trade off the number of participants (i.e., N) with the number of trials on which data are collected from each participant (i.e., n). This article derives a statistical measure of confidence in the outcome of large n experiments in which data are collected from relatively few participants over many trials.

Large N studies have a number of attractive advantages. In experiments in which the interest is primarily in mean statistics, large N guarantees a small standard error and helps alleviate the normality assumptions endemic to parametric statistical inference. In addition, in many experiments, collecting large amounts of data from each participant is impossible, so large N designs are required. This is

the case, for example, in many studies in which participants from certain special populations (e.g., infants, the elderly, or various neuropsychological groups) have been used. Finally, the results of large N studies can be generalized from the particular sample of participants run in the study to the larger population from which they were drawn.

Large n studies offer their own advantages. First, because more data are collected from each participant, it is possible to examine dependent variables that require large sample sizes for accurate estimation. For example, estimating response time distributions requires many hundreds of trials, and articles in which the distributional properties of response times have been examined have frequently used large n designs (e.g., Ashby, Tein, & Balakrishnan, 1993; Maddox, Ashby, & Gottlob, 1998; Palmer, Huk, & Shadlen, 2005; P. L. Smith, Ratcliff, & Wolfgang, 2004). Second, large n designs are especially useful in tasks in which performance on early trials is not stable—for example, because of learning or because participants must adjust to the stimulus conditions. Third, averaging data across participants can sometimes change the psychological structure of the data (Ashby, Maddox, & Lee, 1994; Estes, 1956; Maddox, 1999; Myung, Kim, & Pitt, 2000), so studies whose aim is to test for the existence of such structures must use large n designs and perform single-participant analyses.

As the amount of data collected from each participant increases, experimenters often feel more confident that they have correctly identified the psychological structure (e.g., cognitive strategy) in each participant's data. Historically, however, there has been no convenient method to quantify this confidence. This article proposes a method for quantifying such confidence in large n studies in which nested models are fit to the data of each individual participant.

F. G. Ashby, ashby@psych.ucsb.edu

The method we propose is based on the statistic P_{rep} , which is defined as the probability that the same qualitative pattern of results would occur again if the experiment were replicated (Killeen, 2005a). When data are collected from each participant over many trials, it is common to fit quantitative models to the data from each individual participant, especially models that were designed specifically for the particular task under study. In this case, the most important statistical analysis is often to compare the goodness of fit of a variety of competing models. P_{rep} has been derived for many common inferential statistics, but to our knowledge, P_{rep} has not been derived for statistical tests of model comparison.

If some model (M_2) fits a data set better than another model (M_1), then the qualitative result of interest is that model M_2 again fits better than M_1 in the replicated experiment. In the next section, we will derive the probability of such replication for any pair of nested models. In the third and fourth sections of this article, we will explore the statistical properties of the model comparison P_{rep} and consider an empirical application. The fifth section compares this new P_{rep} with other possible measures of assessing confidence in the model selection process. Finally, we will close with some general comments and conclusions.

P_{rep} for Nested Models

Consider two models, M_1 and M_2 . Suppose that M_1 is a special case of M_2 (i.e., the models are nested). Suppose both models are fit to the data from each individual participant, using the method of maximum likelihood. Let L_i equal the likelihood of the data, given model M_i . We will consider methods of model comparison that favor the model with the smaller penalized fit, in which penalized fit takes the form

$$F_i = -2 \log L_i + Q_i, \quad (1)$$

where Q_i is the penalty incurred by model M_i . The model with the smaller value of F_i is chosen as the best-fitting model. For example, for the Akaike information criterion (AIC) fit statistic,

$$Q_i = 2r_i,$$

where r_i is the number of free parameters of model M_i . For the Bayesian information criterion (BIC) fit statistic,

$$Q_i = r_i \log n,$$

where, as before, n is the number of data trials used in the fitting process.¹

Under the null hypothesis that the simpler model M_1 is correct, the difference

$$\chi^2 = (-2 \log L_1) - (-2 \log L_2) \quad (2)$$

has an asymptotic χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the two models (i.e., equal to $r_2 - r_1$). Thus, the null hypothesis that model M_1 is correct is rejected in favor of the alternative that M_2 is correct if

$$(-2 \log L_1) - (-2 \log L_2) > \chi_{1-\alpha}^2(r_2 - r_1)$$

or, equivalently, if

$$(-2 \log L_1) > (-2 \log L_2) + \chi_{1-\alpha}^2(r_2 - r_1),$$

where $\chi_{1-\alpha}^2(r_2 - r_1)$ is the 100(1- α)% critical value of a χ^2 distribution with $r_2 - r_1$ degrees of freedom. Note that this χ^2 test is a special case of our penalized likelihood test in which $Q_1 = 0$ and $Q_2 = \chi_{1-\alpha}^2(r_2 - r_1)$.

