

Flexibility versus generalizability in model selection

MARK A. PITT, WOOJAE KIM, and IN JAE MYUNG
Ohio State University, Columbus, Ohio

Which quantitative method should be used to choose among competing mathematical models of cognition? Massaro, Cohen, Campbell, and Rodriguez (2001) favor root mean squared deviation (RMSD), choosing the model that provides the best fit to the data. Their simulation results appear to legitimize its use for comparing two models of information integration because it performed just as well as Bayesian model selection (BMS), which had previously been shown by Myung and Pitt (1997) to be a superior alternative selection method because it considers a model's complexity in addition to its fit. In the present study, after contrasting the theoretical approaches to model selection espoused by Massaro et al. and Myung and Pitt, we discuss the cause of the inconsistencies by expanding on the simulations of Massaro et al. Findings demonstrate that the results from model recovery simulations can be misleading if they are not interpreted relative to the data on which they were evaluated, and that BMS is a more robust selection method.

Myung and Pitt (1997) described a method for selecting among mathematical models of cognition, dubbed Bayesian model selection (BMS). It goes beyond more widely used selection methods by not only evaluating a model's ability to fit data (i.e., measuring goodness-of-fit using mean squared error or percent variance accounted for), but also by considering another dimension of a model, its complexity, which conceptually refers to how many different data patterns the model can fit a priori. Massaro, Cohen, Campbell, and Rodriguez (2001) carried out a more extensive series of simulations, the primary purpose of which was to show that prior work demonstrating the superiority of FLMP (fuzzy logical model of perception; Massaro, 1998) relative to WTAV (weighted averaging model) was still valid when BMS was used in place of RMSD (root mean squared deviation), their preferred selection method. Overall, RMSD performed equivalently to BMS, even better in some instances, which is at odds with Myung and Pitt, who found BMS to outperform RMSD. The purpose of the present investigation was to determine the cause of this inconsistency across studies, for an unreliable selection method may be worse than no selection method at all. Our aim was not to debate the merits of FLMP and WTAV as models of information integration. On the contrary, we agree with Massaro et al.'s conclusions regarding the superiority of FLMP.¹

We begin by describing how our theoretical approach to model selection differs from Massaro et al.'s (2001), explaining why we advocate the use of BMS over RMSD.

This work was supported by Research Grant R01 MH57472 from the National Institute of Mental Health. We thank Donald Bamber, Michael Lee, Dominic Massaro, and Peter Grunwald for helpful comments on an earlier draft. Correspondence should be addressed to M. A. Pitt, W. Kim, or I. J. Myung, Department of Psychology, Ohio State University, 1885 Neil Avenue, Columbus, OH 43210 (e-mail: pitt.2@osu.edu, kim.1124@osu.edu, or myung.1@osu.edu).

The remainder of the paper is a reexamination of Massaro et al.'s simulations. With the aid of new analysis techniques, the results of large-scale simulations provide a better understanding of why the two selection methods performed as they did across the various testing situations. In addition, they further elucidate the relative performance of the two selection methods. The paper closes with some thoughts on when and how each method can be used productively.

THE GOAL OF MODEL SELECTION

The goal of model selection is to identify the one model, from a set of competing models, that best captures the regularities underlying the cognitive process of interest. Approaches to the problem can differ greatly depending on how the goal of model selection is defined. The measures that are developed to choose among models may also differ because they are a reflection of the approach.

The approach to model selection espoused by Massaro et al. (2001) places a premium on model flexibility. Models are evaluated on their ability to fit the range of data sets generated in an experiment. For each set of data (i.e., participant), the parameters of the model vary freely to obtain the optimal fit, which is assessed using RMSD as the goodness-of-fit (GOF) measure (Massaro, 1998). The model that yields the best mean fit when averaged across participants is chosen because it is assumed to approximate most closely the underlying cognitive process. In defense of their method, Massaro et al. argued that fitting individual participants' data is necessary to avoid the potential distorting effects of fitting averaged data. They also maintained that allowing parameters to vary freely is a necessary part of modeling.

We are in agreement with Massaro et al. (2001) that care must be taken in fitting averaged data. If response patterns

are similar across participants, there may be no harm in doing so, but averaging should be avoided if there are substantial individual differences (Myung, Kim, & Pitt, 2000). Without a minimal amount of initial exploratory modeling, it is nonsensical to fix the parameters of a model prior to testing it on a given data set. How else could good parameter estimates be obtained?

We have reservations, however, about a flexibility-based approach to model selection and the types of models it favors. To begin with, we take issue with the practice of generating new parameter estimates for each *new* set of data. By allowing parameters to vary anew for fitting each participant's data using a GOF measure (like RMSD), the model is, in effect, being optimally tuned to each set of data. When these RMSD values are then averaged, the result is a measure of the flexibility of the model; that is, what is being measured is the ability of the model to capture the range of response patterns exhibited across participants. Models that provide a good mean fit (i.e., small RMSD) are sufficiently agile to cope with most of the variability in the data, whether it is random error or due to the process of interest. Less flexible models will usually provide poorer mean fits because they are unable to do so. In essence, the model that is most adept at absorbing variability across response patterns (whatever its source) is selected, and it is this model that is assumed to approximate most closely the underlying cognitive process.

The preceding scenario will tend to hold when the models being compared capture the main trends or patterns in the data. If one model is incapable of doing so, then its competitor will more than likely be chosen using RMSD, even if the latter is unquestionably the least flexible of the two. In such a circumstance, however, quantitative selection methods are probably unnecessary for choosing the superior model.

A shortcoming of the flexibility approach becomes apparent when one asks why the most flexible model should be preferred. As flexibility increases, specificity about the form of the underlying process decreases, because a more flexible model can mimic many more response patterns (i.e., potentially unique cognitive processes) than a less flexible one (Myung & Pitt, 1997; Massaro, 1988, makes a related point in comments on connectionist models). If the goal of model selection is to narrow the set of candidate models to one, then emphasizing flexibility works against this goal because models are kept in the running solely by virtue of their ability to fit data, not because they are good approximations to a particular cognitive process. In the limit, the preferred model would simply store the data values, and be nothing more than a data replicator.

The emphasis on model flexibility in Massaro et al.'s (2001) approach leads to nontrivial errors in model selection. As mentioned above, variation in the data due to sampling error, to individual differences, and to the cognitive process itself are undifferentiated in the flexibility approach. All are considered meaningful forms of variation, and models are evaluated on their ability to capture all three, but a GOF measure like RMSD cannot distinguish among them. The following simulation illustrates this problem.

RMSD was used to choose between two models from data that were generated to vary in sampling error, individual differences, and the model (i.e., cognitive process) itself. One model was a 16-parameter version of the original fuzzy logical model perception (FLMP_F; Massaro & Cohen, 1983) and the other was a restricted (i.e., nested) version of this same model with only 8 free parameters (FLMP_R), obtained by fixing half of the 16 parameters to constants. The mathematical equations for FLMP and other models discussed in the present paper are listed in Table 1. The ability of RMSD to recover the correct model was tested in five conditions. In the first and fifth, the data from all simulated participants differed only in sampling error. In the second and fourth, they differed in both sampling error and individual differences, with half the data sets being generated using one set of parameters, and the remainder using the other set. In the third condition, half of the simulated subjects were generated by one model, and half by the other model.

In each condition and for each model (FLMP_F or FLMP_R), a set of parameters was run through the model to create a set of 64 response probabilities in an 8×8 factorial design. From these, a sample of 64 simulated proportions was obtained by introducing sampling error using the binomial probability distribution of sample size 20. This was repeated 50 or 100 times, depending upon the condition, thereby creating the same number of replication samples (i.e., response patterns). Both models were then fitted to each response pattern separately using RMSD.

