Psychonomic Bulletin & Review
2008, 15 (4), 692-712
doi: 10.3758/PBR.15.4.692

Model evaluation using grouped or individual data

Andrew L. Cohen
University of Massachusetts, Amherst, Massachusetts

and

Adam N. Sanborn and Richard M. Shiffrin
Indiana University, Bloomington, Indiana

Analyzing the data of individuals has several advantages over analyzing the data combined across the individuals (the latter we term *group analysis*): Grouping can distort the form of data, and different individuals might perform the task using different processes and parameters. These factors notwithstanding, we demonstrate conditions in which group analysis outperforms individual analysis. Such conditions include those in which there are relatively few trials per subject per condition, a situation that sometimes introduces distortions and biases when models are fit and parameters are estimated. We employed a simulation technique in which data were generated from each of two known models, each with parameter variation across simulated individuals. We examined how well the generating model and its competitor each fared in fitting (both sets of) the data, using both individual and group analysis. We examined the accuracy of *model selection* (the probability that the correct model would be selected by the analysis method). Trials per condition and individuals per experiment were varied systematically. Three pairs of cognitive models were compared: exponential versus power models of forgetting, generalized context versus prototype models of categorization, and the fuzzy logical model of perception versus the linear integration model of information integration. We show that there are situations in which small numbers of trials per condition cause group analysis to outperform individual analysis. Additional tables and figures may be downloaded from the Psychonomic Society Archive of Norms, Stimuli, and Data, www.psychonomic.org/archive.

To determine the processes underlying human performance, researchers often compare the relative fit of competing quantitative models when applied to experimental data. It is sometimes explicit and more often implicit that it is the processes of *individuals* that are of interest. Nonetheless, many researchers apply models not to data from individuals, but rather to data combined across multiple subjects. The hope is that combining the data will provide a clearer view of the underlying processes by reducing error and/or distortions. Such a group analysis is most common when individuals provide very few data per condition. Of course, a group analysis is based on the assumption that the processes used by the individuals are qualitatively similar and that combining across the individuals will not in itself distort the analysis. This assumption is never fully realized. The processes that individuals use may be similar but differ quantitatively, or they may be dissimilar, partitioning the individuals into groups. In the present research, we examine the trade-offs between some of the factors that might favor individual or group analysis when quantitative models of psychological processes are compared. The effort required to carry out both analyses is hardly greater than that required to carry out either alone, so our intent is not to recommend use of one approach instead of the other. We therefore explore the relative utility of the two approaches under different experimental conditions.

A number of researchers (e.g., R. B. Anderson & Tweney, 1997; Ashby, Maddox, & Lee, 1994; Estes, 1956; Estes & Maddox, 2005; Sidman, 1952; Siegler, 1987) have shown that various forms of bias due to data averaging can and do occur. Hayes (1953) gave a compelling example of such a distortion. Assume that a subject is given successive 30-sec opportunities to open a puzzle box. An individual's data will typically be well represented by a step function—successive failures followed by a run of successes. That is, for each individual, learning is all-or-none. But because subjects tend to learn on different trials, the average proportion correct for the group will rise gradually. An analysis of the group data would lead to the biased inference that learning is gradual rather than all-or-none. Just this sort of bias due to grouping was observed and then corrected through more sophisticated analysis, in studies in the 1950s and 1960s of simple learning tasks (e.g., Bower, 1961).

Although there are classes of models for which grouping across individuals does not cause distortions (see, e.g., Estes, 1956), the data from most influential contemporary models (models that tend to be complex, nonlinear, and interactive) change form when combined over individuals. For this and other reasons, the field has seen a trend toward modeling

A. L. Cohen, acohen@psych.umass.edu

individual data. In many cases, individual analysis is well justified—for example, when researchers collect a large number of data from each individual for each condition. Individual data analysis not only reduces distortions caused by the potentially large differences across individuals,[1] but also allows one to discover groups of individuals using different processes or different parameters of the same process.[2]

However, group analysis may be warranted in certain situations. For example, there are experimental situations in which it is difficult or impossible to collect much data from an individual for each condition. Such situations occur, for example, when infants or clinical populations are studied. They also occur when the design prevents multiple observations (e.g., studies of inattentional blindness sometimes obtain one observation per subject; Mack & Rock, 1998), when the number of possible stimuli is limited, or when the critical issue is performance on a particular preasymptotic trial during a long learning process. In such data-poor situations, inference based on individual data may well be so error prone that accepting the distortions produced by grouping is a worthwhile trade-off.

The article is organized as follows. We first briefly review a number of model selection techniques and discuss their appropriateness when there are few data points per individual. We then suggest a simple-to-use simulation procedure that allows us to determine which quantitative analysis choices are best for any particular number of trials per condition and individuals per experiment. Extension of the simulation procedure to model selection is discussed, and a number of model selection techniques are used to compare group and individual analyses in a specific example. To make the results as general as possible, this simulation procedure is then applied to pairs of models from three experimental settings: forgetting, categorization, and information integration. An enormous number of simulation results were obtained, producing information overload for any potential reader. We therefore have placed the complete results of the simulations in the Psychonomic Society Archive of Norms, Stimuli, and Data, giving only particularly useful, diagnostic, and representative results in the body of the article, along with a summary of the main findings.

MODEL SELECTION MEASURES

There are many criteria by which one model may be favored over another, including many that are not quantifiable.[3] Here, we will focus solely on numerical techniques of model comparison. But even with this restriction, comparing models is a complex affair. For example, selecting a model solely on the basis of its relative goodness of fit (the level of agreement between the model's predictions and the observed data) is often inadequate. One problem is that goodness of fit ignores the relative flexibility or complexity of the two models.

The complexity of a model increases with the number of different data sets it can describe. As a simple example, the cubic model ($y = a + bx + cx^2 + dx^3$) is more complex than the quadratic model ($y = a + bx + cx^2$), which is, in turn, more complex than the linear model ($y = a + bx$), on the basis of the number of curves each can fit. In fact, these models are nested: In each instance, the more complex model contains the simpler model as a special case (when d and/or c are set to zero). For nested models, the more complex model can always fit a data set at least as well as the simpler model. Furthermore, if error is added to data produced by a simpler nested model (e.g., the quadratic), the extra terms in the more complex model (e.g., the cubic) will allow the more complex model to fit better than even the simpler model that produced the data.

When used in isolation, the standard goodness-of-fit measures for comparing models, such as sum of squared error, percentage of variance accounted for, or maximum likelihood, do not deal adequately with model complexity. In recent years, a number of sophisticated methods have been developed that do take complexity into account. The normalized maximum likelihood (NML; Barron, Rissanen, & Yu, 1998; Grünwald, 2005; Rissanen, 2001), and Fisher information approximation (FIA; Rissanen, 1996) implementations of minimum description length, Bayesian model selection (BMS; e.g., the Bayes factor; Kass & Raftery, 1995), Bayesian nonparametric model selection (BNPMS; Karabatsos, 2006), various forms of cross-validation (CV; e.g., Berger & Pericchi, 1996; Browne, 2000), and generalization (Busemeyer & Wang, 2000) fall into this category. The difficulty in computing and/or applying the current state-of-the-art methods places these techniques out of the reach of many researchers, so simpler-to-use approximations (which sometimes make unrealistic assumptions about the models) are commonly used. These approximations include the Akaike information criterion (AIC; Akaike, 1973) and the Bayesian information criterion (BIC; Schwarz, 1978). We will consider both categories of methods, although the state-of-the-art methods will be explored only for one special case, since the main target of this article will be the simpler methods that are, or could be, in standard use. Our findings should thereby have considerable practical utility for a large number of practitioners.

In minimum description length, the data set can be considered a message that needs to be encoded for transmission to a receiver. Each model is a method used to encode the message. The message (data and model) with the shortest code is the more parsimonious model and is selected (Grünwald, 2000; Hastie, Tibshirani, & Friedman, 2001; Rissanen, 1978, 1987, 1996). It is not generally possible to determine the shortest code exactly, so it must be approximated. Two methods for approximating this code have achieved recent prominence. The older of these, which we will refer to as FIA (Rissanen, 1996), involves easier computations. Although still perhaps too technical for general use in the field, we will give detailed results for FIA for each of our simulated conditions.

The newer and better-justified method for approximating minimum code length is NML (Grünwald, 2007; Rissanen, 2001; described briefly later). NML requires too many computational resources for us to investigate systematically in all our simulations. A similar computational problem exists for another favored method for model selection known as BMS. BMS is based on the Bayes factor (Jeffreys, 1935, 1961), the ratio of posterior to prior odds. Other computationally intensive methods include various

forms of CV, which represent an empirical attempt to carry out model selection through assessment of predictive accuracy. In CV, model parameters are selected on the basis of a portion of the data (the training data), and prediction accuracy is assessed by employing these parameters to predict the remainder of the data (the testing data). If a model possesses unneeded complexity, it will account for noise in the training data and, therefore, perform poorly on the testing data. The generalization criterion follows a principle similar to that for CV, but instead of selecting some of the data across all conditions for training and testing on the remaining data, training takes place on all of the data from a subset of the conditions, with the remaining conditions used for testing. Our examination of NML, CV, and generalization will be restricted to one special case. (BMS is considered in Appendix A. BNPMS [Karabatsos, 2006] has been proposed too recently for us to assess in this article.)

The AIC and BIC methods are easy to implement and are in common use, so we will report the simulation results for these methods for each of our simulated conditions. The basis for AIC is the idea that the models under consideration are only approximations to the true generating model for a particular set of data (Burnham & Anderson, 2002). AIC selects the model that is closest to the generating model by using the Kullback–Leibler (K–L) distance (Cover & Thomas, 1991), which gives the amount of information lost when the generating model is approximated. It turns out that AIC values are based solely on the number of model parameters, and therefore, AIC fails to take into account such important factors as the correlations among the parameters. BIC is an approximation to the Bayes factor used in BMS and adjusts for both the number of parameters in a model and the number of observations in the data but, like AIC, fails to take into account the correlations among parameters.

Methods such as FIA, NML, BMS, BNPMS, CV, and generalization (and AIC and BIC under certain conditions) have excellent justifications. It is important to note, however, that all of these methods could run into difficulty when applied to data generated from small numbers of trials per condition, exactly the situation in which using group data may be best justified. AIC, for example, was derived as an approximation to the K–L distance between the generating and the candidate models, and it is valid only for large numbers of trials (Burnham & Anderson, 2002). BIC also is a large-sample approximation. For large numbers of trials, BIC approximates the Bayes factor (Kass & Raftery, 1995), but when there are only a few trials per condition, the approximation may not be very accurate. FIA is also a large-sample technique. The behavior of NML, BMS, CV, and generalization with small numbers of observations has not been studied in depth. We now will turn to a model selection procedure that may not suffer from such shortcomings.

THE SIMULATION PROCEDURE

The majority of the present research utilizes the parametric bootstrapping cross-fitting method (PBCM; Wagenmakers, Ratcliff, Gomez, & Iverson, 2004; see also Navarro, Pitt, & Myung, 2004) that is easy to use and allows us to examine model selection and mimicry in a way that applies when the number of data observations is small. Suppose a researcher is interested in inferring which of two models provides a better account of data and needs to decide whether or not to group individuals prior to analysis. To be concrete, suppose that the two models are exponential and power law descriptions of forgetting curves (a case considered in this article). On each trial of a forgetting experiment, a subject studies a list of words and then, after a varying time delay, is asked to recall as many words from the list as possible. This procedure repeats a number of times for each retention interval. The data are the average proportion of words recalled at each retention interval t, $P(\text{Recall}\,|\,t)$. The exponential model assumes that recall drops off according to

$$P(\text{Recall}\,|\,t) = ae^{-bt}, \tag{1}$$

and, for the power model, recall decreases according to

$$P(\text{Recall}\,|\,t) = at^{-b}. \tag{2}$$

For the PBCM procedure, first assume that the experimental data were generated by one of the models—say, the exponential model.[4] Second, an experiment is simulated assuming that the generating model (in this case, the exponential model) produces data for a particular number of individuals and a particular number of trials per condition. To produce variability across individuals, the parameters of the generating model are assumed to differ across individuals. Two methods are used to select parameters to generate data and to select variability in those parameters across individuals. These will be described in detail later, but for now we say only that the *informed* method selects parameters for each individual by sampling from actual fits of parameters to real-life data sets and that the *uninformed* method chooses a parameter value from a range of possible values for that model and then introduces variability across individuals by adding random Gaussian noise to that value. To produce variability within an individual, appropriate measurement error is added to the data generated with a particular set of parameters. This error, in general, will depend on both the form of the true model and the number of trials per condition, but in this article, the form of the error will always be binomial. For example, if the exponential model with a fixed set of parameters predicts that the probability of recall after 2.5 sec is .78 and there are five trials per condition, the number of simulated successful recalls for this condition is binomially distributed, with $p = .78$ and $n = 5$.