In the Appendix, a general expression is derived for the probability of replicating a result (i.e., P_{rep}) that model M_i fits a particular data set better than model M_j (for i and $j = 1$ or 2 and $i \neq j$). In the special case in which the two models differ by only one free parameter, the probability of replicating the result that model M_2 fits better is shown to equal

$$\begin{aligned} P_{\text{rep}} &= P(F_2 < F_1) \\ &= P\left(Z > \sqrt{Q_2 - Q_1} - \sqrt{\chi_{\text{obs}}^2 - 1}\right) \\ &\quad + P\left(Z \leq \sqrt{Q_2 - Q_1} - \sqrt{\chi_{\text{obs}}^2 - 1}\right), \end{aligned} \quad (3)$$

where, as before, Q_i is the penalty associated with model M_i , and χ_{obs}^2 is the observed difference in the unpenalized fits of the two models (i.e., from Equation 2). Note that this probability can be computed from a Z table. Equation 3 assumes $\chi_{\text{obs}}^2 \geq 1$. In the simulations reported below, when $\chi_{\text{obs}}^2 < 1$, we set

$$\sqrt{\chi_{\text{obs}}^2 - 1} = 0.$$

For models that differ by only one parameter, the penalty difference $Q_2 - Q_1 = 1$ for the AIC goodness-of-fit test, $\log n$ for BIC, and for the χ^2 test $Q_2 - Q_1$ equals the 100(1- α)% critical value on a χ^2 distribution with one degree of freedom. Thus, if the difference between the raw fits of models M_1 and M_2 (i.e., χ_{obs}^2) remains constant as the number of trials n increases, P_{rep} will also remain constant for the AIC and χ^2 tests, and it will decrease for the BIC test. Therefore, P_{rep} for these three tests can increase with sample size only if χ_{obs}^2 increases.

Simulations

We will begin our investigations with simulated data that will allow us to examine some basic statistical properties of our P_{rep} estimate. Consider a category-learning task with two categories, A and B, each containing hundreds of exemplars that vary across trials on two stimulus dimensions and in which categorization accuracy is maximized by a linear decision bound. The actual category structures² used in our simulations were identical to those used by Ashby and O'Brien (2007). In that study, all participants receiving full feedback reliably learned the categories, and their performance gradually stabilized over the course of many trials. Performance on later trials was well described by a decision bound model. Decision bound models assume that each participant partitions the perceptual space into response regions that are separated by a decision bound (e.g., Ashby & Gott, 1988; Maddox & Ashby, 1993). On each trial, the participant determines

which region the percept is in and then emits the associated response. Two such models will be used in this article. The general linear classifier (GLC) assumes that participants use a linear decision bound of arbitrary slope and intercept. The GLC has three parameters: the slope and intercept of the linear decision bound and noise variance. The one-dimensional classifier (DIM) assumes that the decision bound is a vertical (or horizontal) line. The DIM has two parameters: the intercept of the decision bound

and the noise variance. Figure 1 shows the responses predicted by these two models in the absence of any noise, given the Ashby and O'Brien (2007) category structures. Stimuli that elicited a Category A response are denoted by pluses, stimuli that elicited a Category B response are denoted by dots, and the decision bound assumed by the model is denoted by the solid line.

Modeling data from hypothetical participants. To simulate the responses of a hypothetical participant in the

Figure 1. Decision strategies used to generate responses in the simulations. (A) Illustration of the GLC decision strategy. (B) Illustration of the DIM strategy. Also shown are the decision bounds for each model. Plus signs indicate Category A responses, and dots indicate Category B responses. The simulated data included perceptual and criterial noise, as well as guessing (see the text for details).

Table 1
True and Estimated P_{rep} From 1,000 Simulated Participants All Responding
According to the Equation 4 Model With the GLC Decision Strategy

	Low Guessing			High Guessing		
	χ^2	AIC	BIC	χ^2	AIC	BIC
True P_{rep}	.946	.946	.946	.787	.787	.787
Mean P_{rep} estimate (from Equation 3)	.949	.954	.947	.797	.820	.785
P_{rep} standard deviation (from Equation 3)	.215	.191	.224	.389	.344	.404

Ashby and O'Brien (2007) category-learning task, we used a simple model that assumes that, with probability p_k , the participant classifies stimulus X on trial k by using the decision bound shown in Figure 1 and, with probability $1 - p_k$, the participant guesses. To ensure that performance gradually stabilized over time, we assumed that p_k gradually increased with k . In particular, we assumed that the probability of responding A to the presentation of stimulus X on trial k is given by

$$P_k(A|X) = p_k P(A|\text{Figure 1 Bound} + \text{Noise}) + (1 - p_k) \frac{1}{2}, \quad (4)$$

where

$$p_k = 1 - e^{-\theta_k} \quad (5)$$

and

$$\theta_k = \theta_{k-1} + \theta_0, \quad (6)$$

where θ_0 is a constant. The term $P(A|\text{Figure 1 Bound} + \text{Noise})$ in Equation 4 denotes the probability of responding A according to the GLC (Figure 1A) or the DIM (Figure 1B) (when the noise variance was set to 21.2). This model is unrealistic in the sense that only one systematic response strategy is ever used. Nevertheless, it does display the realistic property that performance gradually stabilizes with practice.

Simulation 1: How accurate is the Equation 3 estimator? The statistic defined by Equation 3 has a number of properties that could limit its usefulness as an estimator of P_{rep} . First, it uses results from a single set of model fits to estimate χ^2_{obs} , which could make the estimator highly variable. Second, Equation 3 holds only for values of $\chi^2_{\text{obs}} \geq 1$. Finally, there are distributional assumptions underlying Equation 3 (see the Appendix) that are difficult to test directly. In our first set of simulations, therefore, we investigate the accuracy of the Equation 3 estimator.