Of interest in this simulation is how well RMSD discriminates the sources of variability, not the relative performance of the two models. FLMP_F will easily outfit FLMP_R given that the models differ only in that the former has twice as many free parameters. FLMP_R is a restricted ver-

Table 1
Five Models of Information Integration Employed
in the Present Work

Model	Model Equation
FLMP	$P_{ij} = \frac{\theta_i \lambda_j}{\theta_i \lambda_j + (1 - \theta_i)(1 - \lambda_j)}$
FLMP _W	$P_{ij} = \frac{\theta_i^w \lambda_j^{(1-w)}}{\theta_i^w \lambda_j^{(1-w)} + (1 - \theta_i)^w (1 - \lambda_j)^{(1-w)}}$
WTAV	$P_{ij} = w\theta_i + (1 - w)\lambda_j$
LIM	$P_{ij} = \frac{\theta_i + \lambda_j}{2}$
LIM _T	$P_{ij} = \min[\max(\theta_i + \lambda_j - 0.5, 0), 1]$

Note—Each model assumes that the probability of categorizing an input stimulus to one of two categories, denoted by P_{ij} , is a function of two parameters, θ_i and λ_j . The two parameters represent the degree of support for the category of interest given the specific i and j feature dimensions of the stimulus where $i = 1, \dots, q_1, j = 1, \dots, q_2$, and take on values between 0 and 1.

Table 2
Mean Model Recovery Rates and Percentage Wins Across Five Conditions Using RMSD

Data Source	FLMP _R		FLMP _F		Model Fitted	
	θ_a	θ_b	θ_c	θ_d	FLMP _R ($k = 8$)	FLMP _F ($k = 16$)
1. Reduced model with						
(a) Sampling error (SE _R)	100	–	–	–	0.062 (0%)	0.054 (100%)
(b) Sampling error + individual diffs (SE _R + ID)	50	50	–	–	0.063 (0%)	0.054 (100%)
2. Different models (DM)	–	50	50	–	0.092 (0%)	0.053 (100%)
3. Full model with						
(a) Sampling error (SE _F)	–	–	–	100	0.100 (0%)	0.055 (100%)
(b) Sampling Error + individual diffs (SE _F + ID)	–	–	50	50	0.114 (0%)	0.053 (100%)

Note—Sample size = 20. θ_{a-d} refer to the parameter sets used in the simulations. FLMP_F is the first version of FLMP listed in Table 1. FLMP_R is the reduced version, with θ_{4-7} and λ_{4-7} held constant at .2, .6, .7, and .9, respectively.

sion (i.e., a special case) of FLMP_F, so such a finding is not only expected, but also well known.

When looked at from the standpoint of variability, the goal of model selection is to ignore variation due to sampling error and individual differences, and capture only that due to the cognitive process of interest. In the present test, an accurate selection method should ignore sampling error (SE) and parameter variation (i.e., individual differences, ID), but not model variation (different models, DM). Shown in each cell in Table 2 is the mean fit and the percentage of time the particular model yielded the best fit. FLMP_F, the more flexible model because it has eight additional parameters, was always selected regardless of the source of the data, demonstrating that RMSD cannot discriminate between the types of variation. FLMP_F should have been selected in conditions in which it generated the data (SE_F and SE_F + ID, in rows 3a and 3b in Table 2, respectively), but it should not have been chosen when data were generated by the restricted model, FLMP_R. No matter whether data differed in sampling error only (SE_R, row 1a), or both sampling error and individual differences (SE_R + ID, row 1b), FLMP_F always provided the best fit. Finally, RMSD also failed to distinguish the models when the data were equally likely to come from either model (DM condition, row 2). This outcome would be similar regardless of the parameter set chosen.

These results show that flexibility measures such as RMSD can be highly error prone. Put simply, they get sidetracked by the quest to capture variability in the data, and end up selecting the most flexible model, which may not be the best approximation of the cognitive process.

This undesirable property of the flexibility approach to model selection is why we favor an alternative, one in which the selection methods do a better job of achieving the goal of model selection—to infer the form of the cognitive process. The approach that we advocate differs in the conceptualization of the problem, though in practice it can appear, and sometimes perform, quite similarly (the reasons for which will be explained later). In brief, model selection is viewed as a statistical inference problem, analogous to estimating population parameters from sample statistics or replicating (i.e., generalizing) the results in an experiment to a new sample of participants. It begins with

a recognition of the variability problem discussed above: Data are always contaminated by noise. To model a mental process accurately, one must devise a way to disentangle the sources of noise (sampling error, individual variability) from the variation due to the underlying cognitive process.

The problem of error is solved by shifting the focus of model selection from measuring only the precision with which a model can fit a given sample of data to estimating a model’s ability to fit unseen, future data sets that are representative of that same cognitive process. That is, the goal of model selection is to choose the model that generalizes best across all samples, because the one that does has probably captured the cognitive process of interest, and not the random fluctuations (i.e., error) that any one sample will exhibit. This is the essence of generalizability, and a model should be judged on its ability to generalize correctly, not on its adeptness (i.e., flexibility) in fitting only the data in hand.

To measure a model’s generalizability, the selection method must be sensitive to properties of the model in addition to GOF. Collectively they can be thought of as defining the *complexity* of the model, which among other things includes the number of parameters in the model and the way the parameters are combined in the model equation (i.e., functional form). RMSD and other flexibility measures are insensitive to model complexity. For a more detailed discussion of complexity and various measures of generalizability and its theoretical foundations, see Linhart and Zucchini (1986); Myung, Balasubramanian, and Pitt (2000); and Pitt, Myung, and Zhang (2002). One of these measures, cross validation, is mentioned in passing by Massaro et al. (2001, p. 15), but they expressed unease

Table 3
Comparison of Two Models Differing in Complexity Using Two Selection Methods: Goodness of Fit and Generalizability

Selection Method	Model Fitted	
	FLMP _R (true model)	FLMP _F
Goodness of fit	0.062 (0%)	0.054 (100%)
Generalizability	0.078 (81%)	0.084 (19%)

Note—Sample size = 20.

with it and with a central concept of generalizability: prediction. It is used in the following simulation to demonstrate why generalizability, not flexibility, should be the goal of model selection.

In cross-validation, a model's generalizability is assessed by fitting the model to one sample of data, holding those best fitting parameters constant, and then measuring the model's fit to another sample of data generated by the same cognitive process. Parameter retuning on the second data set is not allowed. If the model perfectly captures the underlying process and there is no noise (i.e., sampling error) in the data, the two fits will be equal. Most often, the second fit is worse precisely because the data are noisy. The amount by which it is worse provides an estimate of how much the model overfitted the first data set, capturing the spurious fluctuations due to sampling error.

The two versions of FLMP used in the preceding simulation were again compared. One hundred data sets were generated by FLMP_R and fitted by both models. Mean RMSDs of the fits of each model to the data are shown in the first row of Table 3. This test is equivalent to the SE_R condition in Table 2, and the results are identical. FLMP_F fitted the data best 100% of the time. Because FLMP_R generated the data (i.e., it is the true model), it should have yielded the best fit at least some of the time. The only way FLMP_F could have always provided a better fit is by fitting the variation due to sampling error.

That FLMP_F's superior fit is due to overfitting the data can be seen in the second row of Table 3, where the generalizability of the models was assessed by fitting them to a second set of 100 samples, also generated by FLMP_R. Mean fit was worse for both models, but the change in fit for FLMP_F (0.30) was double that for FLMP_R (0.16), indicating that FLMP_F absorbed twice as much sampling variation as FLMP_R in fitting the first 100 samples. As a result, FLMP_F not only yielded the poorest mean fit, but also provided the best fit least often. FLMP_F's overfitting must be caused by its eight extra parameters because its functional form is identical to FLMP_R's.

This simulation illustrates that the model that provides the best fit may not generalize the best. From the standpoint of generalizability, a suitable method of model selection must balance these opposing goals. On the one

hand, the model must provide a sufficiently good fit to the data to capture the underlying process. On the other hand, the fit must not be so good as to sacrifice generalizability. Measures of GOF, such as RMSD, concern themselves only with the first goal, whereas measures of generalizability try to satisfy both.

It is from the perspective of generalizability that BMS was suggested as a measure of model selection in cognitive psychology (Myung & Pitt, 1997). It is a good measure of generalizability precisely because it is sensitive to the many aspects of model complexity in addition to measuring GOF. RMSD is a poor measure precisely because it is insensitive to model complexity. Massaro et al.'s (2001) wariness of one feature of BMS, what they term "parameter invariance," highlights a fundamental difference in the flexibility and generalizability approaches to model selection. In BMS, a model is selected that does the best job of fitting a set of data across the full range of parameter values. Fits should be averaged if one is interested in the generalizability of the model beyond the data in hand. Sampling error will cause variability in future data sets, which the model, with parameters fixed, should still fit well if it does a good job of capturing only the underlying process, and not also the ever-present noise. Within the flexibility framework, such averaging over the parameter range is counterproductive since it unfairly penalizes models that possess the flexibility necessary to capture all of the variation in the data, regardless of its source (i.e., cognitive process and sampling error).