Third, both models being compared (in this case, the exponential and the power models) are fit to the data just produced. For each model, this is done in two ways: (1) by fitting the model parameters to best account for the data from each simulated individual in the experiment and (2) by fitting the model parameters to best account for the data combined over all of the simulated individuals in the experiment.[5] We find the parameters that maximize the likelihood of the data, given the parameters. The maximum likelihood provides an appropriate method for combining the fits to individual data into a summary measure across all individuals: The summary measure is the product of the individual likelihoods.[6] In an individual analy-

sis, the negative of the log maximum likelihood ($-\log L$) was summed across individuals, but for the group data, a single $-\log L$ fit is obtained to the data combined across individuals.

A fit value of zero means that the model perfectly accounts for the data. The higher the fit value is, the greater the discrepancy between the data and the model predictions. The difference between the two negative log maximum likelihoods for the two models indicates the difference in fit for the two models. For example, if the difference is $-\log L$ for the power model minus $-\log L$ for the exponential model, zero indicates that both models fit the data equally well; for this measure, the exponential and power models fit better when the values are positive and negative, respectively. Because it is difficult to keep track of all the compounded negatives and changes in directions of effects, we will simply indicate in our tables and figures which model is favored by a larger difference score.

To summarize the first three steps: One of the competing models was designated as the generating model; the model was used to simulate experimental data, given a fixed number of trials per condition per individual, a fixed number of individuals in the experiment, and a specific method for varying parameters across individuals; both competing models were then fit to both the individual and the grouped simulated data, using a $-\log L$ fit measure. Thus far, the products of the simulations were two difference-of-fit values, one for the individual data and one for the group data. For the same specific combination of factors, we repeated this procedure for 500 experiments. Each of the 500 simulations produced two difference scores that we could tabulate and graph.

Fourth, we repeated the entire process when the other model under consideration was used to generate the data. Thus, first we carried out the procedure above when one of the two models under examination was used to generate the data (the exponential in this example). Then the entire process was repeated when the other model was used to generate the data (in this example, the power model). To this point, there were now two sets of 1,000 data points (difference-of-fit values), 500 for each generating model. The first set was for the individual data, and the second set was for the group data. Fifth, this entire process was repeated for all the parametric variations we considered. All of the simulations were run as described above, with all combinations of 13 levels of individuals per experiment and 9 levels of trials per condition.

The factors above were also crossed with the two methods of selecting parameters to generate data—that is, two methods of selecting parameter variation across individuals. Both of these methods produced parameter ranges across individuals that were smaller than the range of uncertainty concerning the parameters themselves. The idea was that individuals obeying the same model should vary in their parameters, but not wildly around their joint mean. In addition, we did not wish to explore parameters that were very implausible for humans in a given task. With this in mind, we began by obtaining parameter estimates for individuals from the actual studies on which our simulations were based. (In the actual studies, the numbers of data per individual were large enough to make it plausible that these best-fitting parameter values provided sensible and reasonably stable estimates of parameters for that model used by humans in that task.) Let us suppose that a model has several component parameters that, together, make up its specification (e.g., a and b for a model $y = ax + b$). In the *uninformed* method, a central value was chosen for each component parameter by drawing from a uniform distribution. The range of values included those from past experimental work. Then parameters were chosen for individuals by adding Gaussian random noise with zero mean and a specified variance to each component central value (the variance was small, in comparison with the entire range).[7] The specifics of this sampling procedure will be discussed in Appendix B.

This method does not reflect typical parameter variations between individuals, because component parameter values for any individual are often correlated and not independent. Furthermore, if the level of parameter variation selected for the uninformed method is too low, the differences between individuals may be artificially small, producing little distortion in the data. Thus, we used a second method termed *informed*: The individual's parameter values were sampled with replacement from those parameter sets actually produced by fitting the model to real data for individuals (the real data were obtained in the study we used as a reference point for the simulations). We do not at present know how much the choice of methods for choosing individual differences will affect the pattern of results, but note that some robustness is suggested by the fact that for each of the applications studied in this article, the results of interest were essentially the same for the two different methods for choosing parameters.

In all, the variations of number of observations, number of individuals, and method of parameter selection and variation produced 117,000 simulations[8] that repeated Steps 1–4 above for a pair of compared models (e.g., the exponential and power laws of forgetting).

Sixth, this entire process was repeated for four pairs of models (each case based on some study chosen from the literature): exponential versus power models of forgetting, GCM (two versions) versus prototype models of categorization, and FLMP versus LIM models of information integration (see p. 702 below). In all, 468,000 simulated experiments were carried out.

USE OF THE SIMULATIONS FOR MODEL SELECTION AND MIMICRY

The previous section described how to generate the PBCM simulation results. This section illustrates how to use the simulation results for model selection. Recall that two sets of 1,000 data points were generated for a given set of experimental factors. Each set consisted of 500 simulations in which the first model (the exponential model) was used to generate the data and 500 simulations in which the second model (the power model) was used to generate the data. The first set was generated using average data, and the second set was generated using individual data.

For the purposes of understanding the results, it is perhaps most helpful to provide a graph giving the histogram

of these 500 difference-of-fit scores when one model generated the data and, on the same graph, the histogram of these 500 difference-of-fit scores when the other model generated the data. This histogram comparison graph is produced for a given method of data analysis (say, group), and then another such graph is produced for the other method of data analysis (individual). It is important to note that, at this point, the simulated results are described only in terms of differences in $-\log L$, a measure that does not account for model complexity.

We will illustrate the results and demonstrate how complexity adjustments are introduced by discussing just one simulated case: the comparison of exponential versus power laws of forgetting, with two observations per retention interval per individual, for 34 individuals, with parameter variation produced by sampling from actual data fits (i.e., the informed method). Note that the number of observations per condition is very low, the situation in which we have argued that group analysis might be warranted. The upper graph in Figure 1 shows the pair of histograms for the case in which data are analyzed by group. The lower graph shows the pair of histograms for the case in which the data are analyzed by individual. The horizontal axis gives the difference in $-\log L$ fit for the two models; the white bars give the results when the exponential model generated the data; and the gray bars give the results when the power model generated the data. The difference is arranged so that higher numbers represent better fits for the power model. That is, the gray bars, representing the case in which the true model was the power model, tend to lie further to the right, as they should if the power model generated the data. (Note that the individual and the group histograms have different axes.)

In many ways, the histograms graphed together on one plot invite familiar signal detection analyses (e.g., Macmillan & Creelman, 2005). In particular, to select one of the two models, the researcher needs to determine a decision criterion. Given data from an experiment, if the difference in fit values is above the decision criterion, the power model is selected; otherwise, the exponential model is selected. When the two distributions overlap, as in our example, the accuracy of correct model selection is limited by the degree of overlap (overlap may be thought of as a measure of the extent to which the two models mimic each other). As in signal detection analysis, one may analyze the results to produce a measure of sensitivity that will represent how distinguishable the distributions are from one another, regardless of the placement of the decision criterion. We used the overall probability of correct model selection as a measure and chose the decision criterion that maximized this probability (assuming that both models are equally likely a priori). For each of the histogram-of-difference graphs that we produced, we show this criterion, and label it *optimal*.[9]

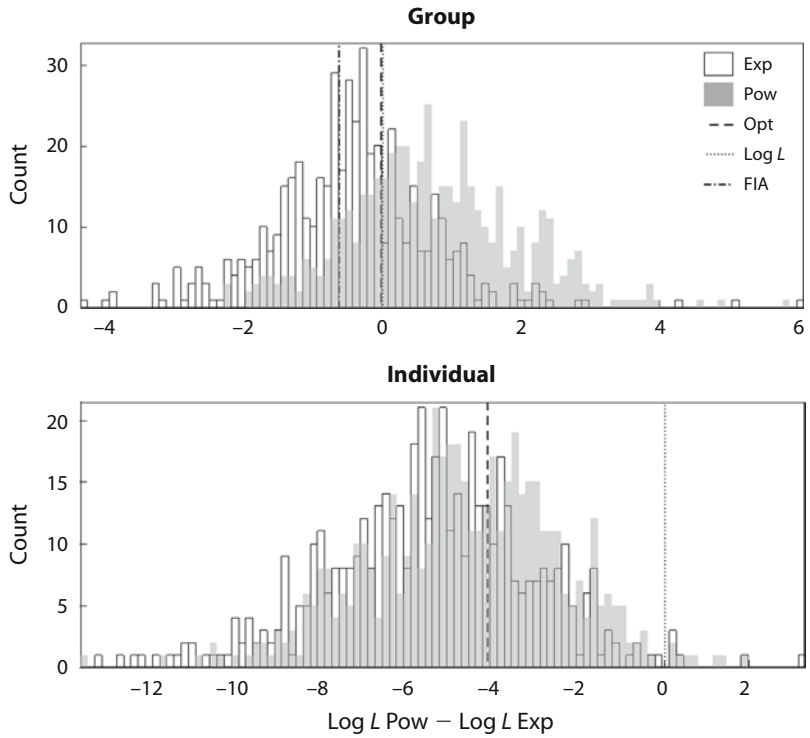

Figure 1. A sample histogram showing the difference of goodness of fit (log likelihood) for the exponential (Exp) and power (Pow) models of forgetting simulated with two trials per condition and 34 individuals per experiment, using informed parameters. The dotted, dashed, and mixed lines are the zero (log *L*), optimal (Opt), and Fisher information approximation (FIA) criteria, respectively. The FIA criterion for the individual analysis was −38.00 and does not show in the figure. The models are fit to the group and individual data in the top and bottom panels, respectively.

The upper graph in Figure 1 illustrates these points for the analysis of group data. Note that, in this case, the optimal decision criterion is close to zero difference. The group analysis histograms overlap quite a lot, limiting accuracy of model selection (perhaps not surprising with only two observations per individual per condition). The lower histograms illustrate the situation in which individual analyses are carried out. These histograms overlap even more than those for group analysis, showing that optimal criterion placement leads to superior model selection for the group analysis: Indeed, using the optimal decision criterion, the correct generating model was selected on 70% of the simulations when group analysis was used and on only 56% of the simulations when individual analysis was used. (Nor is this result exclusive to our method; we will show below that group analysis is also superior for AIC, BIC, FIA, NML, and other methods). These results show, for group and individual analyses, the best possible accuracy of model selection (defined as the probability of selecting the actual generating model) that could be achieved on the basis of the fitting of maximum likelihood estimates.

One might wonder why the group method is superior, especially given previous demonstrations that a mixture of exponentials with different parameters can produce data very similar to those produced by a power law (e.g., R. B. Anderson & Tweney, 1997; Myung, Kim, & Pitt, 2000): This result should bias group analysis to favor the power law. This bias was, in fact, observed in our simulations: Almost all of both individual analysis histograms (lower panel) are to the left of zero difference, showing that the exponential model fit both its own data and the data generated from the power model better. However, the group analysis histograms (top panel) have shifted considerably to the right, showing that group analysis produces a bias favoring the power law. Thus, the superiority of the group analysis occurs despite the bias caused by grouping: The group histograms may have shifted to the right, but they also have separated from each other, producing better performance.

We have identified at least two of the factors that make group analysis superior in the present situation. First is the bias to treat all of the data as coming from 1 subject when the data are noisy as a result of very few observations per condition. The individual analysis assumes (as a result of our combination rule) that the individuals are independent of each other. Our simulation produced subjects that tended to cluster in a region of the parameter space and, as a result, were somewhat dependent. With very noisy data, each individual estimate is poor, and the individual analysis does not take advantage of this dependence between subjects. The group analysis assumes that each subject is the same, so with very noisy data and some subject dependency, the bias of the group analysis is more useful than the flexibility of the individual analysis. Second, there is a separate bias favoring the exponential law when there are few data: A small number of all-or-none trials often results in zero correct scores for the longer delays. Because the exponential distribution has a lower tail than does the power law (at the extreme end of the parameter ranges used), the zero observations produce a better maximum likelihood fit for the exponential. Group analysis is less likely to show this bias than is individual analysis. Whatever biases exist for or against group analyses and for or against individual analyses, they are, of course, present in the methods used to produce the histogram plots, so that the PBCM method of choosing an optimal criterion to optimize model selection accuracy does not get rid of these biases but produces the best answer that can be achieved in the face of these biases.