Using the model of response generation described by Equations 4–6, we simulated the performance of four sets

of 1,000 identical (except for the noise samples) participants over the course of 850 trials each. The four data sets were created by factorially combining two levels of guessing [$\theta_0 = .0023055$ (low guessing) vs. $\theta_0 = .001537$ (high guessing); note from Equations 4–6 that θ_0 is inversely related to amount of guessing] with two decision strategies (GLC vs. DIM). The choice of decision strategy determined which Figure 1 decision bound was used in the Equation 4 model. Next, for each hypothetical participant, we separately fit both the GLC and the DIM models, using the χ^2 , AIC, and BIC goodness-of-fit measures. These fit values were used to determine the best-fitting model for each participant and to estimate a P_{rep} for each participant from Equation 3. Finally, after all the simulations were complete, for each set of 1,000 hypothetical participants, the proportion of participants whose data were best fit by the model assuming the correct decision strategy was computed. This number should be close to the true value of P_{rep} because it tells us the probability that a random sample from the pool of 1,000 hypothetical participants will produce one whose data replicate the finding that the correct model provides the better fit. In addition to the true value of P_{rep} , we also computed the means of all 1,000 P_{rep} estimates that were generated from Equation 3, along with the standard deviations of these 1,000 estimates. By comparing the estimated and true values of P_{rep} , we can evaluate the accuracy of the Equation 3 estimator.

Table 1 shows these results for the hypothetical participants who used the GLC decision strategy. Note that the estimated values of P_{rep} are all very close to the true values. Thus, under these conditions, Equation 3 provides an accurate estimate of P_{rep} . On the other hand, note that the standard deviations of the Equation 3 P_{rep} estimates are large. We will discuss the implications of this result in the next section.

Table 2 shows the results of our simulations for those hypothetical participants who used the simpler DIM decision strategy. The results here are quite different. Note that now the P_{rep} estimates are all less than the true values of P_{rep} . This suggests a systematic bias in our P_{rep} estimate in situations in which the simpler model is correct. Of course,

Table 2
True and Estimated P_{rep} From 1,000 Simulated Participants All Responding
According to the Equation 4 Model With the DIM Decision Strategy

	Low Guessing			High Guessing		
	χ^2	AIC	BIC	χ^2	AIC	BIC
True P_{rep}	.962	.867	.993	.949	.873	.994
Mean P_{rep} estimate (from Equation 3)	.900	.775	.966	.904	.781	.967
P_{rep} standard deviation (from Equation 3)	.118	.147	.069	.120	.146	.074

the simpler model can never provide a better absolute fit to the data than the more complex model. So the only way the simpler model can be favored in goodness-of-fit testing is if the difference in absolute fits between the simple and the complex models is less than the extra penalty paid by the complex model for its extra parameters. For this reason, we should expect a small value of χ^2_{obs} when the simpler model fits better. The bias shown in Table 2 occurs because, for many of these hypothetical participants, χ^2_{obs} was less than 1. As was mentioned above, in this case, the Equation 3 estimator of P_{rep} is undefined. To circumvent this problem, whenever $\chi^2_{\text{obs}} < 1$, we set

$$\sqrt{\chi^2_{\text{obs}} - 1} = 0$$

in Equation 3. This approximation has the effect of underestimating the true value of P_{rep} . Fortunately, however, this systematic bias produces a conservative estimate of P_{rep} —that is, the true P_{rep} will tend to be larger than the estimated P_{rep} .

Overall, the results of the simulations described in Tables 1 and 2 are encouraging. When the more complex model is correct, accurate, albeit variable, P_{rep} estimates should be expected. When the simpler model is correct, the P_{rep} estimates are biased, but biased in the direction that makes use of the P_{rep} estimator conservative.

Simulation 2: The effects of increasing the number of trials. Next, we will investigate the effect of increasing the number of trials on which data are collected (i.e., n) on the P_{rep} estimate. Following the same procedures as in Simulation 1, we simulated the performance of two sets of 50 identical participants over the course of 1,000 trials each (with θ_0 from Equation 6 set to .001537). One set of participants used the GLC decision strategy, and the other set used the DIM strategy. For each hypothetical participant, we fit both the GLC and the DIM separately to 20 subsets of data of increasing sample size. In particular, the models were first fit to the data from the first 50 trials, then to the data from Trials 1–100, then to the data from Trials 1–150, and so on—each time adding data from the next 50 trials. Finally, P_{rep} s were computed from Equation 3 for each set of model fits. Note that because we used the model described by Equations 4–6 to generate these hypothetical data, the amount of guessing decreases with each successive block of 50 trials. Thus, sample size and amount of guessing are confounded in these simulations. We made this choice in an effort to simulate the common finding that the responses of human participants in such experiments gradually become more stable with practice (e.g., Ashby & Maddox, 1992).

Figure 2 shows the means and standard deviations of the estimated P_{rep} s for each block of trials. The top panel summarizes the results of fitting the models to simulated data from participants using the GLC strategy. In this case, P_{rep} is the probability of replicating the result that the GLC fits better than the DIM. The error bars depict the standard deviations of the 50 P_{rep} s from the AIC test. The bottom panel shows corresponding results from fitting the models to simulated data for which the DIM decision strategy was assumed. In this case, P_{rep} is the

probability of replicating the result that the DIM fits better than the GLC. Now the error bars denote the standard deviations from the BIC test.

