Psychonomic Bulletin & Review
2005, 12 (3), 403-408

Risks of drawing inferences about cognitive processes from model fits to individual versus average performance

W. K. ESTES
Indiana University, Bloomington, Indiana

and

W. TODD MADDOX
University of Texas, Austin, Texas

With the goal of drawing inferences about underlying processes from fits of theoretical models to cognitive data, we examined the tradeoff of risks of depending on model fits to individual performance versus risks of depending on fits to averaged data with respect to estimation of values of a model's parameters. Comparisons based on several models applied to experiments on recognition and categorization and to artificial, computer-generated data showed that results of using the two types of model fitting are strongly determined by two factors: model complexity and number of subjects. Reasonably accurate information about true parameter values was found only for model fits to individual performance and then only for some of the parameters of a complex model. Suggested guidelines are given for circumventing a variety of obstacles to successful recovery of useful estimates of a model's parameters from applications to cognitive data.

Theoretical models may be applied to cognitive performance data for any of several distinct purposes. Among these purposes are:

1. To smooth trends in noisy data. In the wake of the pioneering example of Ebbinghaus (1885/1964), this goal characterized the great body of work on curves of practice and forgetting for more than a half century.
2. To select from among a set of alternative models the one that best meets some criterion of success in accounting for relevant empirical data.
3. To obtain the estimates of a particular model's parameters that yield the best prediction of the set of observed performance scores of a group of subjects. This goal is well achieved by the hierarchical linear modeling approach (Bryk & Raudenbush, 1987). A limitation of this approach is that if there are individual differences among subjects of a group with respect to values of the model's parameters, the predictions of the model for any one subject depend to some degree on the performances of the other members of the group.
4. To obtain the estimates of the model's parameters that best serve the purpose of yielding evidence about underlying cognitive processes in individual subjects when applying a particular model to a set of cognitive performance scores.

Although the methods discussed in this article may contribute indirectly to Goals 1 and 2, our focus is strictly on Goal 4. However, in pursuing this goal, an investigator faces a dilemma: Best-fitting[1] models predict true performance of individuals, but individual performance data to which models are applied reflect a composite of true effects and experimental error (Brown & Heathcote, 2003; Estes, 1956, 2002; Maddox & Estes, 2004; Myung, Kim, & Pitt, 2000). To reduce the effects of error, investigators frequently resort to using averaged data for groups of subjects. There is danger, however, that individual differences among subjects with respect to values of a model's parameters may cause averaging to produce distorted inferences about true patterns of individual performance and the cognitive processes underlying them.

The tradeoff between the risks of depending on model fits to error-prone individual performance and the risks of fitting averaged data is the subject of this article. We depart from the tradition of nearly all previous work on this problem, first by concentrating upon the effects of averaging on recovery of the values of a model's parameters from a fit of the model to data, and also by giving prime attention to magnitudes rather than simply the existence of effects.

Effects of Individual Differences and Averaging in Experimental Data

In this section, we treat data analyses based on two ways of fitting a model: (1) estimation of the model's pa-

The research reported was supported by NSF Grants SBR-996206 and BCS-0130512 and NIH Grant MH 59196. Reprint requests or correspondence about this article should be addressed to W. K. Estes, Department of Psychology, Indiana University, Bloomington, IN 47405 or to W. T. Maddox, Department of Psychology, Mezes Hall 330, University of Texas, Austin, TX 78712 (e-mail: wkestes@indiana.edu, maddox@psy.utexas.edu).

rameters separately from the fit of the model to the data of each individual subject in a group and (2) estimation of the model's parameters from the fit of the model to the averaged data of the group. Methods 1 and 2 are henceforth designated FI and FA, respectively.

All analyses are based on applications of array models (Estes, 1994; Estes & Maddox, 2002; Maddox & Estes, 2004). Samples of data analyzed come from two studies of recognition memory and a study of category learning.