It is also unclear how the between-subjects parameter variability will affect the results. For all the simulations in this article, the level of individual variability for the uninformed parameters was fixed. That is, after the mean parameter value was determined, the parameters from each individual were selected from a distribution with a fixed standard deviation. It is a logical step, however, to assume that the advantage of the group analysis will disappear when the dependence between subjects is reduced. In particular, the greater the standard deviation between subjects, the more the distribution over subjects will resemble a uniform distribution, resulting in independent subjects as assumed by the individual analysis. Pilot simulations have shown, however, that increasing or decreasing individual variability produces results that are qualitatively (and often quantitatively) very similar.

Model Selection by Fit Adjusted for Complexity—AIC, BIC, FIA, and NML

AIC, BIC, FIA, and NML select between models by comparing the log-likelihood fit difference with a criterion that adjusts for differences in model complexity. It turns out that the complexity adjustments of these model selection techniques produce a decision criterion that can be directly compared with the PBCM decision criterion. As we will describe shortly, in our example, when there are few data, AIC, BIC, and FIA do not produce a criterion adjustment that approximates the optimal criterion placement as determined by PBCM simulation. The distortion is quite severe, greatly lowering the potential accuracy of model selection and limiting any inference the investigator might draw concerning whether it is better to analyze data by individual or group.

Because AIC and BIC base their complexity adjustments on the number of model parameters, which are equal for the exponential and power models, the adjustment leaves the decision criterion at zero. This is true for both group and individual analyses. For the group histograms, the zero criterion is close to optimal. For the individual histograms, however, the criterion of zero lies almost entirely to the right of both histograms. The exponential model will almost always be selected, and the probability of correct model selection falls to chance.

As was noted earlier, there is bias favoring the exponential law when small amounts of data are analyzed. It might be thought that FIA could correct for this bias. For the group analysis, this criterion slightly misses the optimal criterion placement and only mildly lowers model selection performance to 65% correct model selection. For the individual analysis, however, the FIA complexity adjustment *overcorrects* to an enormous degree, so that both histograms lie far to the right of the criterion. (Indeed, the FIA decision criterion was −38.00 for the individual

analysis, far to the left of any of the simulated data, and so was omitted from the graph.) Now the power model would always be favored, again producing chance performance. Given that the rationale justifying FIA applies for large amounts of data, this failure is perhaps not a complete surprise.

Next, consider NML. NML calculates the maximum likelihood for the observed data, and scales this by the sum of the maximum likelihoods over all possible data sets that could have been observed in the experimental setting. Given that it considers all possible data sets of the size specified by the experimental design, NML is, in principle, applicable to small data sets. The results were simple: For both individual analysis and group analysis, the NML criterion was very close to optimal as determined by PBCM (71% and 58% correct model selection for group and individual analyses, respectively). Thus, for individual analysis, NML did much better than AIC, BIC, and FIA. Nonetheless, NML produced results that were the same as those for the optimal PBCM criterion placement: better performance for the group analysis.

Model Selection by Cross-Validation and Generalization

There are many forms of CV, all based on the principle that a model is fit to part of the data and then *validated* by using the results to predict the remaining data. A model is preferred if it predicts better. Each model was fit to half the data (for individuals, this meant that we fit one of the two observations per delay interval), and then the best-fitting parameters were used to predict the remaining data. For group analysis, a model was preferred if it fit the validation set better. For individual analysis, a model was preferred if the product of likelihoods of the individual validation sets was higher. CV selected the generating model on 69% of the simulations for group data but selected the generating model on only 50% of the simulations (chance) for the individual analysis.

The generalization criterion allows the researcher to choose which conditions should be used for training and which for testing the models under consideration. We chose to train the models on the first three retention intervals and to test using the last two retention intervals. Because the important theoretical difference between the exponential and the power law models lies in the longer retention intervals, the extreme retention intervals were used for testing. Like CV, the model that predicted the testing data better was selected. Generalization performed well for the group analysis (66%), but poorly for the individual analysis (52%). As was found using PBCM, model selection using CV and generalization was better for group data.

Validation of Choice

The comparison of model selection techniques above has equated performance with the accuracy of selecting the generating model. This validation criterion has the merits of being simple and easy to understand. Under the validation criterion of selecting the generating model, the PBCM solution can be considered optimal. However, there are reasons to worry that this validation criterion is not fully adequate, partly on the grounds that, according to some theories, its penalty for complexity may be insufficiently large.

In order to generalize our results, we explore another selection criterion, predictive validation (PV), an empirical method based on predictive accuracy. In this approach, instead of evaluating how well a method selects the generating model, we measure performance by how well a fitted model will predict new data from the same process that generated the original data. This method is a generalization of the method of comparing the recovered parameters with the generating parameters (e.g., Lee & Webb, 2005), but instead of looking at the parameters directly, we examine the probability distribution over all possible data sets that can be produced by a model and a specific set of parameters. Thus, we can compare the generating model and its generating parameters with any candidate model with its best-fitting parameters by examining the correspondence of the probability distributions produced by the two processes. Note that, according to this criterion, it is logically possible that, even if Model A generated the original data, Model B might still be selected because it provides superior predictions of new data generated by Model A. We will present an intuitive overview of the PV technique here; details will be given in Appendix C.

Say we are interested in comparing the exponential and power models of forgetting. To simplify the example, assume for now that the experiment involves only a single subject. (This restriction will be lifted below.) First, a set of parameters is selected for a generating model—say, the exponential model. The manner of parameter selection is identical to that of the informed approach discussed above. Note that, for any one experimental condition, a model with a fixed set of parameters defines a distribution over the possible data. Following the example from above, if the exponential model with a particular parameterization predicts a .78 probability of recall on each of five trials after a 2.5-sec retention interval, the predicted probability of 0, 1, 2, 3, 4, or 5 recalls after 2.5 sec is .00, .01, .06, .23, .40, and .29, respectively. Each experimental condition (in this case, each retention interval) has a comparable distribution. Just as in the PBCM, a simulated data set is produced from the generating model.

Second, the maximum likelihood parameters for these simulated data are found for one of the candidate models—say, the power model. The power model with maximum likelihood parameters will also generate a distribution for each condition. It is unlikely that the predictions of the generating and candidate models will be identical. For any one experimental condition, the differences between the generating distributions and the candidate distributions can be assessed using K–L divergence. K–L divergence produces a distance measure from a generating distribution to a candidate distribution; the smaller the distance, the more similar the candidate and generating distributions will be. To produce an overall measure for an individual, the K–L divergence is found for each condition and summed.

This procedure is then repeated with the exponential model as the candidate model. Because a smaller K–L divergence indicates a greater degree of overlap in the distributions that generate the data, the candidate model with

the smaller K–L divergence is the model that predicts new data from the generating model better.

Now consider the more realistic situation with multiple subjects that can be analyzed as individuals or as a group. For the individual analysis, there are separate generating parameters for each individual, so the distributions produced from a candidate model for each individual can be directly compared with that individual's generating model and parameters. Then the total K–L divergence is found by summing across individuals. For the group analysis, there is no set of best-fitting parameters for each subject, so in order to equate this situation with the individual analysis, each subject is assumed to use the best-fitting group parameters. Thus, each individual subject in the generating process is compared with the distribution defined by the best-fitting group parameters to produce a K–L divergence for each subject. The total K–L divergence is again summed over subjects. Using this procedure for the group analysis allows these K–L divergences to be directly compared with the K–L divergences produced by the individual analysis.

For a given generating model, the result of this methodology is four K–L divergences, one for each of the two candidate models crossed with the two types of analysis (group and individual). Because K–L divergence is a measure of the distance between the data produced by the candidate and the generating models, the combination of candidate model and analysis method that produces the smallest K–L divergence from the generating process is preferred. That is, the four K–L divergences are compared, and the smallest value wins. The entire analysis is repeated numerous times with different generating parameters. Then the analysis is repeated with the other model, the power model, as the generating model.

There is a disadvantage, however, to using the results of PV as a standard for determining whether to use group or individual analysis as a particular model selection method. Recall that the PBCM defines success as selection of the generating model. For any data set, individual or group, from a simulated experiment, a model selection method is accurate if it chooses the generating model. In contrast, PV defines success as selection of the model that predicts new data better, regardless of which model generated the data. This goal could create a situation in which it is not clear which model to prefer. For example, it might be that for data from a simulated experiment, the exponential model applied to group data is the best predictor of new data. Using this result as a standard, a model selection technique would be correct if it selects the exponential model. However, although the exponential model applied to group data may be the best overall predictor, the power model may be the better predictor when only individual data are considered. In this case, it is unclear whether it is appropriate to use the exponential model as the standard for analyzing individual data when the group and the individual PV results disagree.

Given these considerations, we decided to focus on two statistics. First, we ask directly whether PV favors group or individual analysis: How often does group PV analysis outperform individual PV analysis? Second, we use PV to at least partially support the generating model as a selection criterion. We ask, On what proportion of simulated experiments will the group or individual PV criterion match the generating model? If these analyses match each other reasonably well, it should increase our confidence in the generating model criterion that we adopted for the majority of simulations in this article. In this article, PV was computed for only a few interesting model comparisons (in particular, for 34 subjects and two trials per condition with the informed generating method).

Table 1
Predictive Validation Results for 34 Subjects and Two Trials per Condition

Generating Model	Comparison Model	Individual	Group
Exponential	Exponential	.000	.928
	Power	.000	.072
Power	Exponential	.000	.026
	Power	.000	.974
FLMP	FLMP	.436	.564
	LIM	.000	.000
LIM	FLMP	.000	.014
	LIM	.028	.958
GCM–γ	GCM–γ	.668	.008
	Prototype	.324	.000
Prototype	GCM–γ	.000	.000
	Prototype	1.000	.000

Note—FLMP, fuzzy logical model of perception; LIM, linear integration model; GCM, generalized context model.

The results from the PV method are presented in Tables 1 and 2. The tables list the results from each of the model pairs compared. For example, the first section of each table gives the exponential and power law comparison. The generating column indicates which model produced the data. For Table 1, each entry in the generating model column indexes the results for four methods of analysis: the outcome of crossing the two comparison models by individual or group analysis. The values are the proportions of simulated experiments for which a particular analysis method resulted in the best PV score. For example, when the exponential model generated the data, fitting the exponential model to group data outperformed the other three methods on 92.8% of the simulated experiments. For the exponential versus power law of forgetting with few data per individual, the PV results were clear: Regardless of the generating model, group analysis was favored.

Table 2 shows how often the PV result matched the generating model criterion. In Table 2, the same raw PV scores as those in Table 1 are used but the winning method is chosen differently. Instead of the best PV score being chosen from the four competing methods, in Table 2 the individual and group analyses are considered separately. For both the individual and the group analyses, the proportion of matches between the model with the better PV score and the generating model is reported. Table 2 shows that the group analysis matches well with the generating model criterion. Combined with the results of Table 1, showing that the group analysis produced better PV scores than did the individual analysis, the good match between PV and the generating model criterion provides extra support for our use of that criterion.

THE SIMULATIONS

The special case above was explored with a large variety of methods. In this section, we will explore a large number of experimental designs but will focus on just a few model selection methods: PBCM, AIC, and BIC (FIA analyses are included in the tables in the supplemental material, www.psychonomic.org/archive, but are not discussed). AIC and BIC are equivalent to maximum likelihood for all but one of the present model pairs. The overall results are summarized here; some technical details of the simulation procedures are presented in Appendix B; and all the simulation results, including the histograms, are presented in the supplemental material. The simulations were run with all combinations of 1, 2, 3, 4, 5, 6, 7, 8, 10, 14, 17, 20, and 34 individuals per experiment, crossed with 1, 2, 3, 5, 7, 10, 13, 17, and 25 trials per condition, crossed with the uninformed and informed methods of selecting parameter variation, crossed with the four model pairs: the exponential versus power laws, two versions of the generalized context model versus the prototype model, and the fuzzy logical model of perception versus the linear integration model. Because averaging was used to group data across individuals in all of the simulations, the terms *group data* and *average data* will be used interchangeably in the results and figures.