Note first that when the GLC is the correct model (Figure 2, top panel), P_{rep} consistently increases with sample size for all three fit measures (AIC, BIC, and χ^2). Second, note that the AIC test produces the largest P_{rep} s, except when sample size is large, in which case all three tests produce P_{rep} s that are nearly 1.0. Because the GLC is the true model for these data, it will tend to provide a better absolute fit than the incorrect DIM, but because the GLC is more complex, it also carries a higher penalty. The outcome of the competition between better fit and increased penalty depends on sample size. For small n , both models fit well, so the penalties dominate (e.g., with two data points, both models fit perfectly). In this case, the simpler DIM model is frequently chosen as best fitting, and the probability of replicating the result that the more complex (and true) GLC model fits best is low. When n is large, however, the simpler DIM model tends to fit much worse than the GLC and, therefore, the raw fit difference determines model selection. In this case, P_{rep} is high.

Finally, note that the variability of the P_{rep} estimates is extremely low when P_{rep} is near 0 or 1 but quite large when P_{rep} is near .5. This pattern mimics the variance of the binomial distribution, which is greatest when $p = .5$ and decreases toward 0 as p approaches 0 or 1. The large P_{rep} variance when P_{rep} is near .5 is not surprising, given that our estimate of χ^2_{obs} for each simulated participant is based on the difference between a single pair of model fits. As was noted previously, this same effect is seen clearly in Tables 1 and 2. The large variance when P_{rep} estimates are of intermediate magnitude suggests that the statistic may be most useful when the P_{rep} estimates are large. When numerical estimates of P_{rep} are less than, say, .9, the P_{rep} estimate from an individual must be interpreted with caution. In this case, it may be better to compute the mean P_{rep} across all participants, since this averaging procedure will reduce the standard error of the estimate by $1/\sqrt{N}$.

The bottom panel of Figure 2 shows that when the DIM is the correct model, the P_{rep} estimates for the event that the DIM fits better than the GLC are uniformly high for all sample sizes. Even so, BIC consistently dominates the other tests, and AIC always has the lowest P_{rep} s. In fact, the AIC P_{rep} s actually drop by around .1 as sample size increases. When the simpler model is correct, both models will provide good fits to the data (since the complex model can always mimic the structure of the simple model). The simpler model, though, carries a smaller penalty, so it will tend to be favored for all sample sizes.

The failure of the P_{rep} estimates shown in the bottom panel of Figure 2 to increase to 1 is probably due to outliers in the data. As n increases, it is natural to expect the number of outliers to also increase. The effect of outliers on raw fit (i.e., on likelihood L_i) is nonlinear with distance; that is, a mispredicted response that is a distance D from the decision bound increases $-2 \log L_i$ more than two mispredicted re-

Figure 2. Top: P_{rep} estimates for ever-increasing sample sizes for the event that the GLC fit better than the DIM when the data were simulated from the GLC decision model (Figure 1A). The error bars depict the P_{rep} standard deviations for the AIC test. Bottom: P_{rep} estimates for the event that the DIM fit better than the GLC when the data were simulated from the DIM (Figure 1B). The error bars depict the P_{rep} standard deviations for the BIC test.

sponses that are both distance $D/2$ from the decision bound. This nonlinearity imparts an advantage to more flexible models that can minimize the effects of such outliers.

Figure 2 shows that when the more complex model is correct, P_{rep} is highest with AIC, whereas when the simpler model is correct, P_{rep} is highest with BIC. Of course, in real applications, one does not know which model is correct. As a result, a natural question to ask is which test

performs best overall. Figure 3 shows mean P_{rep} s from the top and bottom panels of Figure 2 for each block of data. These must be interpreted with caution due to the differing amounts of variability associated with the P_{rep} s from the top and bottom panels of Figure 2. Note that when sample size is small, all three tests perform equally poorly. At intermediate sample sizes, BIC has a small disadvantage, but at large sample sizes, BIC is superior.

Figure 3. Mean P_{rep} s from the top and bottom panels of Figure 2 for ever-increasing sample sizes.

For the large n studies that are the focus of this article, the simulations indicate that it may be useful to compute P_{rep} on differences between BIC scores of nested models. If one of the two models has more psychological validity than the other, the resulting P_{rep} s are likely to be near 1 and have small variance—a result that should increase confidence in the conclusions of the study. In the next section, this prediction will be tested with some real data.

An Empirical Application

As an empirical application, we will consider the category-learning study reported by Ashby and O'Brien (2007). In one condition of this experiment, each of 4 participants completed five identical experimental sessions of 600 trials each. On each trial, a stimulus was randomly selected from one of two categories and shown to the participant, whose task was to assign it to Category A or B by depressing the appropriate response key. Feedback about response accuracy was given on every trial. Each stimulus was a circular disk containing alternating light and dark bars. The stimuli varied across trials in bar width and bar orientation, and each category contained hundreds of exemplars. The category structure was the same as in Figure 1 (see note 2). In the course of data analysis, the GLC and the DIM were fit to the data from each experimental session of every participant.

The GLC provided a better fit than the DIM for each participant in every session. This was true for BIC, AIC, and the χ^2 test. Thus, one might naturally question the need to run these participants for more than the traditional single session. However, before drawing this conclusion, the P_{rep} s should be examined. These are shown in Table 3. For each of Sessions 2–5, the P_{rep} s for every participant were always greater than .999. Therefore, Table 3 shows only the P_{rep} s for Session 1.