The superiority of generalizability over flexibility is easily shown by rerunning the simulation in Table 2, but substituting BMS for RMSD. As Table 4 shows, model recovery is virtually perfect. The selection method is sensitive to model differences, but not sampling error or individual differences. When variation due to sampling error (SE_R, SE_F) or both sampling error and individual differences (SE_R + ID, SE_F + ID) were present in the data, they were ignored, and the model that generated the data was almost always chosen. When the data came from each model half of the time (DM), BMS was also able to determine which model generated the data, selecting the correct model almost all of the time.

The data across Tables 2–4 might give the impression that BMS is superior to RMSD, which is not what Massaro

Table 4
Mean Model Recovery Rates and Percentage Wins Across Five Conditions Using BMS

Data Source	FLMP _R		FLMP _F		Model Fitted	
	θ_a	θ_b	θ_c	θ_d	FLMP _R ($k = 8$)	FLMP _F ($k = 16$)
1. Reduced model						
(a) Sampling error (SE _R)	100	–	–	–	–41.85 (99%)	–46.41 (1%)
(b) Sampling error + individual diffs (SE _R + ID)	50	50	–	–	–42.44 (100%)	–46.99 (0%)
2. Different models (DM)	–	50	50	–	–51.02 (50%)	–46.96 (50%)
3. Full model						
(a) Sampling error (SE _F)	–	–	–	100	–53.67 (2%)	–49.98 (98%)
(b) Sampling error + individual diffs (SE _F + ID)	–	–	50	50	–55.85 (1%)	–46.58 (99%)

Note—Sample size = 20. θ_{a-d} refer to the parameter sets used in the simulations. FLMP_F is the first version of FLMP listed in Table 1. FLMP_R is the reduced version, with θ_{4-7} and λ_{4-7} held constant at .2, .6, .7, and .9, respectively. BMS equals minus the log of the probability of the data, which is why the values are negative.

Table 5
Mean Model Recovery Rates and Percentage Wins for FLMP and WTAV
Using RMSD and BMS at Seven Noise Levels

Noise (SD)	Model Fitted	Selection Method				% Variance Accounted for	
		RMSD		BMS		Data Source (error-free)	
		Data Source		Data Source		FLMP	WTAV
		FLMP	WTAV	FLMP	WTAV		
0.00	FLMP	0.0367	0.0864	-35.9	-41.9	99.99	94.67
	% win	99	1	96	3		
	WTAV	0.1085	0.0539	-41.0	-37.5	93.39	99.99
	% win	1	99	4	97		
0.05	FLMP	0.0524	0.0920	-38.1	-43.4	99.93	95.12
	% win	98	2	95	4		
	WTAV	0.1127	0.0650	-43.1	-38.1	93.43	99.98
	% win	2	98	5	96		
0.10	FLMP	0.0767	0.1071	-41.9	-46.0	99.69	95.45
	% win	96	6	89	6		
	WTAV	0.1233	0.0881	-45.6	-40.4	93.49	99.94
	% win	4	94	11	94		
0.15	FLMP	0.1021	0.1267	-44.8	-49.3	99.25	95.53
	% win	94	13	78	11		
	WTAV	0.1384	0.1137	-48.9	-43.2	93.51	99.90
	% win	6	87	22	89		
0.20	FLMP	0.1273	0.1478	-48.9	-52.6	98.66	95.39
	% win	88	22	78	11		
	WTAV	0.1558	0.1390	-52.8	-46.5	93.54	99.86
	% win	12	78	22	89		
0.40	FLMP	0.2150	0.2278	-66.7	-70.1	95.86	93.87
	% win	78	47	62	12		
	WTAV	0.2292	0.2281	-70.9	-63.0	93.73	99.90
	% win	22	53	38	88		
0.80	FLMP	0.3089	0.3141	-91.9	-95.2	93.16	91.92
	% win	66	42	61	12		
	WTAV	0.3173	0.3192	-96.9	-86.4	94.02	99.93
	% win	34	59	39	88		

Note—The last two columns contain the mean percent variance accounted for when the specified model was fitted to the data without sampling error. Values along the diagonal should be 100%.

et al. (2001) always found: RMSD performed as well as BMS in at least half of their simulations. In a few instances, BMS performed quite poorly. What is the cause of these seemingly discrepant outcomes? In the remainder of this paper, we show that the discrepancies are more apparent than real, and that BMS's "failures" are readily explainable by a consideration of the details of the simulations and the limitations of all selection methods. We focus first on the three simulations in which BMS underperformed RMSD. These results serve as a backdrop for understanding why both selection methods performed so similarly in the other simulations.

REEXAMINATION OF MASSARO ET AL.'S (2001) SIMULATIONS

The two selection methods are defined as follows:

$$BMS = \ln \int (D|\theta) \pi(\theta) d\theta$$

$$RMSD = \sqrt{\frac{\sum_{i=1}^N (prd_i - obs_i)^2}{N}}$$

where $f(D|\theta)$ is the likelihood function of observed data D , θ is the parameter vector, $\pi(\theta)$ is the prior density of θ , \ln denotes the logarithm of base 10, prd_i and obs_i denote predicted and observed data, respectively, and finally, N is the number of data points being fitted. BMS originated in Bayesian statistics and represents the posterior probability of the model being correct given the data. It takes the logarithm of the mean likelihood, averaged across the full range of parameter values and weighted by the prior density. The model that maximizes BMS should be selected. On the other hand, RMSD represents the square root of the average deviation between predicted and observed data. This selection method prescribes that the model that minimizes RMSD should be preferred.²

The two models that Massaro et al. (2001) compared, FLMP and WTAV, differ in two ways (see Table 1). WTAV has one more parameter (the weight parameter, w) than FLMP. On the basis of the preceding simulations, in which the model with the most parameters provided the best fit to the data, one might think WTAV should outfit FLMP. However, as with most models in cognitive psychology, WTAV and FLMP also differ in their functional forms (i.e., model equations). Because such differences can have nontrivial influ-

ences on model behavior (Myung & Pitt, 1997; Pitt et al., 2002), it is far from obvious which model will perform best.

Evaluating the Selection Methods by Adding Noise to the Parameters

In their Table 4, Massaro et al. (2001) compared RMSD and BMS on their ability to recover the correct model (FLMP or WTAV) when Gaussian-distributed error and sampling error were added to the population proportions (P_{ij} in Table 1). The Gaussian error distribution had one of seven standard deviations spanning a range from 0.00 to 0.80. Performance of the two selection methods was evaluated by measuring how well each could identify which of the two models (FLMP or WTAV) generated a sample of data. The sample was also generated by one of these two models, so the test assessed how well RMSD and BMS "recovered" the data-generating model. Higher accuracy generally indicates better recovery (i.e., ability to discriminate between the models).

Model recovery for both selection methods was measured at each noise level. Their data are reproduced in the first four columns of Table 5 of the present paper. As one might expect, performance of the two selection methods declined as the standard deviation of the noise increased. Massaro et al. (2001) were particularly concerned about the asymmetry that emerged with BMS at the higher noise levels. When the data were generated by WTAV, recovery of the correct model (WTAV) leveled off at 88%, misattributing the data as belonging to FLMP only 12% of the time. When the data were generated by FLMP, correct recovery leveled off at a much lower value, 61%. Massaro et al. conjectured that this asymmetry is due to an inherent bias in the selection method of favoring less complex models. In fact, it is due to an anomaly in the simulation.