Exponential Versus Power Laws of Forgetting

The first two models compared are the exponential and the power models of forgetting discussed above and given in Equations 1 and 2. Recall that the experimental paradigm is based on a design given in Wixted and Ebbesen (1991). On each trial of the experiment, a subject studies a list of words and then, after a time delay of 2.5, 5, 10, 20, or 40 sec, is asked to recall as many words from the list as possible. This procedure is repeated for each of these five retention intervals. Each retention interval is a condition in this experimental paradigm, so the number of trials per condition is the number of lists studied at each delay. Details of the simulation procedure and parameter sampling procedure are given in Appendix B. The results of interest using informed and uninformed parameters were qualitatively very similar; therefore, only the results using uninformed parameters will be discussed.

Two major trends in the data are worth noting. First, except for low trials per condition and individuals per experiment, model mimicry using group data is relatively balanced. That is, the exponential and the power models are approximately matched in their ability to fit data from the competing model. Second, when using individual data, the exponential model is better able to mimic power model data than vice versa. Recall that the exponential model is better at producing extreme data points—that is, data close to 0% recall. It turns out that such data are quite likely when individuals are modeled, giving the exponential model its advantage. The effect, however, is ameliorated somewhat by increasing the number of trials per condition and quite a bit by increasing the number of individuals per experiment.

A summary of the results, using uninformed parameters, is given in Figure 2. The second and third rows of this figure show the proportion of model misclassifications for all combinations of individuals per simulated experiment and trials per condition for the averaged and individual data, respectively: the higher the point, the greater the number of misclassifications. The left column shows the results using the optimal criterion. As can be seen from the right column, misclassification patterns using the log likelihood criterion (zero) were virtually identical. These rows are color coded. The second and third rows are neutral gray at zero misclassifications. The graph for the average data changes to white as the number of misclassifications increases. The graph for the individual data changes to black as the number of misclassifications increases.

The two data sets, average and individual analyses, are directly compared in the first row of Figure 2. This figure gives the proportion incorrect for the averaged data minus the proportion incorrect for the individual data. Thus, to the extent that a point is negative, the average data do a better job of reducing model mimicry over the individual data. Positive points indicate that the individual data reduce model mimicry more than do the average data. Just as in the second and third rows, lighter and darker colors indicate more model mimicry when averaged and individual data, respectively, are used.

All of the points in the first row of this figure are nonpositive, indicating that fitting averaged data always reduces model mimicry. For low numbers of individuals per experiment, the group and individual results are very similar and both models perform poorly. Model selection using both group and individual data improves as the number of trials is increased. The improvement in model selection when the number of individuals is increased improves far more for the group data. With enough trials per condition, the difference between using average and individual data may eventually disappear, but for the ranges explored here, using average data is preferred. These results hold whether the zero (log likelihood) or optimal criterion is used.

In summary, when these models are compared using this experimental design, it is better to use averaged data.

Models of Information Integration: Fuzzy Logical Model of Perception and Linear Integration Model

The main goal of models of information integration is to explain how a decision is formed from multiple sources of information. The stimuli in a typical information integration experiment fall into one of two classes. The first

Table 2
Proportions of Agreement Between Generating Model and Best Predictive Validation Model

	Agreement	
Generating Model	Individual	Group
Exponential	.164	.928
Power	.876	.974
FLMP	.984	1.000
LIM	1.000	.984
GCM–γ	.672	.978
Prototype	1.000	.986

Note—FLMP, fuzzy logical model of perception; LIM, linear integration model; GCM, generalized context model.

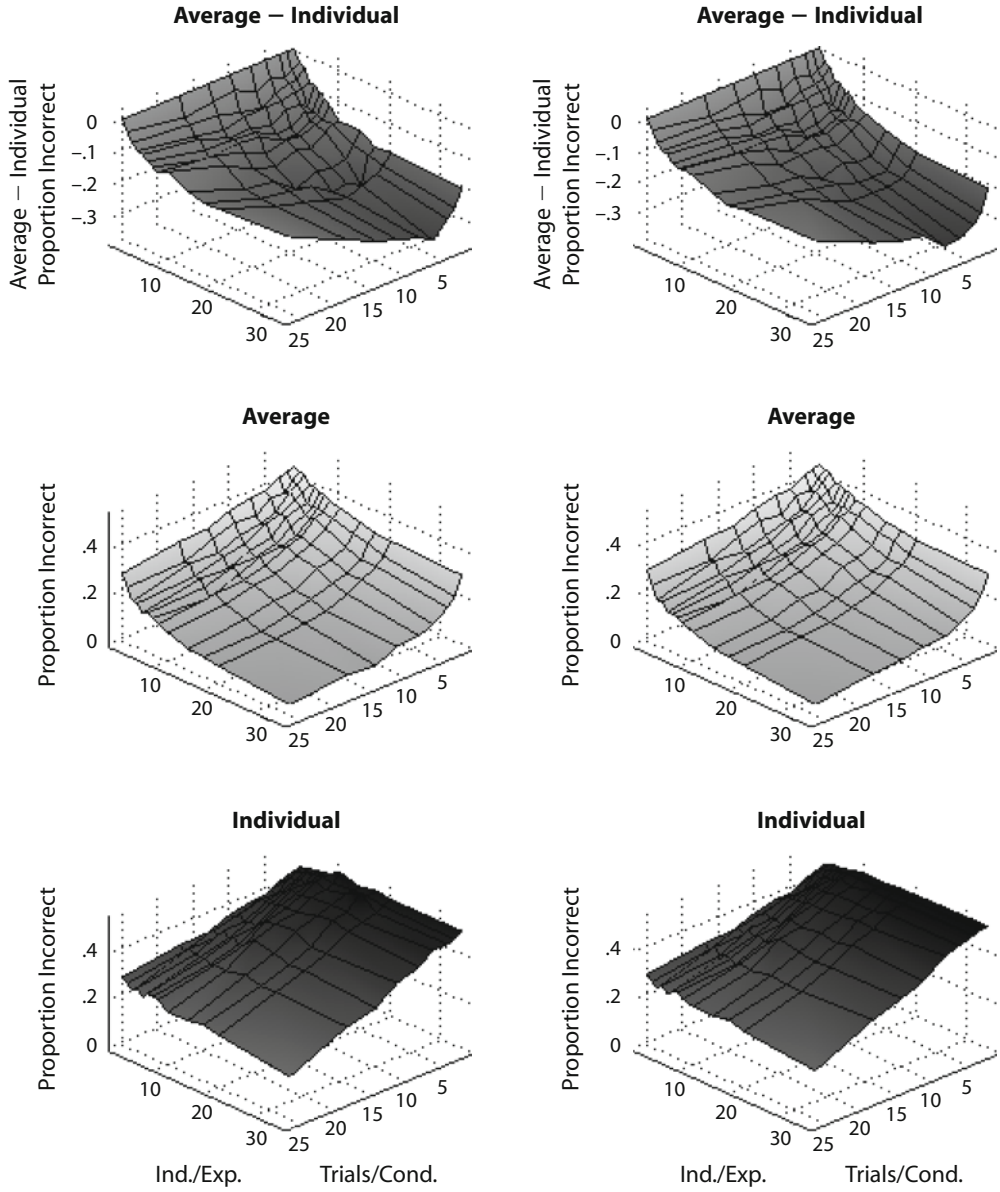

Figure 2. The bottom two rows give the proportions of simulated experiments for which the incorrect model was selected as a function of trials per condition and individuals per experiment when comparing the exponential and power models of forgetting fit to average (middle row) and individual (bottom row) data and using the optimal (left column) and zero (right column) criteria and the uninformed parameter selection method. The top row is the difference of the two bottom rows.

class of stimuli is constructed by factorially combining the different levels of two or more stimulus dimensions (e.g., visual and auditory presentation of a person speaking one of a range of phonemes). The second class consists of each level of each stimulus dimension in isolation (e.g., an auditory presentation of a phoneme without the visual presentation). On each trial, the individual is asked to assign the stimulus to a response class. The data are the proportions of trials for which each stimulus was assigned to a response class.

The present work is based on an experimental design from Massaro (1998). There are two stimulus dimensions: a five-step continuum between the phonemes /ba/ and /da/ for visual (i.e., lip and jaw movement) and auditory speech. As was described above, the 5 stimuli from each of the two dimensions were presented both in isolation (5 auditory and 5 visual stimuli) and factorially combined (25 audiovisual stimuli). On each trial, the subject was asked to categorize the stimulus as /ba/ or /da/. The data are the proportion of times the individual responded /da/, p(/da/). A condition is one of the 35 stimuli.

The two models of information integration under consideration are the fuzzy logical model of perception (FLMP; Oden & Massaro, 1978) and the linear integration model (LIM; N. H. Anderson, 1981). Let A_i and V_j be the ith level of the auditory dimension and the jth level of the

visual dimension, respectively. The psychological evaluation of A_i and V_j, a_i and v_j, respectively, are assumed to lie between 0, a definite mismatch, and 1, a definite match. Whereas the FLMP assumes that the probability of a /da/ response is given by a Bayesian combination of a_i and v_j,

$$p\left(/\text{da}/\,|\,A_i,V_j\right) = \frac{a_i v_j}{a_i v_j + \left(1-a_i\right)\left(1-v_j\right)}, \tag{3}$$

the LIM assumes that these two sources of information are averaged,

$$p\left(/\text{da}/\,|\,A_i,V_j\right) = \frac{a_i + v_j}{2}. \tag{4}$$

Details of the uninformed and informed parameter sampling methods are given in Appendix B. Again, because the data of interest when uninformed and informed parameters were used were very similar, only the results using the uninformed parameters are discussed.

For individual data, in general, model mimicry is low; both models favor their own data, and the optimal criterion is well approximated by a zero criterion. The exception to this pattern is at very low trials per condition, where much of the LIM data are better accounted for by the FLMP. The FLMP advantage in this situation is due to its greater flexibility in fitting extreme data. For the FLMP, setting either parameter to one or zero produces a response probability of one or zero, but for the LIM model, both parameters need to be set to extreme values to produce extreme response probabilities. For all-or-none data with few trials per condition, the FLMP is much better at fitting the resulting data. When average data are used, the advantage of the FLMP with low trials all but disappears.

The proportion incorrect summary graphs for uninformed parameters are displayed in Figure 3. The left and right columns of Figure 3 represent the proportion of incorrect model selections when the optimal criterion and the zero (straightforward log likelihood), respectively, are used. The model selection errors generated by the greater complexity of the FLMP when low trials per condition are used are reflected in the zero-criterion, individual data graph. If a zero criterion is used, model mimicry is greatly reduced at low trials per condition by using average data. Utilizing the optimal cutoff greatly reduces the errors associated with using individual data at low trials per condition and, indeed, produces a slight advantage for using individual data at very low trials per condition and individuals per experiment (where model selection performance is particularly bad). For moderate to high trials per condition and individuals per experiment, however, model mimicry is similar (and very low) when either individual or average data are used.

PV was used only for the case with two observations per individual and 34 subjects. In Table 1, the group analysis showed an advantage over the individual analysis in predicting new data from the same set of simulated subjects, giving more evidence that group analysis is useful when there are few trials per condition. In addition, there was nearly perfect agreement between the models chosen by PV for both group and individual analysis and the generating models (see Table 2). This agreement lends more weight to using recovery of the generating model as a measure of model selection accuracy.

In summary, the use of both average and individual data produces good model selection results, with the following exceptions. When the zero criterion is used, average data should be used for low trials per condition. When the optimal criterion is used, individual data should be used for low trials per condition and individuals per experiment.