If $P_{rep}(i)$ is the probability that the GLC is again better than the DIM when the experiment is replicated for participant i , then the probability of replicating the observed result that the data of all 4 participants are better fit by the GLC is

$$P_{rep}(\text{group}) = \prod_{i=1}^4 P_{rep}(i).$$

For Days 2–5, this product was greater than .999 for all three fit measures. Thus, it is virtually certain that if we were to replicate this experiment, all 4 participants would again produce data that were better fit by the GLC.

In Session 1, the P_{rep} estimates are low enough that we must worry about their variability. To reduce the standard error of these estimates, we can use the mean P_{rep} across the 4 participants as an estimate of the average P_{rep} for each participant. Assuming that this is the correct value for each participant, the probability of replicating the result that the GLC fits best for all 4 participants during Session 1 is .662 when the BIC fit statistic is used, .771 for AIC, and .613 if the models are compared via a χ^2 test. Had we run this experiment as a traditional large N study (i.e., where each participant completes a single experi-

Table 3
 P_{rep} s From Session 1 of the Category-Learning Experiment of Ashby and O'Brien (2007)

Participant	BIC	AIC	χ^2
1	.940	.996	.983
2	1.000	1.000	1.000
3	.667	.753	.555
4	1.000	1.000	1.000
Mean	.902	.937	.885
Product	.627	.750	.546

Note— P_{rep} s from Sessions 2–5 were all greater than .999.

mental session), with, say, 30 participants, we could estimate the probability that the GLC would have provided the better fit for all participants to be only .045 for the BIC test, .142 for the AIC test, and .026 for the χ^2 test. Thus, it is highly likely that the DIM would have provided the best fit for at least some of the participants in such a study.

Of course, one significant advantage of large N , single-session experiments is that the results can be generalized to the population from which the participants were selected. Within ANOVA, the authority to generalize to the population depends critically on three assumptions. First, the unique abilities of the participants (i.e., the treatment effects) must be normally distributed. Second, all members of the population must be equally likely to be sampled, and third, all samples must be drawn independently. Because these assumptions are surely not met in most large N studies, the validity of any generalization to the population is typically compromised in some unknown way. Even if the assumptions are met, however, one might legitimately question the validity of any generalization that might be made to, say, the population of hundreds of millions of western young adults from any results that might be obtained from a sample of 30, or even 100, participants.

In the present case, our P_{rep} analysis suggests that in a single-session version of the Ashby and O'Brien (2007) study, the data from some participants would have favored the GLC, whereas the data from other participants would have favored the DIM. Even if the number of participants better fit by the GLC was significantly greater than the number better fit by the DIM, we would be left with two very different but equally plausible conclusions. One possibility would be that all the participants were using a GLC strategy but that the DIM fit better for some participants because of noise. Another possibility, however, would be that there are at least two subpopulations of participants, who learn in qualitatively different ways. It would be impossible to decide between these two alternatives on the basis of such an experiment. In contrast, the multisession small N study summarized in Table 1 definitively rules out the possibility that these 4 participants belong to more than one heterogeneous subpopulation. It is possible, of course, that such subpopulations exist but that they were not represented among the participants selected for this study. It is important to realize, however, that this same omission, although less likely, could also occur in many large N studies. For example, a subpopulation with a low base rate in North America might easily be missed in a study that includes 100 undergraduates enrolled in the local introductory psychology course.

Comparisons With Other Methods of Assessing Confidence in Model Fitting

Given the goodness-of-fit measures that we considered in this article, we know of two alternative measures of assessing confidence in the outcome of a model-fitting exercise. In the case of the χ^2 test, one could compute the p value of the test—that is, the probability of observing an outcome as extreme or more extreme than the obtained value of χ^2_{obs} , given that the more restricted model is correct. Unfortunately, however, for the present purposes, p values

have two significant weaknesses. First, the AIC and BIC tests have no p values (i.e., these statistics are not used to test a null hypothesis). Second, the use of p values to assess confidence has been highly criticized, and in fact, P_{rep} was proposed specifically as an alternative that overcomes many of the weaknesses associated with p values (Killeen, 2005a). As just one well-known example, when the effect size is small, increasing sample size will typically reduce the p value of a test to an arbitrarily small value. The widely held belief, however, is that the probability of replicating small effects is low (i.e., P_{rep} would be low in this case), so most researchers probably would not express a high degree of confidence in the outcome of such an experiment.

A Bayesian and much preferred alternative to the p value is based on the difference in BIC scores. If the prior probability that model M_1 is correct is equal to the prior probability that model M_2 is correct, then under certain technical conditions (e.g., Raftery, 1995), it can be shown that³

$$P(M_i | \text{Data}) \doteq \frac{1}{1 + \exp\left[-\frac{1}{2}(\text{BIC}_2 - \text{BIC}_1)\right]}, \quad (7)$$

where $P(M_i | \text{Data})$ is the probability that model M_i is correct (for $i = 1$ or 2), assuming that either model M_1 or M_2 is correct. Thus, for example, if model M_1 is favored by a BIC difference of 2, the probability that model M_1 is correct is approximately .73, whereas if the BIC difference is 10, this probability is approximately .99.