The design of the experiment for which we give the most detailed analysis, henceforth termed the *list strength experiment*, is essentially that used in much recent research on tests of models for recognition memory (Murdock & Kahana, 1993; Ratcliff, Clark, & Shiffrin, 1990; Shiffrin, Huber, & Marinelli, 1995; Shiffrin, Ratcliff, & Clark, 1990). A second experiment, termed the *familiarity–recollection experiment*, deals with recognition memory in relation to issues raised by Jacoby (1991); its analysis bears on the role of model complexity. The third, *category learning*, experiment is relevant to generality across experimental paradigms. Essentials of all three experiments are described in Appendix A.

Parameter estimation: List strength experiment. The first task we address is a comparison of parameter estimation from FI and FA for the list strength experiment. The relevant results in terms of mean estimates of parameters are shown in Table 1. The four parameters—s_m, s_nm, a_1, and a_2—are defined in Appendix B.

Of these, s_m and s_nm pertain to similarity comparisons between test items and their representations in memory, and a_1 and a_2 to probability that an item presented on a study trial has its representation stored in memory. A notable finding exhibited in Table 1 is the very large overall difference in magnitudes of mean parameter values between the estimates from FI and those from FA in both replications.

The question of generality of this result immediately arises. We cannot give any final answers, but we can assemble some relevant evidence about the way FI versus FA comparisons are affected by various changes in properties of the model or data under consideration.

Table 1
Mean Estimates of Array Model Parameters in Fits to Individual or Average Data of List Strength Experiment

		Estimated Value	
Replication	Parameter	FI	FA
R1	s_m	.059	.008
	s_nm	.029	.008
	a_1	.248	.000
	a_2	.526	.000
R2	s_m	.037	.000
	s_nm	.043	.001
	a_1	.248	.000
	a_2	.350	.000

Note—FI, fit with individuals; FA, fit with average.

Parameter estimation: Familiarity–recollection experiment. Two models were applied to the same data set from an experiment on recognition memory similar to that of Yonelinas (1994). The simpler of these is of the same type as the one applied to the list strength experiment but has only three parameters to be estimated. The more complex one is a dual-process model developed as an alternative to the "remember–know" model of Jacoby (1991) and Yonelinas (1994), and it has five parameters to be estimated (see Appendix B).

The salient results for parameter estimation by FI versus FA model fits are given in Table 2. The gist of these results is that estimates are uniformly smaller in magnitude for FA and that the difference is much greater for the more complex model. In particular, extremely low estimates occur only for the more complex model. It is interesting to note that, considering only the three parameters that are common to the two models (s_1, s_2, and B), FI versus FA differences are much larger when the parameters are embedded in the more complex model.

Parameter estimation: Category learning experiment. The data were fitted using both FI and FA for a simple exemplar-based model with three parameters, denoted s_m, s_nm, and a (Appendix B). Mean estimates of the three parameters, in the order just given, were .6781, .0958, and .9905 for FI and .0001, .0637, and .9740 for FA.

Estimates from FA are uniformly smaller than those from FI, as in all of the preceding analyses, with differences ranging from small to extremely large.

Effects of Averaging and Individual Differences in Artificial Data

Constructing an artificial data set. The critical question left untouched by the results of the preceding section is how model fits by FI or FA compare with respect to recovering true parameter values. For real experimental data, the question cannot be answered, because even if the model used is "correct" for the situation, the true values of its parameters for the real subjects are necessarily unknown. However, some progress can be made by using artificial, computer-generated data for which true parameter values are known.

To obtain a data set appropriate for our purposes, we used a tactic described previously by Maddox and Estes (2004).