Models of Categorization: Generalized Context Model and Prototype Model

In a typical categorization task, an individual is asked to classify a series of stimuli that vary along multiple feature dimensions. The individual receives corrective feedback after each trial. Two categorization models are considered here: the generalized context model (GCM; Nosofsky, 1986) and the prototype model (Reed, 1972). Both models assume that the stimuli are represented by points in a psychological M-dimensional space. Let x_{im} be the psychological value of stimulus i on dimension m. Then, the psychological distance between stimulus i and j is

$$d_{ij} = \sum_{m\in M} w_m \left(x_{im} - x_{jm}\right)^2, \tag{5}$$

where the w_ms are parameters representing the attention weight given to dimension m. Each $w_m \geq 0$ and the w_ms sum to 1. The similarity between stimulus i and j is an exponentially decreasing function of distance in the space (Shepard, 1987),

$$s_{ij} = \exp\left(-c \cdot d_{ij}\right), \tag{6}$$

where c is an overall sensitivity parameter. In a two-category experiment, the GCM assumes that the probability that stimulus i is classified into Category A is given by the summed similarity of stimulus i to all stored exemplars of Category A divided by the summed similarity of stimulus i to all stored exemplars from both Categories A and B,

$$P(\text{A}\,|\,i) = \frac{\sum_{a\in\text{A}} s_{ia}}{\sum_{a\in\text{A}} s_{ia} + \sum_{b\in\text{B}} s_{ib}}. \tag{7}$$

The prototype model assumes that category decisions are based on the relative similarity of the test stimulus to the single central category prototype from each category. Let P_A and P_B be the prototypes (usually the mean category member) for Categories A and B, respectively. Then,

$$P(\text{A}\,|\,i) = \frac{s_{iP_\text{A}}}{s_{iP_\text{A}} + s_{iP_\text{B}}}. \tag{8}$$

To account for the finding that subjects tended to respond more deterministically than is predicted by the GCM (Ashby & Gott, 1988), Ashby and Maddox (1993) proposed a version of the GCM that allows the model to predict more or less response determinism than is predicted by the base-

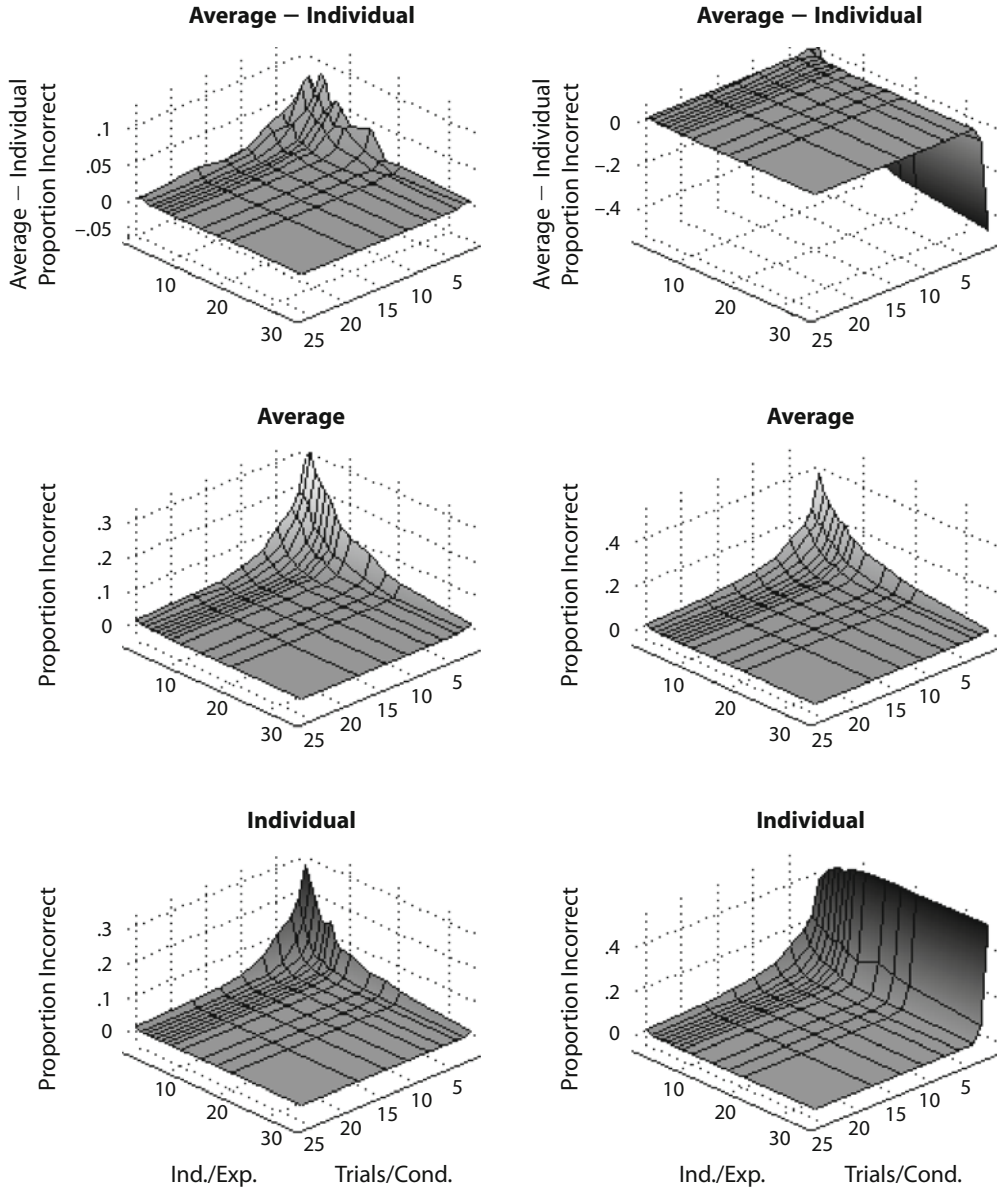

Figure 3. The bottom two rows give the proportions of simulated experiments for which the incorrect model was selected as a function of trials per condition and individuals per experiment when comparing the fuzzy logical model of perception and the linear integration model of information integration fit to average (middle row) and individual (bottom row) data and using the optimal (left column) and zero (right column) criteria and the uninformed parameter selection method. The top row is the difference of the two bottom rows.

line version given in Equation 7. This version of the GCM adds a *response-scaling* parameter, γ, to Equation 7,

$$P(\mathrm{A} \mid i) = \frac{\left(\sum_{a \in \mathrm{A}} s_{ia}\right)^{\gamma}}{\left(\sum_{a \in \mathrm{A}} s_{ia}\right)^{\gamma} + \left(\sum_{b \in \mathrm{B}} s_{ib}\right)^{\gamma}}. \qquad (9)$$

As γ increases, responses become more deterministic; that is, the more probable response category is more likely to be selected. When $\gamma = 0$, all categories are selected with equal probability. The GCM without and with the response-scaling parameter will be referred to as GCM–R (GCM–restricted) and GCM–γ, respectively. The same response-scaling parameter can be added to the prototype model. However, Nosofsky and Zaki (2002) showed that the addition of a response-scaling parameter to the prototype model formally trades off with the sensitivity parameter and, so, is redundant. The category members, classifications, and prototypes used as stimuli in the simulated categorization task are taken from the 5/4 category structure of Medin and Schaffer (1978, their Figure 4).

GCM–γ versus prototype. The analyses will start by comparing the prototype model with the GCM–γ. The

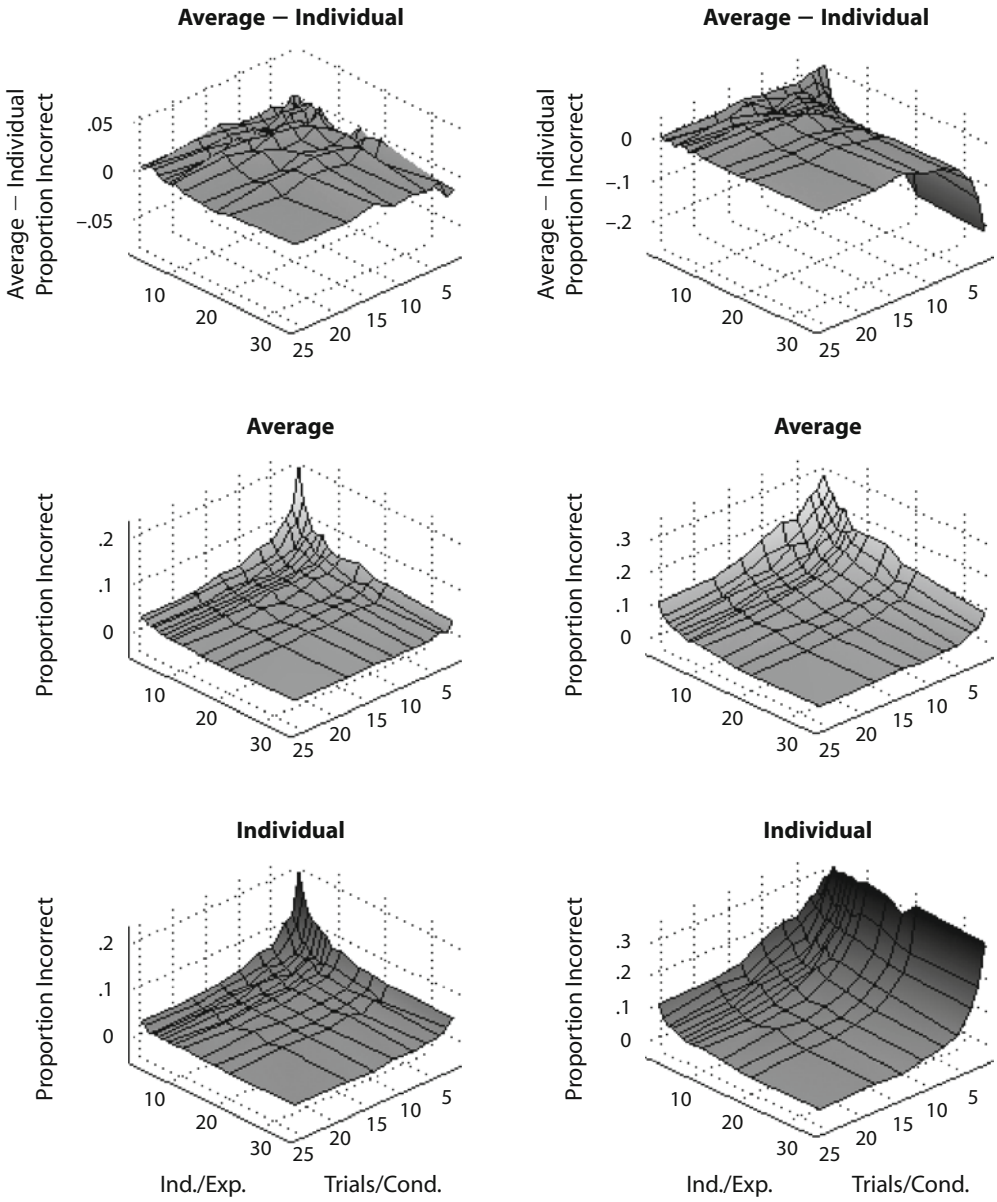

Figure 4. The bottom two rows give the proportions of simulated experiments for which the incorrect model was selected as a function of trials per condition and individuals per experiment when comparing the generalized context model–γ and prototype model of categorization fit to average (middle row) and individual (bottom row) data and using the optimal (left column) and zero (right column) criteria and the uninformed parameter method. The top row is the difference of the two bottom rows.

parameter-sampling scheme is detailed in Appendix B. Again, although the distributions for the data generated using the uninformed and informed parameter selection methods have different shapes, the results were otherwise very similar. The discussion will focus on the results generated from uninformed parameters.

The GCM–γ is far more complex than the prototype model when fitting individual data for low trials per conditions and individuals per experiment. This advantage decreases as the number of trials increases. For group data, the GCM–γ is again more complex than the prototype model for low numbers of trials and individuals. The relative complexity of the two models is reduced with increases in both trials and individuals. Indeed, with 25 trials and 34 individuals, the two models rarely mimicked each other.

A summary of the selection errors for the zero and optimal criteria is given in Figure 4 for both the optimal (left column) and the zero (right column) criteria when uninformed parameters are used. When the zero criterion is used, it is typically safer to analyze average data, especially when the number of trials per condition is low. For moderate to high trials per condition, both averaged and individual data give approximately the same result. When the optimal criterion is used, both averaged and individual

data essentially produce the same results for all levels of the factors explored here.

Both AIC and BIC penalize the GCM–γ for its extra parameter (BIC's penalty depending on the number of observations), implemented as a shift of the zero criterion to another specified position. For individual analysis, AIC does no better than the zero cutoff, and BIC does even worse than the zero cutoff for very small numbers of trials per condition. We do not recommend the use of AIC and BIC with individual analysis.