The addition of Gaussian-distributed error to P_{ij} necessitated truncating the error distribution at 0 and 1 so that simulated response probabilities would stay within the range of 0 and 1. An unwanted side effect of truncating the error distribution in this way is that the distribution itself becomes asymmetric. For example, when simulated probabilities are generated by introducing truncated Gaussian errors around a given response probability, say .80, values above 1 will be obtained more often than values below 0, so truncated probabilities will include more 1s than 0s. As a result, the mean of these probabilities will not in general be the same as the original probability, unless the original value is exactly equal to .5. Consequently, the Gaussian truncation procedure resulted in the creation of distorted response patterns that are no longer FLMP response patterns. As such, FLMP could not have provided a perfect fit to the data even in the absence of error. A prerequisite for evaluating model recovery is that the model recover its own data perfectly in this situation. Otherwise, the results of the model recovery simulation will not accurately reflect the performance of the model selection method.

Evidence that demonstrates the distorting effect of error truncation is provided in the last two columns of Table 5. At each noise level, FLMP and WTAV were fitted to the

response patterns generated by each model when Gaussian error and sampling error should have been zero. The expectation of each observed proportion (i.e., response value) generated by each model was estimated by sampling each P_{ij} and the two noise distributions 10,000 times and then averaging the resulting values. If an error distribution is symmetrical, the averaged value will be zero. Asymmetrical distributions will add a nonzero value, and thus distort P_{ij} . If both error distributions were symmetrical, the effects of error would have been eliminated and each model should have fitted its own data perfectly, which would be reflected in values of 100% in the two cells along the diagonal at each noise level. Scanning across the seven noise levels, one can see that FLMP was affected much more by noise truncation than WTAV. Even at the .80 noise level, WTAV absorbed 99.93% of the variance. In contrast, FLMP provided an increasingly poor fit as noise level increased, until by the .80 noise level, the model absorbed only 93.16% of the variance, which is even less than that for WTAV (94.02%). As noise level increased, the data-generating FLMP model became less and less FLMP-like, whereas WTAV changed little.

It is this asymmetry in the effects of noise truncation on the two data-generating models that caused BMS to discriminate between the two models less well when the data were generated by FLMP rather than WTAV. The parallel between the performance of BMS and the FLMP-likeness of the data-generating model is striking. When the data were generated by FLMP, correct model recovery under BMS dropped the most between the .20 and .40 noise levels, from 78% to 62%. Between these same noise levels is where the largest change in the data-generating model is found. Percent variance accounted for dropped 2.80% (from 98.66% to 95.86%). Prior to this point, the largest drop was 0.59%.

Additional evidence that demonstrates that truncation of the noise distribution was the cause of the asymmetry in how BMS performed is shown in Table 6. The simulations in Table 5 were rerun using binomially distributed noise, which not only eliminates the truncation problem, but also is a more appropriate method of simulating sampling variation given the experimental design (two response choices). When binomially distributed noise is used, noise level cannot be varied by changing the standard deviation of the distribution. Instead, the sample size must be varied, with smaller sample sizes being the equivalent of more noise (i.e., there is less information about the identity of the model underlying the data). Model recovery using BMS (right pair of columns in Table 6) shows no such bias in favoring the simpler model (i.e., WTAV), as Massaro et al. (2001) argued, performing exceptionally well all the way down to a sample size of 10. Only with a sample size of 5 does BMS begin to have difficulty, and it is just the opposite of what Massaro et al. hypothesized: BMS recovered the correct model more often when FLMP generated the data than when WTAV generated the data. Model recovery using RMSD (left pair of columns in Table 6) performed as it did in Table 5, exhibiting a consistent bias to favor FLMP as noise level increased.

Table 6
Mean Model Recovery Rates and Percentage Wins for FLMP
and WTAV Using RMSD and BMS at Five Sample Sizes

Sample Size	Model Fitted	Selection Method			
		RMSD		BMS	
		Data Source		Data Source	
		FLMP	WTAV	FLMP	WTAV
50	FLMP	0.0275	0.0829	-75.60	-139.56
	% win	100	0	99	0
	WTAV	0.1036	0.0453	-141.07	-92.93
	% win	0	100	1	100
24	FLMP	0.0372	0.0917	-61.23	-96.63
	% win	99	0	100	0
	WTAV	0.1075	0.0628	-92.45	-76.74
	% win	1	100	0	100
15	FLMP	0.0452	0.0995	-51.29	-76.85
	% win	99	13	100	1
	WTAV	0.1129	0.0820	-71.31	-66.73
	% win	1	87	0	99
10	FLMP	0.0576	0.1147	-45.44	-64.73
	% win	96	21	98	5
	WTAV	0.1231	0.1000	-59.03	-58.55
	% win	4	79	2	95
5	FLMP	0.0741	0.1385	-33.27	-46.09
	% win	96	49	98	26
	WTAV	0.1390	0.1407	-40.78	-44.57
	% win	4	51	2	74

Replication and Extension of Myung and Pitt (1997)

In Tables 1 and 2 of their paper, Massaro et al. (2001) extended a model recovery simulation carried out by Myung and Pitt (1997) in which BMS and RMSD were compared across three parameter sets using two models, FLMP and LIM. Massaro et al. included two additional parameter sets and the WTAV model. The results of Myung and Pitt were replicated. Across the nine conditions, BMS always recovered the correct model most often. RMSD, in contrast, recovered the correct model in only three of the nine conditions (always and only when FLMP generated the data). When LIM generated the data, RMSD's lack of sensitivity to model complexity caused it to choose FLMP most often. When WTAV generated the data, FLMP also won out, but twice by very little (~4%).

These same biases emerged with RMSD when model recovery was tested using the two new data sets: The cor-

rect model was chosen only when FLMP generated the data. BMS performed better, recovering the correct model in four of the six cases. In two cases, however, BMS chose the wrong model most often. One was with their Parameter Set 4. BMS's performance was absolutely appalling, choosing WTAV instead of FLMP 92% of the time when FLMP generated the data. With Parameter Set 5, BMS performed more like RMSD, choosing FLMP when WTAV actually generated the data, but only by a slight margin (4%).

The results with Parameter Set 5 are in fact not a failure of BMS, but rather the numerical integration method. Massaro et al. (2001) used the simple Monte Carlo (SMC) method to estimate the marginal likelihood for BMS. This method can require a very large number of iterations to settle on a solution. Massaro et al. used a cutoff of 1 million. Even this can be insufficient when the models are difficult to distinguish, causing the solution to be an under- or overestimate of the true value. In these (and

Table 7
Mean Marginal Log Likelihood Estimates
and Percentage Wins for Three Models

Model Fitted	Data Source			MCMC Replication
	FLMP	LIM	WTAV	WTAV
FLMP	-12.64	-15.53	-15.85	-16.52
%win	94	12	52	11
LIM	-21.93	-15.28	-17.74	
%win	0	88	0	
WTAV	-16.01	-16.25	-15.98	-15.88
%win	6	0	48	89

Note—The first three columns contain the data from Table 2 of Massaro et al. (2001), in which Parameter Set 5 was used. The last column is a replication of the simulation using MCMC when the data were generated by WTAV.

most) situations, Markov chain Monte Carlo (MCMC) methods (Gilks, Richardson, & Spiegelhalter, 1996) will yield more accurate results.³

When MCMC was used in place of SMC, BMS performed as expected. Shown in the first three columns of Table 7 are the BMS model recovery results of Massaro et al. (2001) using Parameter Set 5. BMS's failure can be seen in the third column, where WTAV data were thought to have been generated by FLMP slightly more often than WTAV (52% vs. 48%). The last column in Table 7 contains a replication of these simulations using MCMC. Mean marginal likelihood values reversed between FLMP and WTAV, and the correct model was chosen 89% of the time.

The abysmal failure of BMS to choose the correct model when Parameter Set 4 was used is not due to a convergence problem with the numerical integration method. Rather, as the following simulation shows, it was due to a parameter set that yielded FLMP response patterns that closely resembled WTAV response patterns. In a sense, the data tricked BMS into choosing the wrong model.

A massive simulation was undertaken in which the entire parameter space was sampled to identify FLMP response patterns that are very WTAV-like. The same 2×8 design of the preceding simulations was used. Ten thousand sets of parameters were generated (10 parameters for FLMP plus a weight parameter for WTAV) using a uniform density on $[0,1]$. Each of these parameter sets was then run through both models to create 10,000 FLMP response patterns and 10,000 WTAV response patterns. Because we were interested in identifying FLMP response patterns that resemble WTAV response patterns, WTAV was fitted to each of the 10,000 FLMP response patterns using RMSD as the measure of fit. Next, all parameter sets were sorted as a function of this fit, with those yielding the smallest RMSD (i.e., best fit) at the top of the list, and the worst fit at the bottom, essentially ranking the parameter sets by the confusability of the response pattern. Those at the top of the list yielded FLMP response patterns that were most like WTAV response patterns; those at the bottom of the list yielded response patterns that were more uniquely FLMP-like.