The gist of the method was to yoke the simulation to a real experiment. We started with the parameter values estimated for each of the 48 subjects in each of replications R1 and R2 of the list strength experiment and generated "true" scores for 48 hypothetical subjects using the computer program that embodied the array model. However, these scores could not be used, as they stood to constitute the "observed" data for the simulations because they were error free. Thus, we added to the score for each hypothetical subject on each trial an error variable assumed to come from a population with a mean of 0 and standard deviation σ_{ij}. To obtain data at two levels

Table 2
Mean FI Versus FA Parameter Estimates for Single- and Dual-Process Models

	Dual-Process Model		Single-Process Model	
Parameter	FI	FA	FI	FA
s_1	.037	.004	.027	.025
s_2	.727	.615	.402	.373
B	2.558	0.669	2.019	1.528
r	.160	.105		
r'	.115	.003		

Note—FI, fit with individuals; FA, fit with average.

of error, we drew values of the error variable from this population with σ_{ij} set equal to .10 for one simulation and to .14 for a second simulation in each replication.

With the simulated performance scores in hand, we fitted the array model to the artificial data with our parameter-search program for FI and FA just as we had done previously for the real data of the list strength experiment.

Estimation of parameters. For the data of the R1 and R2 replications at both levels of error, parameter estimates obtained with the two methods of model fitting are summarized in Table 3. The format of Table 3 differs from that of Table 2 by the addition of a column of mean true values (i.e., the values that were used in generation of the artificial data). On the whole, the mean parameter estimates obtained for the 48 individual subjects by FI move toward the true values as the error level (σ_{ij}) drops from .14 to .10, and the estimates exhibit patterns of relative value over the four parameters that agree reasonably well with those of the true values.

The same trends do not hold for the estimates produced by FA. FA estimates uniformly fall into a very narrow range, as was found in the analysis of the real data (Table 1), and do not conform well to the pattern that holds for the true values.

To bring out the relation between the FI–FA difference and the number of subjects entered into the model fits, we expanded the analysis of the $\sigma_{ij} = .10$ segment of the artificial data set to give results separately for subgroups of 6, 12, and 24 subjects, as well as for the full 48. The results, exhibited in Table 4, show that the correspondence between patterns of FI parameter estimates and of true (Tr) values and the frequency of extremely small estimates by FA both increase with group size.

In view of our goal of obtaining useful parameter estimates for individual subjects, it is important to go beyond comparisons of means by examining comparisons for individuals. One approach is illustrated in Figure 1, which shows for each individual simulated subject in the artificial data set the difference between the true value of a parameter and the value estimated by FI compared with this difference for the estimate by FA.

For a similarity parameter (upper panel), estimates by FI are closer to true values for more than 3/4 of individuals than are any estimates by FA. For a storage probability parameter (lower panel), estimates by FI are close to the true values for about 1/3 of individuals, but estimates by FA for almost none.

Discussion

Summary of findings. Regarding the advantages and disadvantages of FI and FA for applications of cognitive models, an overall conclusion from the work reported in this article is that one should think of this issue in terms of tradeoffs rather than of sweeping generalizations. As is pointed up most sharply in Figure 1, we have found FI to be far superior to FA in yielding accurate recovery of true parameter values for individual subjects; but we have also found FI to yield the largest disparities between true and estimated values.

One of the most pervasive factors influencing parameter recovery by either FI or FA has proved to be model complexity. From mathematical analyses, it is known that, under some conditions, the two types of model fitting are indistinguishable: (1) when the model is linear in its parameters, as in linear regression; (2) when the model includes no more than two free parameters and is simple enough in structure that "model-preserving" forms of averaging can be used (Estes, 1956; Myung, Kim, & Pitt, 2000). When these conditions are not met, however, as in the case of most models used in research on memory, categorization, and decision making, hazards of indiscriminate use of FA become acute (Ashby, Maddox, & Lee, 1994; Maddox, 1999).

Table 3
True Values and Estimates of Array Model Parameters for Artificial Data With Error σ Equal to .10 or .14

			Estimate From Simulation			
			FI		FA	
Replication	Parameter	True	.10	.14	.10	.14
R1	s_{m}	.023	.058	.064	.015	.017
	s_{nm}	.010	.026	.038	.006	.007
	a_1	.389	.175	.148	.004	.001
	a_2	.529	.320	.218	.008	.003
R2	s_{m}	.020	.048	.036	.018	.020
	s_{nm}	.007	.025	.026	.008	.009
	a_1	.375	.212	.139	.012	.003
	a_2	.538	.312	.208	.018	.005

Note—FI, fit with individuals; FA, fit with average.