To compare this measure with P_{rep} , we replicated the Simulation 2 analyses shown in Figure 2, except that we replaced P_{rep} on the ordinate with the Equation 7 Bayesian probability. The results are shown in Figure 4. A comparison of Figures 2 and 4 shows that both P_{rep} and the Bayesian probability behave in a qualitatively similar fashion. For data generated from the GLC, both probabilities are initially near zero, and then both climb steeply to an eventual asymptote near 1. In contrast, when the data were generated from the simpler DIM, both probabilities are uniformly high for all sample sizes.

A more detailed comparison of these alternative probabilities can be made by plotting the two probabilities against each other. In the case of the DIM data, such a comparison is not useful, because both probabilities range over such a small set of values. Figure 5 therefore plots P_{rep} against the Equation 7 Bayesian probability for the Simulation 2 data that were generated from the GLC. Except for very large values of the Bayesian probability, note that this plot is essentially linear, indicating an extremely high correlation between these two confidence measures. It is also important to note that for all sample sizes, P_{rep} is more conservative than the Bayesian probability, since the Bayesian probability is always larger than P_{rep} .

A critical assumption of Equation 7 is that the two models have the same a priori probability of being correct. The requirement that priors must be specified in Bayesian approaches has long been a source of controversy and is likely the primary reason that many statisticians remain skeptical of Bayesian methods (see, e.g., Dennis, 2004; Killeen, 2005b). For example, in our empirical application, virtually all popular theories of category learning

Figure 4. Bayesian probability (from Equation 7) that the model generating the data is the correct model for the data from Simulation 2. Compare with Figure 2.

predict that the GLC is more likely to be correct for these category structures than the DIM. This includes prototype theory (e.g., Homa, Sterling, & Trepel, 1981; Posner & Keele, 1968; Reed, 1972; Rosch, 1973; J. D. Smith & Minda, 1998), exemplar theory (Brooks, 1978; Estes, 1986; Hintzman, 1986; Lamberts, 2000; Medin & Schaffer, 1978; Nosofsky, 1986), and the multiple systems theory of COVIS (Ashby, Alfonso-Reese, Turken, & Waldron, 1998). All of these theoretical predictions are based on many empirical studies—that is, on prior observations of behavior in studies like the one that was the focus of our empirical application. Thus, the typical Bayesian assumption that the two models have the same a priori probability of being correct is inconsistent with prior data from similar tasks. But to our knowledge, no objective method has been proposed for adjusting the prior probabilities in a way that reflects current theoretical beliefs in the field. For this reason, we believe that the P_{rep} estimate proposed here, which does not require specifying prior probabilities, is a useful alternative to Bayesian approaches.

Another critical assumption of the Equation 7 Bayesian probability is that one of the M_1 or M_2 models is correct. This same assumption was required in our P_{rep} derivation. Although this is a common assumption in traditional methods of model selection, it obviously would be desirable to have a method that does not require assuming that

one of the specified models is correct. For a promising approach that relaxes this assumption, see Golden (2000) and Vuong (1989).

Conclusions

In traditional statistical methodology (e.g., ANOVA), confidence in the observed results is often assessed by computing the p value of the observed test statistic and/or the power of the statistical test. In most cases, adding more participants to a study will improve these measures more than will increasing the amount of data collected from each participant, even when both approaches add the same amount of data. Thus, traditional statistical methods are biased in favor of experiments with large numbers of participants. This article has proposed a method for computing confidence in the results of experiments in which data are collected from a few participants over many trials. In such experiments, it is common to fit a series of mathematical models to the resulting data and to conclude that the best-fitting model provides the best account of the participants' behavior. In this case, the main statistical conclusion of the study is that some model— M_2 , say—fit better than another model, M_1 . In this article, we derived the probability of replicating this result (i.e., P_{rep}) in the case in which M_2 and M_1 are nested models. We also reported a set of simulations and empirical applications that confirmed

Figure 5. Simultaneous values of the BIC P_{rep} and the Bayesian probability of Equation 7 for the data from Simulation 2 that were generated from the GLC.

the utility of this P_{rep} statistic in studies in which data are collected from a few participants over many trials.

If P_{rep} is high, one can be confident that the main result of the study, perhaps that model M_2 provides a better account of participants' behavior than model M_1 , would be supported if the experiment were replicated. Killeen (2005a) argued persuasively that the P_{rep} probability is more meaningful than the traditional p value. With enough participants, the p value can be small, even for small effect sizes that are unlikely to replicate.

On the other hand, it is also important to acknowledge the weaknesses and limitations of the P_{rep} statistic proposed here. For example, Figure 2 shows that the P_{rep} estimate is highly variable, especially when it is of intermediate magnitude. In addition, the P_{rep} statistic derived in this article requires nested models and maximum likelihood parameter estimation, and as such, it applies to only a restricted subset of modeling applications. Finally, it is important to note that the nested-models P_{rep} is susceptible to the same criticisms and controversies that were raised in response to Killeen's (2005a) original P_{rep} proposal (e.g., Cumming, 2005; Doros & Geier, 2005; Killeen, 2005b; Macdonald, 2005). At least some of that controversy arose because of Killeen's (2005a) proposal that P_{rep} should replace an established statistical methodology (i.e., null hypothesis significance testing). In the present application, a p value could be computed for the χ^2 significance test on nested models, but the more popular AIC and BIC tests do not produce p values. Thus, in modeling applications, P_{rep} offers a solution to a problem that is outside the scope of traditional null hypothesis significance testing.