The rank-ordered parameter sets were then divided into six bins to reveal the primary trends in the data. One hundred parameter sets were randomly selected from each bin to carry out a couple of analyses. In the first analysis, all 10,000 parameter sets were compared with Parameter Set 4 of Massaro et al. (2001) using RMSD to find the closest match. Not surprisingly, the most similar parameter set is one of the most confusable, coming from Bin 2. It and another similar parameter set are listed in Appendix A along with Parameter Set 4. Also listed are two additional parameter sets that caused BMS to recover the wrong model. What is common among them is that they yield virtually parallel response functions, a characteristic of WTAV, not FLMP, which is known for its football-shaped response patterns (Massaro, 1998).

Shown in the upper graph in Figure 1 are the FLMP response functions using Parameter Set 4 and the FLMP response functions using one of the most similar param-

eter sets. The similarity of the functions in shape and actual values suggests that Parameter Set 4 produced confusable data and would most likely have belonged to Bin 1 or 2 were it one of the 10,000 generated in the simulation. WTAV accounted for 99.81% of the variance when fitted to data generated by Parameter Set 4 and 99.92% when fitted to the other parameter set. BMS's failure in this instance is therefore not surprising. In fact, the response pattern is so much like WTAV, one might well wonder whether BMS's performance should be considered a failure at all: There is every reason to believe the data were generated by WTAV. Given how WTAV-like the FLMP response pattern is, one might also wonder why RMSD performed so well on these highly confusable response patterns, recovering the correct model (FLMP) 77.5% of the time. (In contrast, BMS performed "correctly" only 8% of the time.) This question is answered below.

In the second analysis of the simulation data, the 100 parameter sets in each of the six bins were used in six model recovery simulations, as in Table 5. For each set of parameters, a single data sample (sample size of 20) was generated from each model, and RMSD and BMS were evaluated on their abilities to determine whether FLMP or WTAV generated the data. Shown in the lower graph in Figure 1 are the percentages of errors made by each selection method in each bin. Also shown are the percentages of errors in which both methods failed on the same parameter set (i.e., the intersection of RMSD and BMS errors). Mean WTAV fit (percent variance accounted for) to the 100 FLMP response patterns in each bin is shown on the y-axis for reference.

The data provide additional insight into the similarities and differences of the two selection methods, and the conditions in which they are likely to fail and succeed. Bin 1 contained the 500 parameter sets that yielded the most WTAV-like FLMP response patterns. As can be seen, both selection methods fared quite poorly. The error rate for BMS was 55% and that for RMSD was 80%. As the parameter sets yielded more distinguishable response patterns, errors decreased for both selection methods, but much more so for BMS than RMSD. The divergence of the two selection methods is most apparent in Bin 3, where RMSD dropped from 80% to 60% errors and BMS dropped a proportionately much larger amount (from 45% to 18%). Error rates in Bins 4–6 did not fluctuate greatly from the values in Bin 3.

These data clearly demonstrate how much more error-prone RMSD can be than BMS. The differences are quite large when averaged over bins. RMSD failed 65.0% of the time (weighted averaged) and BMS failed 18.9% of the time. Both methods failed on the same parameter sets in 15.7% of the tests (hashed bars). Comparison of the bars across bins shows that the hashed bar is only slightly below the BMS bar, indicating that there were very few parameter sets in which BMS failed and RMSD did not (3.2%). Looked at another way, if BMS failed, RMSD was almost certain to fail. The reverse, however, was not true.

The data in Figure 1 lump together the different types of errors that a selection method made. These data were

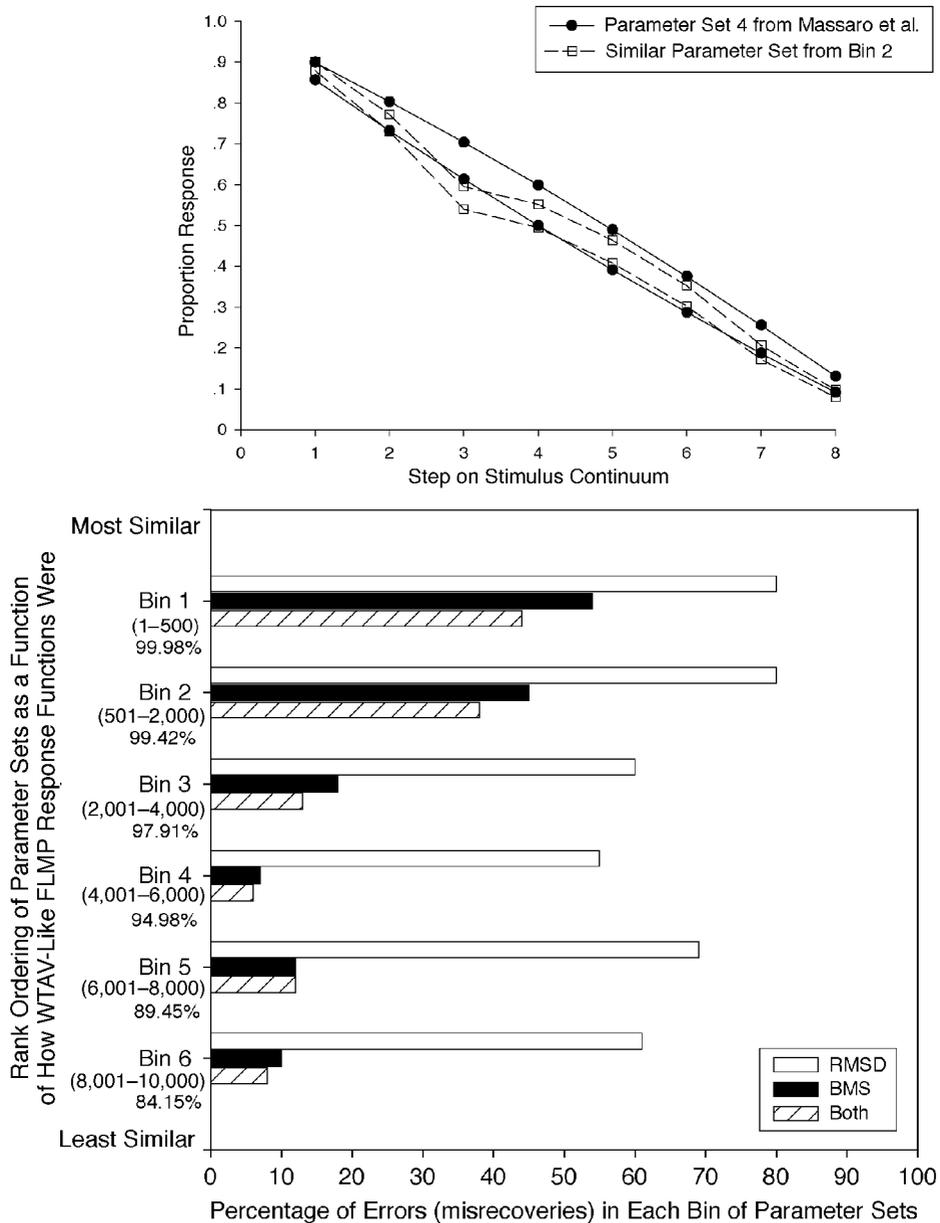


Figure 1. Top graph is a plot of two FLMP response functions in a 2×8 experimental design. One set of functions was produced using Parameter Set 4 from Table 2 of Massaro et al. (2001). The other was produced using the parameter set from the current simulation that most closely matched Parameter Set 4 (see Appendix B for parameter values). In the bottom graph, the percentage of misrecoveries by RMSD, BMS, and those common to both are plotted for each of the six bins of parameter sets. The percentages along the y-axis are the mean percent variance accounted for when WTAV was fitted to 100 FLMP response patterns in each bin.

therefore broken down to reveal the types of errors, and are shown in Table 8 in the familiar format used in preceding tables. RMSD exhibited its typical error pattern: When WTAV generated the data, there was a strong bias to choose FLMP, the more complex model. That this bias remained stable across bins reminds us that the bias is an inherent property of the selection method, having nothing to do with the parameter set or the discriminability of the

resulting response patterns. When FLMP generated the data, RMSD performed surprisingly well in Bin 1 (71% accuracy) given how WTAV-like the response patterns were. Performance continued to improve across bins.