Table 4
Mean Parameter Estimates for Artificial Data With Error σ Equal to .10 as a Function of Group Size

	Number of Subjects											
	6			12			24			48		
Parameter	Tr	FI	FA	Tr	FI	FA	Tr	FI	FA	Tr	FI	FA
s_m	.02	.09	.03	.02	.03	.03	.02	.05	.03	.02	.05	.02
s_{nm}	.00	.04	.02	.01	.02	.01	.01	.03	.02	.01	.03	.01
a_1	.49	.12	.00	.36	.26	.08	.33	.23	.00	.38	.21	.01
a_2	.64	.22	.01	.50	.39	.11	.50	.35	.01	.53	.34	.01

Note—Tr, true value; FI, fit with individuals; FA, fit with average.

Analyses of both experimental and artificial data in this study complement the work of Myung (2000) on the role of complexity in model selection by showing that differences between results of parameter recovery from FI versus FA depend strongly on model complexity, as do differences in degree of approximation of parameter estimates to true values.

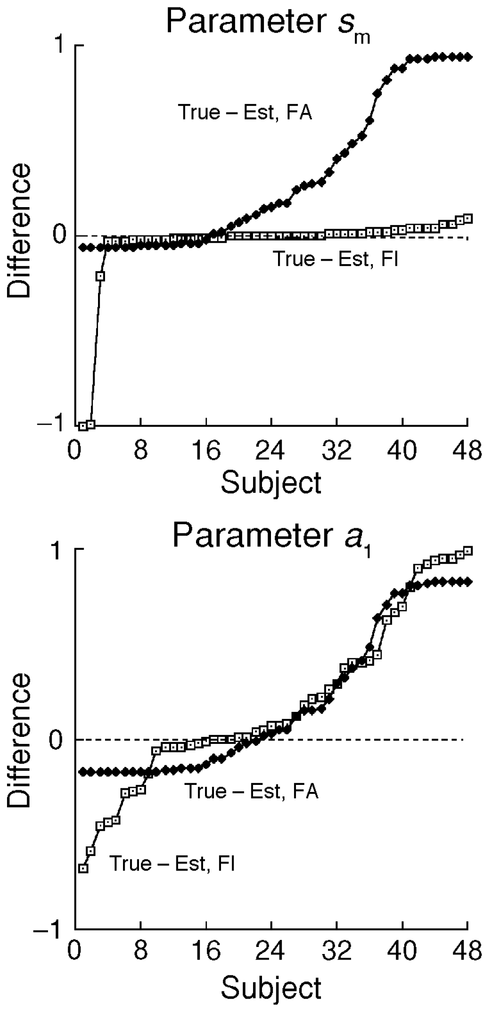

Figure 1. Difference for each subject in the artificial data between true parameter value and estimate by FI or FA. Parameters s_m and a_1 are represented in the upper and lower panels, respectively.

In contrast, FI versus FA comparisons seem insensitive to the nature of the response scale (ratings on a continuous scale versus binary choices) and to number of conditions in an experiment, but they do depend strongly on group size.

The question might be raised whether FA is being unfairly penalized in the analyses reported because, for some reason, the model-fitting program does not produce as close agreement between observed and predicted performance scores for FA as for FI. This possibility can be readily checked for any situation of interest. In the case of our analysis of an artificial data set, for example, the mean square of disparities between observed and predicted scores was .008 for FA. For the 48 subjects in FI, this measure ranged from .005 to .035, signifying a poorer, not a better, fit in FI than in FA.

We are concerned with gaining information about parameters of a model from fits to noisy data not as an end in itself, but because this information contributes to evaluation of the usefulness of a model for elucidation of the cognitive processes underlying performance. For instance, given a model that has accrued empirical support, two versions of the model are fitted to a data set, one version including and the other not including a parameter corresponding to a particular process. Superiority of the former version is taken to support inclusion of the given process in the model. In the studies reported in this article, we found that information relevant to this objective was obtained by analyses of type FI but virtually never by those of type FA.