PV was used for the case of 34 subjects and two trials per condition. Tables 1 and 2 show an interesting result for this particular situation. In Table 1, the individual analysis was more effective than the group analysis at predicting new data from the same simulated subjects. However, in Table 2, the group PV analysis picked the generating model more often than did the individual PV analysis. Combined, these results mean that the group analysis produced worse PV scores than did the individual analysis, yet was able to determine the generating model better. Because the group and individual analyses almost always agreed on the prototype model when it generated the data, this strange result is due to the GCM–γ-generated data. Of the 500 simulations in which the GCM–γ generated the data, there were 159 cases in which the group analysis picked the GCM–γ model and the individual analysis picked the prototype model. For 157 of these 159 cases, the prototype model had a better PV score than did the GCM–γ model. This result is likely due to the enhanced flexibility of the GCM–γ model with small amounts of data, as compared with the prototype model. With few trials, the GCM–γ can fit the extreme values produced by all-or-none trials better than the prototype model can. However, new data generated from the same process are unlikely to have the same extreme values, so the prototype model (although not the generating model) is a better predictor of new data.

In summary, regardless of the parameter-sampling scheme or criterion choice, similar results are given when average and individual data are used, except when the number of trials per condition is low. In this case, it is safer to use averaged data.

GCM–R versus prototype. An additional set of simulations were run with the response scaling parameter, γ, of the GCM–γ fixed at one, so that the response probability always matched the probability of Equation 7. We will refer to this restricted model as the GCM–R. Data were generated from the prototype model as in the previous simulation. Details of the parameter-sampling procedure for the GCM–R are given in Appendix B.

As in the GCM–γ and prototype model simulations, the results using the zero and optimal criteria do differ, but because the differences here do not have a large impact on how often the correct generating model is selected, the analysis will focus on the optimal criterion. There are, however, some important differences between the results when uninformed and informed parameter selection methods are used. When uninformed parameters are used, model mimicry is relatively low for both individual and average data, except for low trials per condition and individuals per experiment, where the GCM–R was slightly more complex. Similar results were found when informed parameters were used, but, interestingly, the prototype model was the more general model.

The probability of selecting the incorrect model for the uninformed (left column) and informed (right column) parameter selection methods is shown in Figure 5. For moderate to high trials per condition, model selection was excellent regardless of parameter selection method or the use of average and individual data. The average analysis outperformed the individual analysis for low trials per condition when the uninformed parameters were used, but a reversal was found when informed parameters were used. This result can be taken as a demonstration that one cannot always expect that a reduction of data will produce an increasing benefit for group analysis.

CONCLUSIONS

The results of our simulations illustrate some of the complexities of model selection and model mimicry. They afford a few reliable conclusions and some suggestions for action and provide some pointers toward further research. It must be kept in mind that our main concern was the great body of researchers who need to draw conclusions on the basis of fairly simple and easy-to-apply techniques, which we implemented as maximum likelihood parameter estimation for each model. Our primary goal was investigation of the relative merits of analysis by individuals or analysis by group (data combined across individuals).

First, and most clearly, we have seen that neither analysis by individual nor analysis by group can be recommended as a universal practice. The field began in the 1800s with experiments utilizing just one or only a few subjects and, therefore, using individual analysis. As studies began using larger groups of subjects, typically with fewer data per subject, group analysis became common. More recently, practice has moved toward individual analysis, followed by combination of the results across individuals. We have seen, most clearly in the case of the power/exponential model comparison, that the last practice is not always justified. Furthermore, we have seen cases in which group analysis is superior, whether one's goal is selecting the better model or obtaining accurate parameter estimates (although our main focus was on the former).

Our simulations show a tendency for the relative advantage of group analysis to increase as the number of data per subject drops. There are, of course, many potential benefits of individual data analysis that we have not explored in our simulations, such as the possibility of partitioning individuals into groups with similar parameters or into groups obeying different models. We note that such refinements of individual analysis will likely be ineffective in the situations that we have found in which group analysis is superior: experiments with few data per individual. On the other hand, one cannot unconditionally recommend group analysis, because such analysis is subject to well-known distortions in many of the settings in which individuals operate with different parameters. These factors often operate in opposition. Our findings demonstrate the perhaps surprising result that when there are very few data per individual, individual analysis is subject to noise and bias that sometimes produce distortions

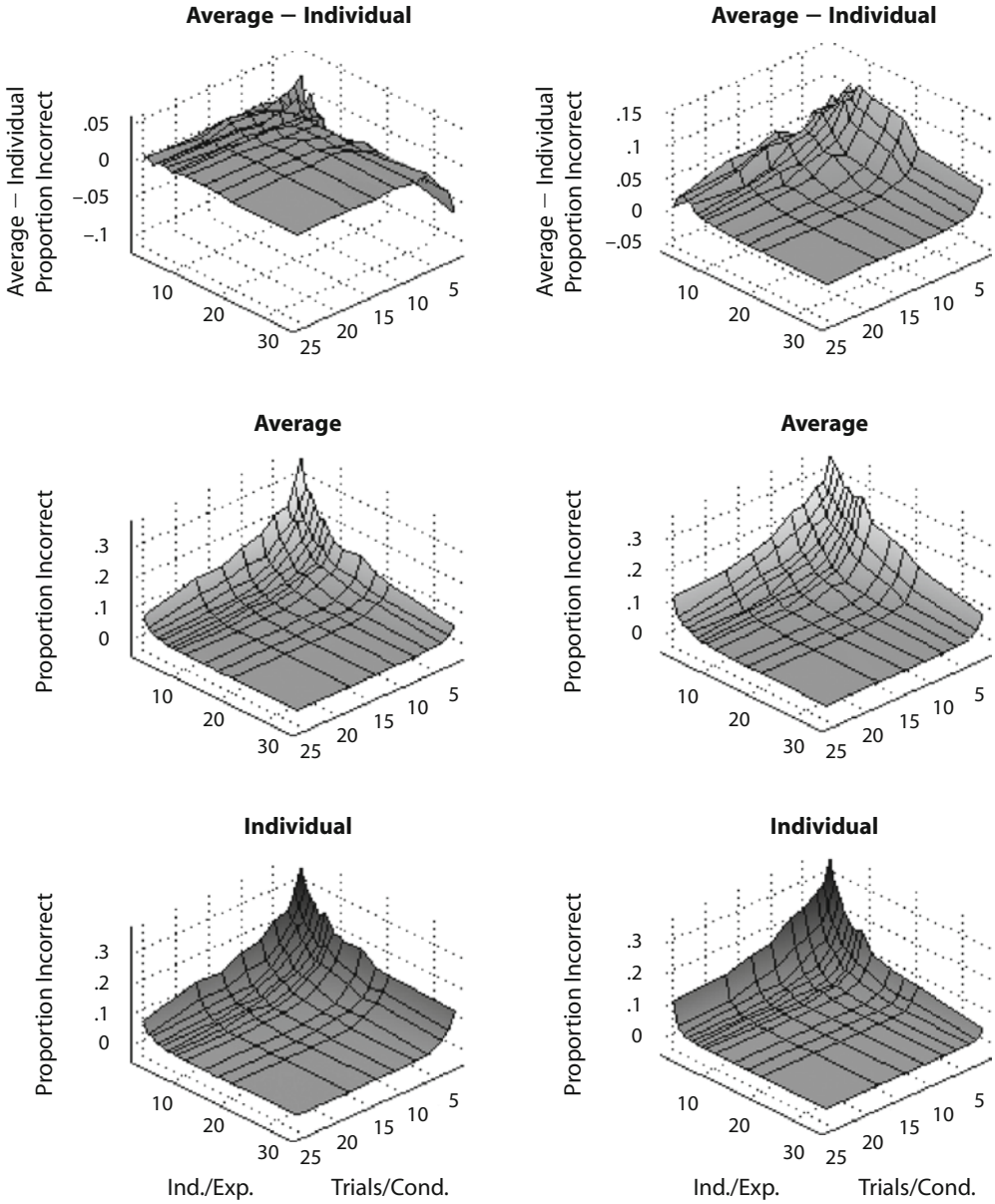

Figure 5. The bottom two rows give the proportions of simulated experiments for which the incorrect model was selected as a function of trials per condition and individuals per experiment when comparing the generalized context model–restricted and prototype model of categorization fit to average (middle row) and individual (bottom row) data and using the uninformed (left column) and informed (right column) parameter selection methods and the optimal criterion. The top row is the difference of the two bottom rows.

even more serious than those produced by grouping. In such cases, group analysis is the least bad strategy.

These findings do not produce a clear recommendation for practice, because there does not seem to exist a generally optimal approach, at least for the cases we analyzed in detail in which an investigator is limited to the use of maximum likelihood parameter estimation. Superficially, it appears tempting to recommend more advanced methods of model selection, such as FIA, NML, BMS, and so forth. Our results, however, lead to a cautionary note: The use of a single advanced method could produce poor results, because we have seen different answers produced by different methods. In contrast with a number of the theoretically based methods for balancing simplicity and fit, PBCM provides a good deal of extra information, giving the distributions of expected outcomes when each model is in fact "true." The PBCM simulation method has the advantages of relatively easy implementation, ease of comparison of individual and group analysis for small data sets, and a measure of how well an analysis will work for a particular experimental design but requires the distribution of subject parameters to be specified. In addition, PBCM requires one to accept the goal of selection of the actual generating model. Experts in the field of model selection will often prefer other goals (ones under-

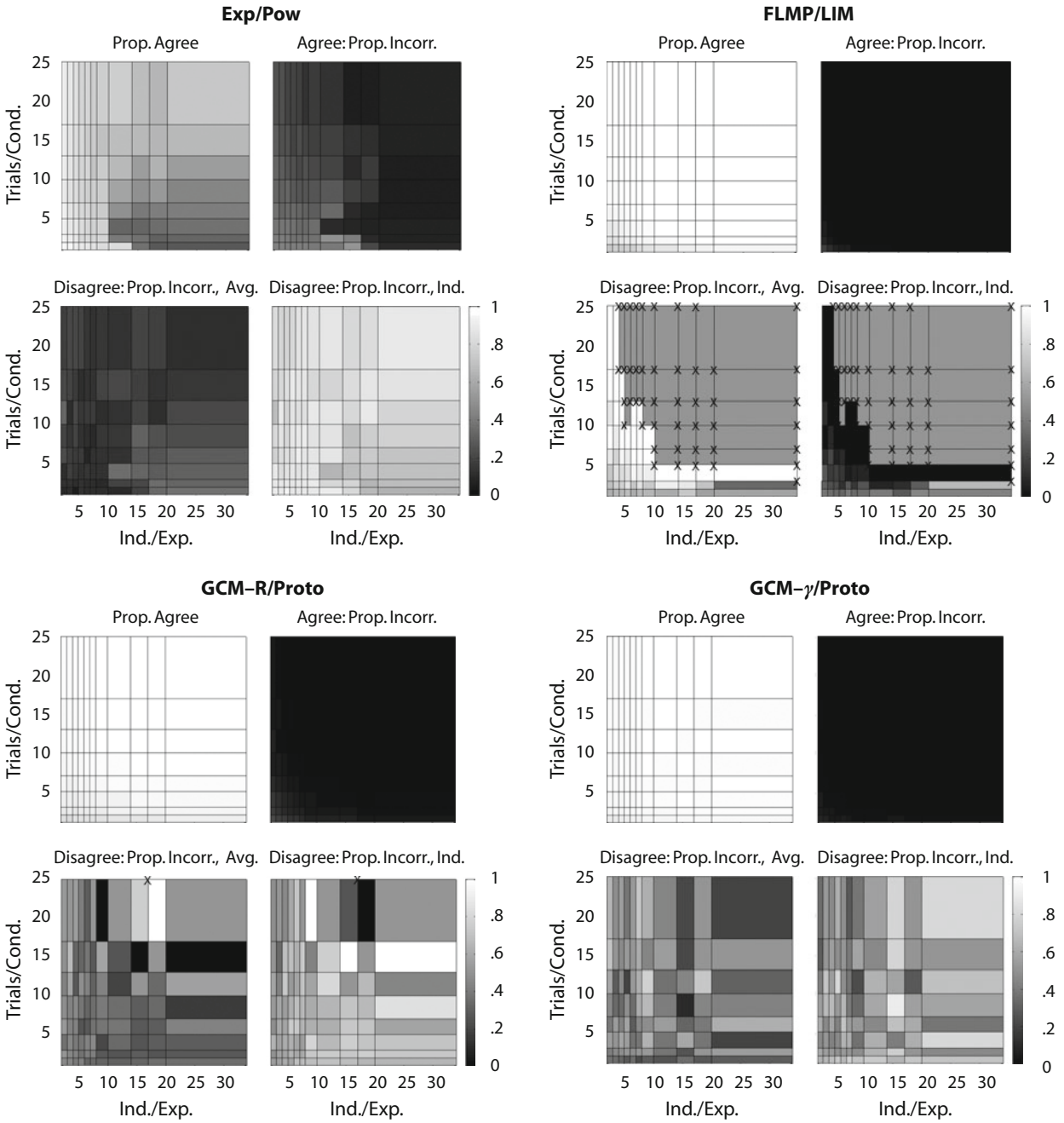

Figure 6. Proportions of simulations for which average and individual analyses agree, proportions of incorrect model selections when average and individual analyses agree, and proportions of incorrect model selections using average and individual analyses, respectively, when average and individual model analyses do not agree for the four pairs of models and levels of trials per condition and individuals per experiment discussed in the text. Xs indicate conditions for which average and individual analyses always agreed. Exp, exponential; Pow, power; FLMP, fuzzy logical model of perception; LIM, linear integration model; GCM, generalized context model; Proto, prototype.

lying such techniques as NML and BMS). In our opinion, there are multiple goals of model selection that cannot all be satisfied simultaneously, even when focus is restricted primarily to the goal of balancing simplicity and fit. Thus the "best" goal for model selection remains an elusive target, a matter for debate, and a subject for ongoing research.