When BMS was the selection method, model recovery was good and consistent across all bins when WTAV generated the data, selecting the correct model 90% or more of the time. When FLMP generated the data, BMS per-

Table 8
Mean Model Recovery Performance of RMSD and
BMS in Each of the Six Bins

Rank-Ordered Parameter Set	Model Fitted	RMSD		BMS	
		Data Source		Data Source	
		FLMP	WTAV	FLMP	WTAV
Bin 1 (1–500)	FLMP	0.0541	0.0647	–15.60	–18.31
	% win	71	66	50	5
	WTAV	0.0564	0.0653	–15.84	–16.64
	% win	29	34	50	95
Bin 2 (501–2,000)	FLMP	0.0530	0.0630	–15.40	–18.27
	% win	66	69	60	5
	WTAV	0.0568	0.0642	–15.81	–16.57
	% win	34	31	40	95
Bin 3 (2,001–4,000)	FLMP	0.0482	0.0627	–15.08	–18.36
	% win	85	52	85	3
	WTAV	0.0652	0.0632	–16.41	–16.59
	% win	15	48	15	97
Bin 4 (4001–6,000)	FLMP	0.0440	0.0615	–14.91	–17.96
	% win	100	55	97	4
	WTAV	0.0795	0.0645	–17.44	–16.20
	% win	0	45	3	96
Bin 5 (6,001–8,000)	FLMP	0.0335	0.0577	–13.88	–17.37
	% win	100	69	98	10
	WTAV	0.0967	0.0611	–18.51	–16.14
	% win	0	31	2	90
Bin 6 (8,001–10,000)	FLMP	0.0228	0.0562	–12.98	–17.11
	% win	100	61	100	10
	WTAV	0.1390	0.0598	–21.91	–16.04
	% win	0	39	0	90

Note—Parameter sets were rank-ordered as a function of how WTAV-like the response function was that FLMP produced using each parameter set.

formed as expected given the rank ordering of the parameter sets. Model recovery was at chance in Bin 1, which is where it should be when FLMP response patterns are highly similar to WTAV response patterns. As the parameter sets yielded response patterns that were more typical of FLMP (Bins 2–6), recovery quickly improved, becoming virtually perfect by Bin 4.

In Bin 1, RMSD performed unusually well when FLMP generated the data, better than BMS (71% vs. 50%). Given that this outcome is found in what is surely the most difficult condition in the simulation (i.e., where the response patterns of the two models were most confusable), it can give the impression that RMSD will perform best when it really counts, and seems at odds with the high error rates in Figure 1. However, in the evaluation of RMSD's performance in this condition, or in any model recovery test, it is important to keep in mind that performance cannot be attributed solely to the model recovery abilities of the selection method. Rather, performance is partially due to the method's inherent bias in favoring the more complex model, which just happens to be FLMP in this case. That is, RMSD's bias contributes to its successes as well as its failures. The bias is omnipresent. It inflates performance when the bias works in favor of the more complex model (i.e., FLMP), but in this situation it masquerades as accurate recovery, making the bias invisible. This gives the appearance of RMSD performing better than it really does. The magnitude of the bias is visible when it works

against the less complex model (i.e., WTAV). This condition is a more accurate reflection of the method's inability to recover the correct model.

The severity of RMSD's bias is visible in the adjacent column in Table 8, where WTAV generated the data. FLMP was incorrectly chosen as the data-generating model two thirds of the time, and this bias is evident over all bins. If the 71% recovery rate in Bin 1 were due solely to RMSD's good ability to discriminate between the models when the response patterns were highly confusable, then recovery in other cells in Table 8 should have been equally good, if not perfect. That they are not clearly indicates that the 71% recovery rate in Bin 1 is an artifact of bias.

Another interpretation of the results in Bin 1 is that the 50% correct recovery rate by BMS when FLMP generated the data is to be expected under the circumstances. The most WTAV-like FLMP response patterns were present in this bin. Recovery should have been near chance because FLMP data closely mimicked WTAV data. The correct model was unrecoverable because the data provided little or no information about the true identity of the data-generating model. As one would expect, this situation quickly corrected itself across bins as the FLMP data became less WTAV-like.

The data in Figure 1 and Table 8 together should also dispel beliefs about the omnipotence of any model selection method. No method, no matter how good it is, should perform well when confronted with data from Bins 1 and 2.

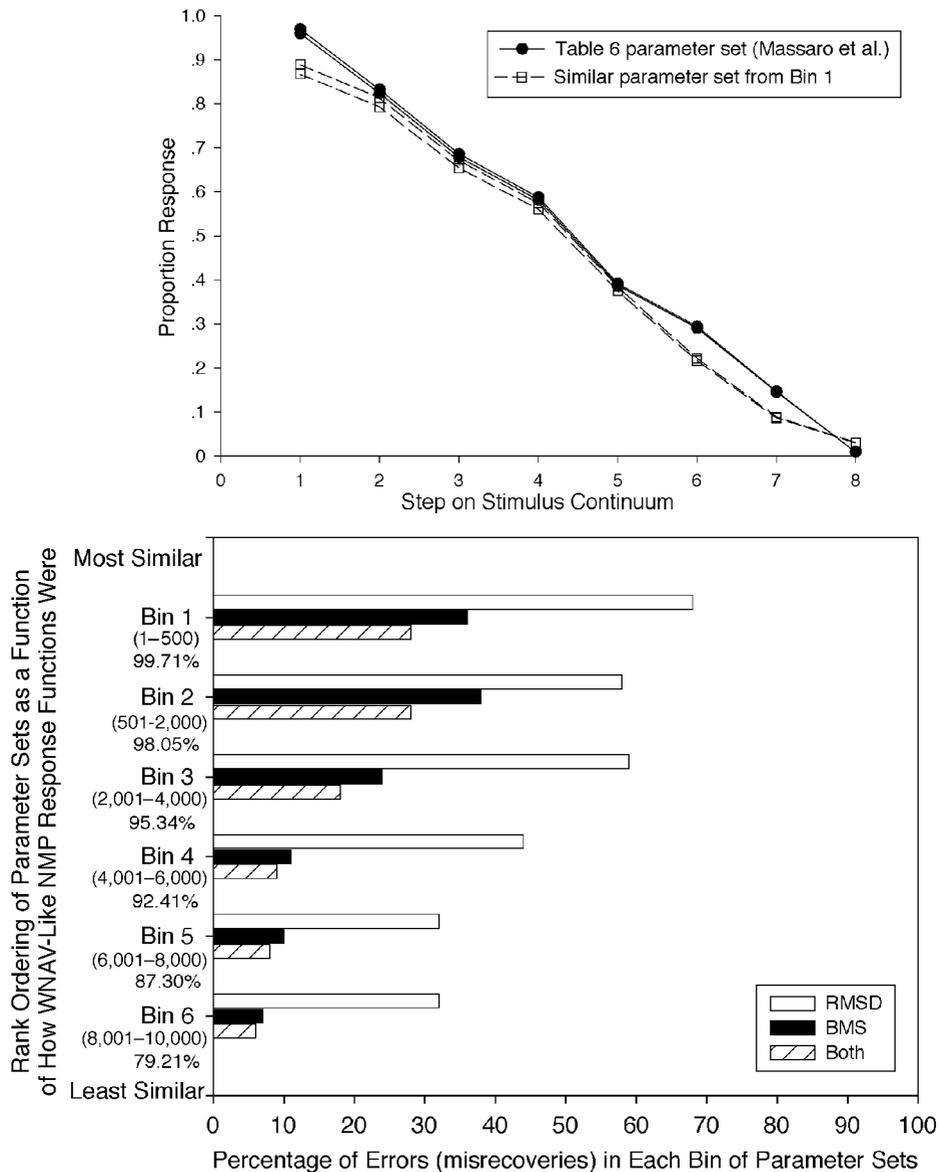


Figure 2. Top graph is a plot of two NMP response functions in a 2×8 experimental design. One set of functions was produced using the parameter set Massaro et al. (2001) used in the simulation in their Table 6. The other was produced using the parameter set from the current simulation that most closely matched theirs. In the bottom graph, the percentage of misrecoveries by RMSD, BMS, and those common to both are plotted for each of the six bins of parameter sets. The percentages along the y-axis are the mean percent variance accounted for when WNAV was fitted to the 100 NMP response patterns in each bin.