Analyzing fits of several models in different experimental designs yielded evidence suggesting some generality for our findings across research situations, but this result should not be overinterpreted. The main point we wish to emphasize is that our results point up first the failures of parameter recovery that can be produced by averaging, and also the desirability of studying the application of a new model to data in the manner illustrated in this article before the model is used as a tool for gaining information about the cognitive processes that underlie performance.

On the use of artificial data. Although many aspects of FI versus FA comparisons have been effectively illustrated by model fits to experimental data, for the purpose of evaluating the capability of model fits to yield information about true parameter values, it is essential to analyze fits to artificial data for which true parameter values are known. This procedure needs to be used with some care, however, and we list here some constraints that we have found essential.

1. To yield results that are of practical relevance to an investigator's purposes, the data need to be generated by the model under consideration for a hypothetical experiment of the same design as the one used in the investigator's relevant research, with the number of hypothetical subjects matching the number of real subjects, with the same performance measure, and with values of the model's parameters chosen so that the real and artificial performance scores fall in about the same range with respect to magnitude.

2. It has become well accepted in cognitive modeling that performance scores obtained in cognitive research must represent composites of true scores and error (Brown & Heathcote, 2003; Estes, 2002; Myung, Kim, & Pitt, 2000; Pitt, Myung, & Zhang, 2002). As in our related work (Maddox & Estes, 2004), we realized this requirement by adding to each model-generated performance score an error component drawn from a normal distribution with a variance chosen to keep the error component from being so small as to have negligible effects or so large as to cause scores to stray out of the allowable range.

To construct artificial data sets appropriate for some kinds of performance measures (e.g., binary choices or reaction times), replacement of the normal error distribution with binomial, gamma, or other distributions would not be expected to affect any of the conclusions of this study. In all cases, it might be necessary to truncate the distributions to handle the range problem. Furthermore, we have found it good practice to check on the comparability of real and artificial data sets by examining the results of analyses that can be done similarly on real and artificial data (Maddox & Estes, 2004).

Guidelines. Some of our findings have implications for cognitive modeling that we have translated into guidelines for our own work. For example:

1. In all research involving model fitting, use adequate numbers of subjects, even at the cost of conducting fewer experiments.

2. Avoid the use of FA for preliminary appraisals of models in small pilot studies or in applications to results of other investigators that have been published only in the form of predictions of models for averaged data.

3. When drawing conclusions from applications of FI, go beyond dependence on parameter estimates averaged over subjects and examine the full range of estimates obtained for individuals (as can be accomplished by the type of analysis illustrated in Figure 1).

REFERENCES

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.

Brown, S., & Heathcote, A. (2003). Bias in exponential and power function fits due to noise: Comment on Myung, Kim, and Pitt. *Memory & Cognition*, **31**, 656-661.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, **101**, 147-158.

Ebbinghaus, H. (1964). *On memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: Dover. (Original work published 1885)

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

Estes, W. K. (1994). *Classification and cognition*. New York: Oxford University Press.

Estes, W. K. (2002). Traps in the route to models of memory and decision. *Psychonomic Bulletin & Review*, **9**, 3-25.

Estes, W. K., & Maddox, W. T. (2002). On the processes underlying stimulus-familiarity effects in recognition of words and non-words. *Journal of Experimental Psychology*: *Learning*, *Memory*, *& Cognition*, **28**, 1003-1018.

Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory & Language*, **30**, 513-554.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, **61**, 354-374.

Maddox, W. T., & Estes, W. K. (2004). Predicting true patterns of cognitive performance from noisy data. *Psychonomic Bulletin & Review*, **11**, 1129-1135.

Murdock, B. B., & Kahana, M. J. (1993). Analysis of the list-strength effect. *Journal of Experimental Psychology*: *Learning*, *Memory*, *& Cognition*, **19**, 689-697.