In many cases, the most appropriate experiment requires collecting data from each participant over many trials (e.g., in studies of learning, or when participants must adapt to stimulus conditions). We hope that the P_{rep} statistic proposed here will allow researchers to have as much confidence in the results of such studies as in those of more traditional experiments that collect limited amounts of data from many participants.

AUTHOR NOTE

This research was supported in part by Public Health Service Grant MH3760-2. We thank Peter Killeen, Michael Lee, and Todd Maddox for their helpful comments. Correspondence concerning this article should be addressed to F. G. Ashby, Department of Psychology, University of California, Santa Barbara, CA 93106 (e-mail: ashby@psych.ucsb.edu).

REFERENCES

- ASHBY, F. G., ALFONSO-REESE, L. A., TURKEN, A. U., & WALDRON, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*, 442-481.
- ASHBY, F. G., & GOTT, R. E. (1988). Decision rules in the perception and categorization of multidimensional stimuli. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *14*, 33-53.
- ASHBY, F. G., & MADDOX, W. T. (1992). Complex decision rules in categorization: Contrasting novice and experienced performance. *Journal of Experimental Psychology: Human Perception & Performance*, *18*, 50-71.
- ASHBY, F. G., MADDOX, W. T., & LEE, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144-151.
- ASHBY, F. G., & O'BRIEN, J. B. (2007). The effects of positive versus negative feedback on information-integration category learning. *Perception & Psychophysics*, *69*, 865-878.
- ASHBY, F. G., TEIN, J. Y., & BALAKRISHNAN, J. D. (1993). Response time

- distributions in memory scanning. *Journal of Mathematical Psychology*, **37**, 526-555.
- BROOKS, L. (1978). Nonanalytic concept formation and memory for instances. In E. Rosch & B. B. Lloyd (Eds.), *Cognition and categorization* (pp. 169-211). Hillsdale, NJ: Erlbaum.
- CUMMING, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, **16**, 1002-1004.
- DENNIS, B. (2004). Statistics and the scientific method in ecology. In M. L. Taper & S. R. Lele (Eds.), *The nature of scientific evidence: Statistical, philosophical, and empirical considerations* (pp. 327-359). Chicago: University of Chicago Press.
- DOROS, G., & GEIER, A. B. (2005). Probability of replication revisited: Comment on "An alternative to null-hypothesis significance tests." *Psychological Science*, **16**, 1005-1006.
- EBBINGHAUS, H. (1913). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Columbia University, Teachers College. (Original work published 1885)
- ESTES, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.
- ESTES, W. K. (1986). Array models for category learning. *Cognitive Psychology*, **18**, 500-549.
- FEDER, P. I. (1968). On the distribution of the log likelihood ratio test statistic when the true parameter is "near" the boundaries of the hypothesis regions. *Annals of Mathematical Statistics*, **39**, 2044-2055.
- GOLDEN, R. M. (2000). Statistical tests for comparing possibly misspecified and nonnested models. *Journal of Mathematical Psychology*, **44**, 153-170.
- HINTZMAN, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review*, **93**, 411-428.
- HOMA, D., STERLING, S., & TREPEL, L. (1981). Limitations of exemplar-based generalization and the abstraction of categorical information. *Journal of Experimental Psychology: Human Learning & Memory*, **7**, 418-439.
- JOHNSON, N. L., KOTZ, S., & BALAKRISHNAN, N. (1994). *Continuous univariate distributions* (Vol. 2, 2nd ed.). New York: Wiley.
- KILLEEN, P. R. (2005a). An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345-353.
- KILLEEN, P. R. (2005b). Replicability, confidence, and priors. *Psychological Science*, **16**, 1009-1012.
- LAMBERTS, K. (2000). Information-accumulation theory of speeded categorization. *Psychological Review*, **107**, 227-260.
- MACDONALD, R. R. (2005). Why replication probabilities depend on prior probability distributions: A rejoinder to Killeen (2005). *Psychological Science*, **16**, 1007-1008.
- MADDOX, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354-374.
- MADDOX, W. T., & ASHBY, F. G. (1993). Comparing decision bound and exemplar models of categorization. *Perception & Psychophysics*, **53**, 49-70.
- MADDOX, W. T., ASHBY, F. G., & GOTTLÖB, L. R. (1998). Response time distributions in multidimensional perceptual categorization. *Perception & Psychophysics*, **60**, 620-637.
- MEDIN, D. L., & SCHAFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
- MYUNG, I. J., KIM, C., & PITT, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, **28**, 832-840.
- NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
- PALMER, J., HUK, A. C., & SHADLEN, M. N. (2005). The effect of stimulus strength on the speed and accuracy of a perceptual decision. *Journal of Vision*, **5**, 376-404.
- POSNER, M. I., & KEELE, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, **77**, 353-363.
- RAFTERY, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology*, **25**, 111-163.
- REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.
- ROSCH, E. (1973). Natural categories. *Cognitive Psychology*, **4**, 328-350.
- SMITH, J. D., & MINDA, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **24**, 1411-1430.
- SMITH, P. L., RATCLIFF, R., & WOLFGANG, B. J. (2004). Attention orienting and the time course of perceptual decisions: Response time distributions with masked and unmasked displays. *Vision Research*, **44**, 1297-1320.
- VUONG, Q. H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica*, **57**, 307-333.
- WALD, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, **54**, 426-482.