How could it when information is essentially misleading? In this regard, Massaro et al.'s (2001, p. 6) observation that selection method performance appears to be data dependent is right on the mark. One implication of the data-dependent nature of model selection is that data patterns can always be found that will cause a selection method to fail. Recovery results from one or two atypical parameter sets can be misleading, especially when they produce response patterns that are representative of the typical performance of the competing model. Only by sampling the

entire parameter space, as we did in this simulation, can a more accurate picture of the true recovery ability of a selection method be obtained. Of course, simulations like this are only suggestive of general tendencies. They are approximations, not proofs, of what is likely to be found. Looked at another way, knowledge of the parameter sets and response patterns is necessary to interpret model recovery results correctly. In a 2×2 matrix of model recovery data, values of 100% along the diagonals is not always the correct prediction. By this criterion, what ap-

Table 9
Reproduction of the Model Recovery Data in Table 6
of Massaro et al. (2001)

Model Fitted	Selection Method			
	RMSD		BMS	
	Data Source		Data Source	
	NMP	WNAV	NMP	WNAV
NMP	0.0517	0.0154	-23.19	-30.69
% win	100	90	22	0
WNAV	0.0764	0.0167	-13.81	-13.81
% win	0	10	78	100

pears to be a failure in model recovery can actually be reasonable selection behavior given the characteristics of the response patterns (e.g., when they are highly similar). Conversely, good model recovery in situations like this can serve as a red flag that the selection method is performing incorrectly, as was the case with RMSD.

A Test of Newton's Law

The third situation in which Massaro et al. (2001) found BMS misrecovered the correct model a high percentage of the time was when Newton's law (NMP) was compared with a weighted additive version of Newton's law (WNAV). The data from this simulation are reproduced in Table 9. RMSD again performed as expected, exhibiting a strong bias to select the more complex model (NMP in this case) when WNAV generated the data, but correctly chose NMP 100% of the time when it generated the data. BMS exhibited the opposite bias, choosing WNAV (78% of the time) instead of NMP when NMP generated the data, but correctly choosing WNAV when WNAV generated the data.

Massaro et al. (2001) again attributed BMS's failure to an inherent bias in the selection method to favor the less complex model. In fact, their simulation is another example that underscores the importance of knowing the confusability of the response patterns that a parameter set generates. When the simulations in the preceding section were rerun using these two models, a very similar outcome emerged. As before, 10,000 parameter sets were generated, sorted as a function of how WNAV-like the resulting NMP response patterns were, divided into six bins, and then analyzed as before.

The data are shown in the two graphs in Figure 2. In the upper graph, two NMP response patterns are plotted, one using the parameter set of Massaro et al. (2001) and the other the parameter set from the current simulation that most closely matched theirs (parameter values are listed in Appendix A). The best matching parameter set came from Bin 1. The similarity of the two response patterns confirms our suspicion that the parameter set Massaro et al. chose produced a particularly confusable (i.e., very WNAV-like) response pattern, which is why BMS chose WNAV so frequently (78%) when NMP actually generated the data. That RMSD produced *no* errors (NMP was chosen 100% of the time) with this parameter set is odd, and again indicates that RMSD's bias to favor complex

models had a significant influence on model selection. RMSD should have made errors precisely because the underlying model is not easily identifiable from the data. All selection methods should.

Overall model recovery performance of the two selection methods across bins is shown in the lower graph. Both made the most errors when the parameter sets came from Bins 1 and 2, where NMP response patterns were most WNAV-like. As confusability of these response patterns decreased, so did error rate. In each bin, including Bin 6, in which the parameter sets yielded the most discriminable data, RMSD made more errors than BMS. Overall, RMSD made an average of 46.5% errors. BMS made almost one third less (16.0%). As in Figure 1, the frequency with which both methods erred on the same parameter sets (10.9%) was slightly less than BMS's error rate, replicating the findings that if BMS failed to recover the correct model, RMSD almost certainly failed as well. In sum, there are actually very few parameter sets (5.1%) that led RMSD to outperform BMS.⁴

By knowing that the parameter set used in the model recovery test in Table 9 created an NMP response pattern that was very WNAV-like, expectations regarding model recovery when NMP generated the data should change accordingly. Like BMS, RMSD should have been misled and chosen WNAV as the data-generating model much of the time. Instead, RMSD exhibited a strong bias to choose the more complex model (NMP). The extent of this bias is visible when WNAV generated the data; RMSD incorrectly chose NMP 90% of the time. When the RMSD's two outcomes are considered together, its performance seems paradoxical: How can the selection method discriminate perfectly between two models given what are very confusable response patterns, and yet fail so miserably on another set of response patterns, which in all likelihood are not nearly as confusable?

These two large-scale simulations demonstrate that model recovery results must be interpreted with a thorough understanding of the models and the data on which recovery is evaluated. In Massaro et al.'s (2001) simulations, the errors that BMS made, if they can even be called that, stem from a limitation of the method in a very specific context: The parameter sets that were used yielded response patterns that mimicked the typical behavior of the competing model, causing misrecovery. All selection methods should perform poorly in this situation simply because there is not enough information in the data to distinguish between the models. That RMSD did not, and instead frequently chose the more complex model, is strong evidence that bias played a significant role in guiding model selection.

HOW TRUSTWORTHY IS THE SELECTION METHOD?

When RMSD and BMS Perform Similarly

The results in the preceding sections might lead one to think that BMS should always outperform RMSD, yet in

three of the model selection tests that Massaro et al. (2001) carried out (5×5 bimodal integration data, 8-alternative bimodal integration data, and Pitt, 1995, data), the two selection methods performed equivalently. Why? Part of the answer, as Massaro et al. discussed, is that FLMP and WTAV are better matched in complexity than, for example, FLMP and LIM. Whether these two selection methods will perform similarly will also be determined by the data. To understand why, consider the data in the lower graph in Figure 1 again. Just as there were many parameter sets that yielded response patterns on which one or both methods failed, there were also many parameter sets whose response patterns led to correct recovery by both methods. A rough estimate of these can be gleaned from the graph. In each bin, the space to the right of the bars (i.e., 100 minus the error rate) is an estimate of the percentage of response patterns that each selection method would recover correctly. Collapsed over bins, RMSD and BMS performed correctly on 45% and 79.1% of the response patterns, respectively. Although these estimates will change as a function of other factors (e.g., experimental design, sample size), they demonstrate that there is plenty of opportunity for RMSD and BMS to yield the same (correct) answer.

Because BMS takes into account model complexity and RMSD does not, the data in Figures 1 and 2 also provide information on how much model complexity influenced model selection in this experimental setup. The difference between the BMS and RMSD bars is a rough estimate of this value. If the models were equal in complexity, then the two bars in each bin should be equal in length because complexity would have had no effect on model recovery using RMSD. The RMSD bars will lengthen relative to the BMS bars as the complexity of one of the models increases. The difference between these two bars can therefore also provide an indication of how much more complex one model is than the other.

The frequency with which the two selection methods will perform similarly makes it reasonable to ask whether they are interchangeable. The similar performance of RMSD and BMS led Massaro et al. (2001) to conclude that RMSD is likely to yield accurate and reliable results in their testing situations. To the extent that such experiments advance the science, these testing situations will undergo both small and large changes, particularly when better models are introduced in place of ones shown to be inferior. An advantage of using BMS is that it will perform far more accurately across these situations, as the simulations in the next section demonstrate.

Generalizing Model Selection Performance

Differences in the reliability of the two selection methods is further demonstrated in Table 10, where BMS and RMSD were compared in their ability to recover different pairs of models across samples sizes. Shown in each cell of Table 10 is the mean of the given selection criterion across simulated samples and the percentage of time the particular model was selected under the selection method.