There is, however, an easy-to-use heuristic that can increase a researcher's confidence in their model selection results. Because the computer program capable of analyzing the combined group data will be capable of analyzing individual data without extra programming, and vice versa, both methods should be used for each data set. When both methods lead to the same conclusion, it is a reasonable decision strategy to accept this conclusion. There is, of course, no guarantee that the conclusion so reached will be correct. In fact, the accuracy of such joint

acceptance decisions could well be significantly lower than the decision reached with the better method by itself. This was seen, for example, in the case of the power law versus exponential, with few data per individual, where individual analysis was essentially at chance. Applying a joint agreement heuristic in such a situation is equivalent to randomly selecting one half of the group analysis cases. It is clear that such a procedure can only reduce the efficacy of group analysis. Nonetheless, the joint acceptance heuristic may be the best strategy available in practice.

We tested this joint acceptance heuristic on the simulated data discussed previously with uninformed parameter selection. The results are given in Figure 6 for the four model comparisons. The *x*- and *y*-axes of each panel are the different levels of individuals per experiment and trials per condition, respectively, discussed above. Simulations with one individual per experiment are not included in the figures, because average and individual analyses always agree. For each model comparison, the upper left panel gives the proportion of simulations for which the average and individual analyses agreed. The upper right panel indicates the proportion of simulations for which the incorrect model was selected when average and individual analyses agreed. The lower left and lower right panels give the proportion of simulations in which average and individual data, respectively, were used for which the incorrect model was selected when the average and individual analyses disagreed. Light and dark shadings indicate high and low proportions, respectively.

First, note that, in general, agreement between group and individual analyses was high. The only exception was for the exponential and power simulations, especially for a large number of subjects with few trials. Second, when average and individual analyses agreed, model selection was nearly perfect for most model pairs. The accuracy for the exponential and power simulations was less than that for the other model pairs but was still good. Third, the simulations do not suggest a clear choice for which analyses to favor when average and individual analyses disagree. For the exponential and power simulations, average analysis clearly performs better, and interestingly, for few subjects and trials, group analysis is even better when it disagrees with individual analysis. For the FLMP and LIM, individual analysis is preferred when there is disagreement. There is no obvious recommendation for the GCM and prototype simulations. In our simulations, the joint acceptance heuristic tends to outperform even the best single method of analysis.

The field of model selection and data analysis methodology continues to evolve, of course, and although there may be no single best approach, the methods continue to improve. Researchers have, for example, already begun to explore hierarchical analyses. Hierarchical analyses hold the potential of combining the best elements of both individual and group methods (see, e.g., Navarro, Griffiths, Steyvers, & Lee, 2006). The technical methods of model selection also continue to evolve in other ways. We can, for example, expect to see methods combining elements of BMS and NML (e.g., Grünwald, 2007). We end by noting that although advanced research in model selection techniques tends not to focus on the issues of accessibility and ease of use, it is not unreasonable to expect the eventual production of packaged programs placing the best state-of-the-art methods within reach of the average researcher.

AUTHOR NOTE

All the authors contributed equally to this article. A.N.S. was supported by a National Defense Science and Engineering Graduate Fellowship and an NSF Graduate Research Fellowship. R.M.S. was supported by NIMH Grants 1 R01 MH12717 and 1 R01 MH63993. The authors thank Woojae Kim, Jay Myung, Michael Ross, Eric-Jan Wagenmakers, Trisha Van Zandt, Michael Lee, and Daniel Navarro for their help in the preparation of the manuscript and Dominic Massaro, John Paul Minda, David Rubin, and John Wixted for the use of their data. Correspondence concerning this article should be addressed to A. L. Cohen, Department of Psychology, University of Massachusetts, Amherst, MA 01003 (e-mail: acohen@psych.umass.edu).

REFERENCES

AKAIKE, H. (1973). Information theory as an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second International Symposium on Information Theory* (pp. 267-281). Budapest: Akademiai Kiado.

ANDERSON, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.

ANDERSON, R. B., & TWENEY, R. D. (1997). Artifactual power curves in forgetting. *Memory & Cognition*, **25**, 724-730.

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. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, **37**, 372-400.

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.

BARRON, A. R., RISSANEN, J., & YU, B. (1998). The MDL principle in modeling and coding. *IEEE Transactions on Information Theory*, **44**, 2743-2760.

BERGER, J. O., & PERICCHI, L. R. (1996). The intrinsic Bayes factor for model selection and prediction. *Journal of the American Statistical Association*, **91**, 109-122.

BOWER, G. H. (1961). Application of a model to paired-associate learning. *Psychometrika*, **26**, 255-280.

BROWNE, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, **44**, 108-132.

BURNHAM, K. P., & ANDERSON, D. R. (1998). *Model selection and inference: A practical information-theoretic approach*. New York: Springer.

BURNHAM, K. P., & ANDERSON, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York: Springer.

BUSEMEYER, J. R., & WANG, Y.-M. (2000). Model comparisons and model selections based on generalization criterion methodology. *Journal of Mathematical Psychology*, **44**, 171-189.

COVER, T. M., & THOMAS, J. A. (1991). *Elements of information theory*. New York: Wiley.

ESTES, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin*, **53**, 134-140.

ESTES, W. K., & MADDOX, W. T. (2005). Risks of drawing inferences about cognitive processes from model fits to individual versus average performance. *Psychonomic Bulletin & Review*, **12**, 403-408.

GRÜNWALD, P. [D.] (2000). Model selection based on minimum description length. *Journal of Mathematical Psychology*, **44**, 133-152.

GRÜNWALD, P. D. (2005). Minimum description length tutorial. In P. D. Grünwald, I. J. Myung, & M. A. Pitt (Eds.), *Advances in minimum description length: Theory and applications* (pp. 23-80). Cambridge, MA: MIT Press.

GRÜNWALD, P. D. (2007). *The minimum description length principle*. Cambridge, MA: MIT Press.

HASTIE, T., TIBSHIRANI, R., & FRIEDMAN, J. (2001). *The elements of sta-*

tistical learning: Data mining, inference, and prediction. New York: Springer.
HAYES, K. J. (1953). The backward curve: A method for the study of learning. *Psychological Review*, **60**, 269-275.
JEFFREYS, H. (1935). Some tests of significance, treated by the theory of probability. *Proceedings of the Cambridge Philosophical Society*, **31**, 203-222.
JEFFREYS, H. (1961). *Theory of probability* (3rd ed.). Oxford: Oxford University Press.
KARABATSOS, G. (2006). Bayesian nonparametric model selection and model testing. *Journal of Mathematical Psychology*, **50**, 123-148.
KASS, R. E., & RAFTERY, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.
LAGARIAS, J. C., REEDS, J. A., WRIGHT, M. H., & WRIGHT, P. E. (1998). Convergence properties of the Nelder–Mead simplex method in low dimensions. *SIAM Journal of Optimization*, **9**, 112-147.
LEE, M. D., & WEBB, M. R. (2005). Modeling individual differences in cognition. *Psychonomic Bulletin & Review*, **12**, 605-621.
MACK, A., & ROCK, I. (1998). *Inattentional blindness*. Cambridge, MA: MIT Press.
MACMILLAN, N. A., & CREELMAN, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Erlbaum.
MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
MEDIN, D. L., & SCHAFFER, M. M. (1978). Context theory of classification learning. *Psychological Review*, **85**, 207-238.
MINDA, J. P., & SMITH, J. D. (2002). Comparing prototype-based and exemplar-based accounts of category learning and attentional allocation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 275-292.
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.
NAVARRO, D. J., GRIFFITHS, T. L., STEYVERS, M., & LEE, M. D. (2006). Modeling individual differences using Dirichlet processes. *Journal of Mathematical Psychology*, **50**, 101-122.
NAVARRO, D. J., PITT, M. A., & MYUNG, I. J. (2004). Assessing the distinguishability of models and the informativeness of data. *Cognitive Psychology*, **49**, 47-84.
NOSOFSKY, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
NOSOFSKY, R. M., & ZAKI, S. R. (2002). Exemplar and prototype models revisited: Response strategies, selective attention, and stimulus generalization. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **28**, 924-940.
ODEN, G. C., & MASSARO, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, **85**, 172-191.
REED, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, **3**, 382-407.
RISSANEN, J. (1978). Modeling by the shortest data description. *Automatica*, **14**, 465-471.
RISSANEN, J. (1987). Stochastic complexity. *Journal of the Royal Statistical Society B*, **49**, 223-239, 252-265.
RISSANEN, J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, **42**, 40-47.
RISSANEN, J. (2001). Strong optimality of the normalized ML models as universal codes and information in data. *IEEE Transactions on Information Theory*, **47**, 1712-1717.
SCHWARZ, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461-464.
SHEPARD, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, **237**, 1317-1323.
SIDMAN, M. (1952). A note on functional relations obtained from group data. *Psychological Bulletin*, **49**, 263-269.
SIEGLER, R. S. (1987). The perils of averaging data over strategies: An example from children's addition. *Journal of Experimental Psychology: General*, **116**, 250-264.
WAGENMAKERS, E.-J., RATCLIFF, R., GOMEZ, P., & IVERSON, G. J. (2004). Assessing model mimicry using the parametric bootstrap. *Journal of Mathematical Psychology*, **48**, 28-50.
WIXTED, J. T., & EBBESEN, E. B. (1991). On the form of forgetting. *Psychological Science*, **2**, 409-415.

NOTES

1. Recent work, such as Bayesian hierarchical modeling (e.g., Lee & Webb, 2005), allows model fitting with parameters that vary within or across individuals.

2. Of course, even in the case of well-justified individual analyses, there is still the possibility of other potential distortions, such as the switching of processes or changing of parameters across trials (Estes, 1956). Those possibilities go beyond the present research.

3. We recognize that quantitative modeling is carried out for a variety of purposes, such as describing the functional form of data, testing a hypothesis, testing the truth of a formal model, and estimating model parameters. Although it is possible that the preferred method of analysis could depend on the purpose, we believe that the lessons learned here will also be of use when other goals of data analysis are pursued.

4. This assumption underlies other common model selection techniques, such as BIC (Burnham & Anderson, 1998).

5. It may be best to use a common set of parameters to fit the data from each individual. That is, rather than fitting the models either to a single set of averaged data or to individuals, using different parameters for each individual, it is possible to find the single set of model parameters that maximizes the goodness of fit for each of the individuals. This analysis was repeated for each pair of models discussed below, but the results from this third type of analysis were almost identical to those when average data were used. They will not be discussed further.

6. It is unclear how to combine measures such as sum of squared error across individuals in a principled manner.

7. Note that this is a very simple form of uninformed prior on the parameters. To be consistent with past work (Wagenmakers et al., 2004), the priors would have to be selected from a more complex prior, such as the Jeffreys's prior. Because of its ease of use, we opted for the uniform prior. Furthermore, as will be seen, the results, for the most part, seem to be unchanged by the choice of prior.

8. 13 individuals/experiment $\times$ 9 trials/condition $\times$ 2 parameter selection methods $\times$ 500 simulations.