NOTES

1. The BIC statistic is defined in a variety of different ways. For example, it is common to define L_i as the ratio of the likelihood of model M_i to the likelihood of the saturated model in which there is one free parameter to every data point, or as the ratio of the likelihood of model M_i to the likelihood of the null model in which there are no free parameters (e.g., Raftery, 1995). The important point, however, is that in all these definitions, the model with the smaller BIC is preferred, and the numerical value of the difference in BICs for two models is the same.
2. Category exemplars were generated using the randomization technique (Ashby & Gott, 1988) in which each category is defined as a bivariate normal distribution along the two stimulus dimensions. Category A had an orientation mean of 34 and a bar width mean of 56, whereas Category B had means on these two dimensions of 56 and 34, respectively. In both categories, the variance on each dimension was 235, and the covariance between dimensions was 19. The optimal category bound has a slope of 1 and an intercept of 0, and a participant responding A to any exemplar above this bound and B to any exemplar below the bound would achieve 86% correct. See Ashby and O'Brien (2007) for more details.
3. It is straightforward to derive the following equation from Equation 22 in Raftery (1995).

APPENDIX

Consider the case in which two models, M_1 and M_2 , are being compared and in which model M_1 is a special case of model M_2 (i.e., M_1 is nested under M_2). If the null hypothesis that M_1 is the correct model is true, it is well known that the statistic

$$\chi^2 = (-2 \log L_1) - (-2 \log L_2)$$

has an asymptotic χ^2 distribution with $r_2 - r_1$ degrees of freedom. If, on the other hand, the null hypothesis is false, then under the one extra assumption that model M_2 is instead true, the χ^2 statistic has an asymptotic non-central χ^2 distribution with $r_2 - r_1$ degrees of freedom and noncentrality parameter δ (e.g., Feder, 1968; Wald, 1943). The numerical value of δ is equal to the true (i.e., population expected value) difference in these model fits minus the number of degrees of freedom—that is, to

$$\delta = E[(-2 \log L_1) - (-2 \log L_2)] - (r_2 - r_1), \quad (\text{A1})$$

where E is expected value.

Consider a data set for which $F_2 < F_1$ —that is, for which the more complex model M_2 provides the better fit. The statistic P_{rep} is defined as the probability that model M_2 would provide the better fit again if the experiment were replicated. Thus,

$$\begin{aligned} P_{\text{rep}} &= P(F_2 < F_1) \\ &= P(F_1 - F_2 > 0) \\ &= P[(-2 \log L_1) - (-2 \log L_2) - (Q_2 - Q_1) > 0] \\ &= P[\chi^2(r_2 - r_1, \delta) > (Q_2 - Q_1)], \end{aligned}$$

where $\chi^2(r_2 - r_1, \delta)$ is a random variable with a noncentral χ^2 distribution with $r_2 - r_1$ degrees of freedom and noncentrality parameter δ given by Equation A1. Now $E[(-2 \log L_1) - (-2 \log L_2)]$ can be estimated by the observed difference in the (nonpenalized) fit values of models M_1 and M_2 , which we denote by χ_{obs}^2 . Thus,

$$\hat{\delta} = \chi_{\text{obs}}^2 - (r_2 - r_1). \quad (\text{A2})$$

Therefore, following standard practice (e.g., Killeen, 2005a), P_{rep} can be approximated by

$$P_{\text{rep}} \doteq P[\chi^2(r_2 - r_1, \hat{\delta}) > (Q_2 - Q_1)]. \quad (\text{A3})$$

This probability can be computed using a variety of different numerical approximations to the cumulative distribution function of the noncentral χ^2 distribution (e.g., Johnson, Kotz, & Balakrishnan, 1994).

In the case in which model M_2 has only one more parameter than model M_1 , the P_{rep} approximation of Equation A3 simplifies considerably. Under these conditions, the noncentral χ^2 has one degree of freedom (i.e., $r_2 - r_1 = 1$) and can be expressed as

$$\chi^2(1, \delta) = (Z + \mu)^2,$$

where Z has a standard normal distribution (i.e., with a mean of 0 and a variance of 1) and $\delta = \mu^2$ (e.g., Johnson et al., 1994). Following Equation A2, μ can be estimated by

$$\hat{\mu} = \sqrt{\chi_{\text{obs}}^2 - 1}.$$

Therefore, when the nested models differ by only one free parameter,

$$\begin{aligned} P_{\text{rep}} &= P[\chi^2(1, \delta) > Q_2 - Q_1] \\ &= P[(Z + \mu)^2 > Q_2 - Q_1] \\ &\doteq P[(Z + \hat{\mu})^2 > Q_2 - Q_1] \\ &= P[Z > \sqrt{Q_2 - Q_1} - \hat{\mu}] + P[Z \leq -\sqrt{Q_2 - Q_1} - \hat{\mu}] \\ &= P[Z > \sqrt{Q_2 - Q_1} - \sqrt{\chi_{\text{obs}}^2 - 1}] + P[Z \leq -\sqrt{Q_2 - Q_1} - \sqrt{\chi_{\text{obs}}^2 - 1}], \end{aligned} \quad (\text{A4})$$

which can be computed from a Z table.