Myung, I. J. (2000). The importance of complexity in model selection. *Journal of Mathematical Psychology*, **44**, 190-204.

Myung, I. J. (2003). Tutorial on maximum likelihood estimation. *Journal of Mathematical Psychology*, **47**, 90-100.

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.

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

Ratcliff, R., Clark, S. E., & Shiffrin, R. M. (1990). List-strength effect: I. Data and discussion. *Journal of Experimental Psychology*: *Learning*, *Memory*, *& Cognition*, **16**, 163-178.

Shiffrin, R. M., Huber, D. E., & Marinelli, K. (1995). Effects of category length and strength on familiarity in recognition. *Journal of Experimental Psychology*: *Learning*, *Memory*, *& Cognition*, **21**, 267- 287.

Shiffrin, R. M., Ratcliff, R., & Clark, S. E. (1990). List-strength effect: II. Theoretical mechanisms. *Journal of Experimental Psychology*: *Learning*, *Memory*, *& Cognition*, **16**, 179-195.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology*: *Learning, Memory, & Cognition*, **20**, 1341-1354.

NOTE

[1] By the "fit" of a model to data we mean the goodness of agreement between observed and predicted performance scores as measured either by the computed likelihood of the data given the model or by the mean of squared discrepancies between observed and predicted scores. The likelihood measure has theoretical advantages if the results are being used for model selection (Myung, 2003), but for the purposes of this article, we find the squared discrepancies (MS_e) method preferable.

APPENDIX A
Sources of Experimental Data

All data used in illustrative analyses came from unpublished studies conducted in our research program.

List strength experiment. A standard study–test paradigm was employed with 48 university undergraduates as subjects. Each participated in the same design in which list length (4 or 8 items), study duration (400 or 1,200 msec per item), and study frequency (one or two occurrences of an item in the study list) were crossed factorially. The entire design was replicated for each subject with two different samples of pseudoword stimuli from the same population, the replications being labeled R1 and R2. The recognition test for each list included all items of the study list plus an equal number of new items drawn randomly from the master list. For each test item, subjects gave a rating on a 0–100 scale of confidence that the item was "old" (i.e., that it came from the study list).

Familiarity–recollection experiment. This experiment on study–test recognition memory was concerned with the question of whether recognition depends both on familiarity of test items and on recollection of occurrences of the items in study contexts. Following Yonelinas (1994), presentation of two study lists successively was followed by separate tests on the two lists. Performance scores were binary choices ("old" or "new" responses to test items). The subjects were 30 university undergraduates.

Category learning experiment. The experiment constituted five blocks of trials, each composed of 20 observation trials followed by 15 test trials. The stimuli were pseudowords, and the categories were verb or noun. Performance scores on the tests were subjects' ratings (on a scale of 0 to 100) of confidence that the test item belonged to a specified category. The subjects were 37 university undergraduates.

APPENDIX B
Parameters of the Array Model

The array model for category learning is described by Estes (1994), and the versions applicable to recognition memory by Estes and Maddox (2002) and Maddox and Estes (2004). For the applications in this article, parameters to be estimated were as follows.

List strength experiment and artificial data. The parameters were s_{m}, s_{nm}, a_1, and a_2. The first two represented similarities, on a 0–1 scale, between a test item and an item representation in memory: s_{m} applied when the study frequency of a test item matched that of the memory element it was being compared with, and s_{nm} applied when there was no match. Parameters a_1 and a_2 were probabilities that a studied item was stored in memory when study time was short or long, respectively.

Familiarity–recollection experiment. In the simpler model, the parameters were s_1, s_2, and B, where s_1 and s_2 applied to tests on List 1 or List 2, respectively, and B was a reference parameter.

In the more complex model, the parameters were s_1, s_2, B, r, and r', the last two being interpretable as probabilities of correct or incorrect recollection on test trials.

Category learning experiment. The parameters were s_{m}, s_{nm}, and a, where a was the storage probability on study trials.

(Manuscript received May 5, 2004;
revision accepted for publication November 23, 2004.)