9. Note that we use the term *optimal* guardedly. The optimality depends not only on the details of our simulation procedure, but also, more critically, on the goal of selecting the model that actually generated the data. There are many goals that can be used as targets for model selection, and these often conflict with each other. If FIA, for example, does not produce a decision criterion at the "optimal" point, one could argue that FIA is nonetheless a better goal for model selection. The field has not yet converged on a best method for model selection, or even on a best approximation. In our present judgment, the existence of many mutually inconsistent goals for model selection makes it unlikely that a single best method exists.

ARCHIVED MATERIALS

The following materials associated with this article may be accessed through the Psychonomic Society's Norms, Stimuli, and Data archive, www.psychonomic.org/archive.

To access these files, search the archive for this article using the journal name (*Psychonomic Bulletin & Review*), the first author's name (Cohen), and the publication year (2008).

FILE: Cohen-PB&R-2008.doc.
DESCRIPTION: Microsoft Word document, containing Tables A1–A16 and Figures A1–A12.

FILE: Cohen-PB&R-2008.rtf.
DESCRIPTION: .rtf file, containing Tables A1–A16 and Figures A1–A12 in .rtf format.

FILE: Cohen-PB&R-2008.pdf.
DESCRIPTION: Acrobat .pdf file, containing Figures A1–A12.

FILE: Cohen2-PB&R-2008.pdf.
DESCRIPTION: Acrobat .pdf file, containing Tables A1–A16.

AUTHOR'S E-MAIL ADDRESS: acohen@psych.umass.edu.

(*Continued on next page*)

APPENDIX A
Bayesian Model Selection

The use of BMS for model selection raises many deep issues and is a topic we hope to take up in detail in future research. We report here only a few interesting preliminary results. Let Model A with parameters θ be denoted A_θ, with associated prior probability of $A_{\theta,0}$. In the simplest approach, the posterior odds for Model A over Model B are given by

$$\sum_\theta \left[A_{\theta,0} / B_{\theta,0} \right] \left[P\left(D \mid A_\theta\right) / P\left(D \mid B_\theta\right) \right],$$

where D is the data. The sum is replaced by an integral for continuous parameter spaces. Because BMS integrates likelihood ratios over the parameter space, the simulated difference of log maximum likelihoods is not an appropriate axis for exhibiting results. Our plots for BMS show differences of log (integrated) likelihoods.

As is usual in BMS applications, the answers will depend on the choice of priors.[A1] Consider first a flat uniform prior for each parameter in each model, ranging from 0.001 to 1.0 for both coefficients and 0.001 to 1.5 for both decay parameters (these ranges were chosen to encompass the range of plausible possibilities). This approach produces the smoothed histogram graphs in Figure A1. The natural decision statistic is zero on the log likelihood difference axis. For these priors, performance was terrible: The probability of correct model selection was .52 for group analysis and .50 (chance) for individual analysis. It seems clear that the choice of priors overly penalized the exponential model, relative to the power law model, for the data sets to which they were applied.

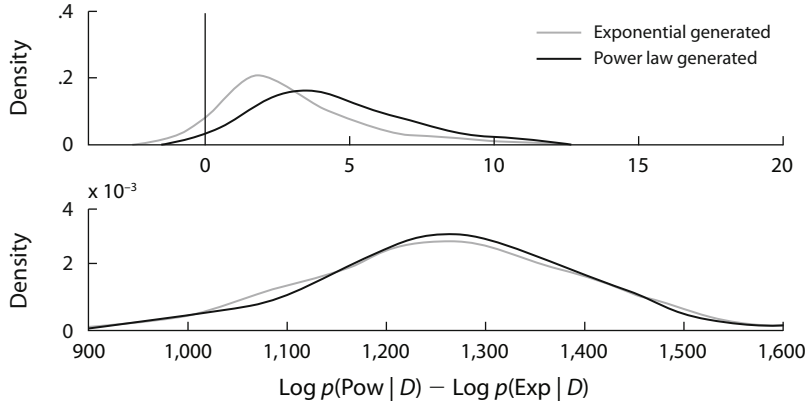

Figure A1. Histograms showing the distribution of Bayesian model selection results for the exponential and power law models with 34 subjects and two trials per condition. The group data results are shown in the top panel, and the individual data results are shown in the bottom panel. A uniform prior was used in this simulation, with the same parameter range for both models.

By changing the priors, we can greatly change the results of BMS. For example, employing priors that matched the parameter distributions used to generate the data increased model selection accuracy to levels comparable to those of PBCM and NML. We did not pursue these approaches further because, in most cases, in practice one would not have the advance knowledge allowing selection of such "tailored" priors. In future research, we intend to explore other Bayesian approaches that could circumvent this problem. In particular, various forms of hierarchical Bayesian analysis look promising. For example, one could assume that individual parameter estimates are drawn from a Gaussian distribution with some mean and variance (and covariance). It seems likely that such an assumption would produce a tendency for individual parameter estimates to move closer to the group mean estimates—in a sense, interpolating between individual and group analysis. Other possible approaches are the BNPMS method recently proposed by Karabatsos (2006) and modeling individual subjects with a Dirichlet process (Navarro, Griffiths, Steyvers, & Lee, 2006).

NOTE

A1. Note that all inference implicitly involves prior assumptions. That this assumption is made explicit is one of the strengths of Bayesian analysis.

APPENDIX B
Simulation Details

General Methodology

All of the simulations were run as described in the main text, with all combinations of 1, 2, 3, 4, 5, 6, 7, 8, 10, 14, 17, 20, and 34 individuals per experiment and for 1, 2, 3, 5, 7, 10, 13, 17, and 25 trials per condition. This range spans the range of likely experimental values and invites analyses in which trials per condition and individuals per experiment are varied but the total number of trials per experiment is held roughly constant. All of the simulations reported below were run in MATLAB, using the simplex search method (Lagarias, Reeds, Wright, & Wright, 1998) to find the maximum likelihood parameters.[B1] To reduce potential problems with local minima, each fit was repeated three times with different, randomly selected starting parameters. The ranges of possible parameters for both the generating and the fitting models were constrained to the model-specific values given below. Each of the two distributions was transformed into an empirical cumulative density function (cdf), which functions were, in turn, used to find the optimal criterion, the criterion that minimizes the overall proportion of model misclassifications. If the two distributions are widely separated, there may be a range of points over which the proportion of misclassifications does not vary. In this case, the optimal criterion was the mean of this range. To get a feel for the stability of the optimal criterion, this procedure was repeated five times. On each repetition, 400 of the points were used to generate the optimal criterion, and the remaining 100 points (0–100, 100–200, etc.) were used to determine the number of misclassifications. In general, the standard deviation across these five repetitions was well below 0.05, suggesting very stable estimates, and will not be discussed further.

As was mentioned earlier, the overall fit measure when the models were applied to individual data was generated by adding the difference of log likelihood fits (or equivalently, multiplying the likelihood ratios) for each of the individuals. Assuming that the individuals are independent of one another, this (logarithm of the) likelihood ratio is the probability of the data given the first model and its best-fitting parameters over the probability of the second model and its best-fitting parameters.

Models of Forgetting: Power and Exponential

The proportions of incorrect model selections for the exponential and power models are given in the online supplemental material (www.psychonomic.org/archive).

Uninformed. For the uninformed method of parameter selection, the parameters a and b for the power and exponential models were restricted to lie in the ranges 0.001–1.0 and 0.001–1.5, respectively. These values were selected both to encompass the range of values seen in human studies and to ensure that the recall probabilities stayed in the range of 0–1. For each simulated experiment, a single "mean" parameter value was randomly selected from each of these ranges. The parameter values for each individual were then selected from a normal distribution with the mean as the "mean" parameter value and standard deviation as 5% of the overall parameter range. If an individual's parameter value fell outside of the acceptable parameter range, a new sample was drawn from the distribution (e.g., the distribution was resampled). It is important to note that the reported results (and the results for all of the simulations) hold only for this parameter range, since changing the range may change the results.

Informed. The best-fitting parameters (using the maximum likelihood measure) for each model were found for each of the 8 subjects in Wixted and Ebbesen (1991). The simulation parameters were then randomly selected from these two sets of parameters (one for each model) with replacement and without added noise.

Models of Information Integration: FLMP and LIM

The proportions of incorrect model selections for the FLMP and LIM models are given in the online supplemental material (www.psychonomic.org/archive).

Uninformed. The FLMP and LIM have 10 parameters each, $a_{1..5}$ and $v_{1..5}$, and each parameter is restricted to lie in the range of 0.001–0.999 (0 and 1 are excluded to prevent taking the log of zero in the maximum likelihood calculation). When uninformed parameters were simulated, a single vector of "mean" parameter values was randomly selected for each simulated experiment. Because of the restricted range of the parameters and the resampling procedure (see below), it is not possible to set a variance for each parameter and expect that the parameter distribution will conform to this variance. Instead, the parameter values for each individual were selected from a normal distribution with the mean as the "mean" parameter value and a standard deviation of 0.16. If an individual's parameter value fell outside of the acceptable parameter range, the value was resampled. After resampling, the average standard deviation for a single parameter in the one trial per condition and 34 individuals per experiment simulation was 0.13 for both models. The standard deviation changed very little over parameter position,[B2] ranging from an average of 0.121 to 0.145.

Informed. The parameters were sampled from the best-fitting parameters of the data from the 82 individuals in Massaro (1998).

(*Continued on next page*)

APPENDIX B (Continued)

Models of Categorization: GCM–R, GCM–γ, and Prototype

The proportion of incorrect model selections for the GCM–γ and prototype models is given in the online supplemental material (www.psychonomic.org/archive). To prevent infinite log values when the maximum likelihood parameters were found, all probabilities were restricted to lie in the range of 0.001–0.999.

Uninformed. For the GCM–R, GCM–γ, and prototype models, data when the uninformed prior was used were generated as follows. For both the GCM–γ and the prototype models, c and γ were restricted to be in the range of 0.1–15. Each of the four raw (i.e., before they were constrained to sum to 1) w_ms began in the range of 0.001–0.999. For each simulated experiment, single "means" were randomly selected for the c, γ, and four w_m parameter values from each of these ranges. Each w_m was then normalized by the sum of all of the w_ms. For c and γ, the parameter values for each individual were selected from a normal distribution with the mean as the "mean" parameter value and standard deviation as 5% of the overall parameter range, respectively. The individual w_m parameters were selected using the same method, but, because of the extremely restricted range and normalizing procedure, the standard deviation was 1% of the overall parameter range. If an individual's parameter value fell outside of the acceptable parameter range, the value was resampled. The weights were then renormalized. Because of the normalization and resampling procedures, the standard deviation of the w_m will not be exactly 0.01. For example, the average standard deviation for a single w_m in the one trial per condition and 34 individuals per experiment simulation was 0.0097.

For the GCM–R, selections of the c and the w_m parameters are unchanged from the GCM–γ, but γ is fixed at 1.

Informed. The best-fitting parameters for each model were also found for each of the 48 subjects in Minda and Smith's (2002) Experiment 2, who participated in eight trials per condition. This set of parameters was used as the pool from which informed parameters were drawn with replacement.

NOTES

B1. For some of the models below, closed form solutions exist for the maximum likelihood parameters, but we opted to perform a search for the parameters.

B2. Although, in this context, the order of the parameters does not matter, they were sorted from lowest to highest value for each dimension.

APPENDIX C
Predictive Validation

For all of the models examined in this article, the distributions discussed in the text are binomial. For two binomial distributions, the K–L divergence between the probability of a success under the generating model, p_G, and the probability of success under the comparison model, p_C, is

$$\sum_i \frac{n!}{i!(n-i)!} p_G^i \left(1 - p_G\right)^{n-i} \left[i \log \frac{p_G}{p_C} + (n-i) \log \frac{1-p_G}{1-p_C} \right], \tag{C1}$$

where i is the number of successes. The quantity in Equation C1 is summed over all conditions and subjects to produce the total K–L divergence between the generating and the comparison models. The generating models in this article used different parameters for each individual. To make as close a comparison as possible, the data were taken from the informed PBCM simulations described in the main text. For individually fitted models, the K–L divergence was summed over the fitted and generating parameters for each individual. To make the models fitted to the group data comparable, the K–L divergence was summed over each generating individual against the single set of group parameters.

(Manuscript received June 4, 2007;
revision accepted for publication December 4, 2007.)