A comparison of FLMP and WTAV is shown in the middle of Table 10 for reference. Just as Massaro et al. (2001) found in some of their simulations, RMSD and BMS performed equivalently and they did so across sample sizes. This outcome can give the impression that the two selection methods are interchangeable, when in fact it is situation specific, as demonstrated by the data on the left and right sides of the table. On the left, WTAV was compared with FLMP_w, a geometrically weighted version of FLMP (see Table 1 for its model equation), making the models equal in the number of parameters (11). BMS outperformed RMSD, but only when the data were generated by WTAV. As sample size increased, RMSD's performance approached that of BMS's. The same result was found on the right side of the table, where FLMP was compared with LIM_T, a truncated version of LIM, defined in

Table 10
Mean Recovery Rate and Percentage Wins Comparing RMSD and BMS on Three Pairs of Models at Three Sample Sizes

Sample Size	Model Fitted	Selection Method				Model Fitted	Selection Method				Model Fitted	Selection Method				
		RMSD		BMS			RMSD		BMS			RMSD		BMS		
		Data Source	Data Source	Data Source	Data Source		Data Source	Data Source	Data Source	Data Source		Data Source	Data Source	Data Source	Data Source	
50	FLMP _w	0.025	0.0424	-40.67	-41.68	FLMP	0.0252	0.0793	-32.63	-55.58	FLMP	0.0252	0.0462	-32.63	-40.47	
	% win	100	18	98	6	% win	100	0	99	0	% win	100	6	100	0	
	WTAV	0.0899	0.0369	-54.01	-37.51	WTAV	0.1027	0.0369	-61.14	-37.51	LIM _T	0.0888	0.0361	-57	-32.24	
	% win	0	82	2	94	% win	0	100	1	100	% win	0	94	0	100	
	24	FLMP _w	0.036	0.0544	-30.44	-32.17	FLMP	0.0373	0.0865	-26.41	-38.32	FLMP	0.0373	0.0573	-26.41	-30.34
		% win	99	39	99	17	% win	100	2	100	1	% win	99	22	99	2
WTAV		0.0952	0.0533	-36.8	-30.65	WTAV	0.1079	0.0533	-40.04	-30.65	LIM _T	0.0955	0.0526	-36.77	-26.8	
% win		1	61	1	83	% win	0	98	0	99	% win	1	78	1	98	
12		FLMP _w	0.0518	0.0714	-23.51	-24.83	FLMP	0.0511	0.0967	-20.67	-27.42	FLMP	0.0511	0.073	-20.67	-23.3
		% win	100	61	100	32	% win	99	10	99	7	% win	100	52	97	10
	WTAV	0.1064	0.0747	-26.46	-24.38	WTAV	0.1183	0.0747	-27.94	-24.38	LIM _T	0.1061	0.0753	-25.34	-21.78	
	% win	0	39	0	68	% win	10	90	1	93	% win	0	48	3	90	

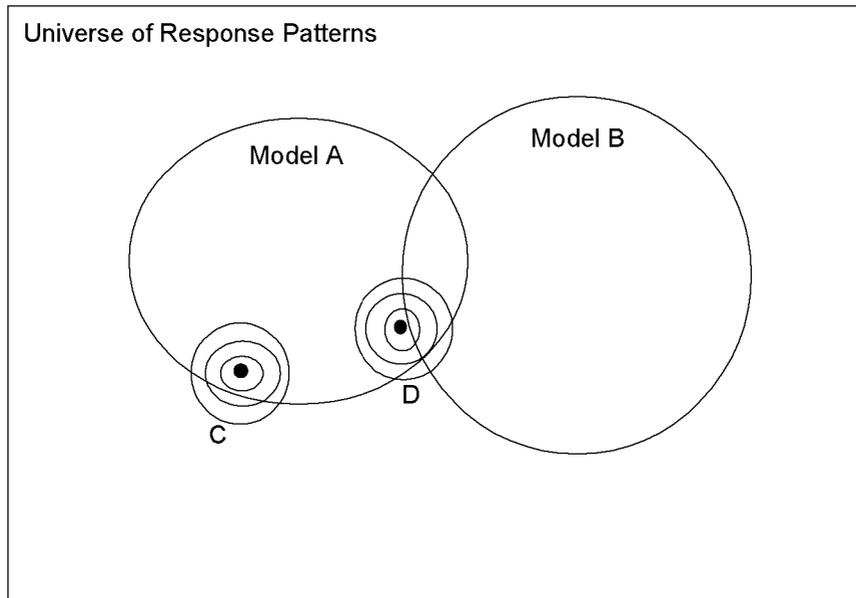


Figure 3. Illustration of how model recovery is influenced by data and sample size. Two models, A and B, occupy a region of the space of all possible response patterns (rectangle area). Two points represent different response patterns, with the concentric circles denoting the amount of sampling error in the data (larger rings indicate more error).

Table 1. RMSD misrecovered the correct model, in particular for Sample Size 12, when LIM_T generated the data. BMS performed much more robustly.

The left-hand and right-hand model recovery results are also clear demonstrations of why a model's functional form must be considered in model selection. Each pair of models had the same number of parameters, so the biases found in model recovery in favor of FLMP (see RMSD data for Sample Sizes 12 and 24) must be due to functional form.

The law of large numbers is responsible for the improvements in recovery across sample size. As it grows, sampling error diminishes, making it easier to discriminate between the models. Essentially, there is less error to confuse the selection methods. The fact that both methods perform similarly as sample size increases reminds us that BMS and RMSD are asymptotically equivalent: Both will perform identically given an infinitely large sample of data. This fact is one reason why the two theoretical approaches to model selection that were discussed in the first part of the paper, flexibility and generalizability, can yield the same answer. They differ primarily in how sampling error is treated, as potentially meaningful variation (flexibility) or as meaningless variation (generalizability). When sample size is large enough, sampling error becomes negligible to the point where its contribution to model selection can be ignored. There is no more noise for a flexible model to absorb erroneously, and when there is no more noise, models will generalize perfectly. In other words, as sample size increases, and data become more precise, complexity plays a smaller and smaller role in model selection. However, because small samples are the norm in experimentation, being constrained by a host of factors,

including experimental design and choice of stimuli (not to mention modelers being forced to use whatever data are available), the accuracy and trustworthiness of BMS make it the safer choice a priori.

Perhaps not surprisingly, sample size and data (i.e., response patterns) interact to influence model selection. We close this section with an example that illustrates this point. The rectangle in Figure 3 represents the space of all possible response patterns. Models A and B occupy a region in this space. When the experimental data fall in a region that is close to only one model (Point C), model selection should be successful regardless of whether there is a small or large amount of noise (concentric rings). When the data fall close to both models, it is much more difficult to discriminate the source of the data when error is large, but easier when error is smaller. Thus, model selection methods can be highly data dependent at small sample sizes because of the presence of noise. As sample size increases, this dependency diminishes (see Table 10). In the limit (i.e., asymptotically), model selection is data independent.

CONCLUSION

Selecting among models of cognition given a limited amount of data is a difficult problem. In psychology, it is particularly challenging because the mental process being studied is not directly observable and our only tie to it is noisy data. By making generalizability the goal, the problem of noise is mitigated and model selection becomes statistical inference based on fit and complexity. The superiority of this approach is demonstrated by the robustness of its selection methods, such as BMS (see Myung, Balasub-

ramanian, & Pitt, 2000; Pitt et al., 2002). The intuitiveness of generalizability makes it the approach of choice in other fields, such as computer science, statistics, and econometrics (Grunwald, in press; Rissanen, 1983; Vitanyi & Li, 2000; see also Hansen & Yu, 2001), where a related selection method, minimum description length, is proving quite valuable. We believe it should be preferred in psychology as well.

The simulations presented here, along with reanalyses of those carried out in Massaro et al. (2001), reveal severe limitations of RMSD and clearly demonstrate the superiority of BMS as a selection method. Nevertheless, we are not claiming that BMS is infallible or even that it is bias free. It may indeed be biased to select the simpler of two models (Grunwald, in press), but even if BMS is biased, its biases are far smaller than those of RMSD, as BMS's consistently good recovery performance demonstrates. In this regard, it is important to understand that BMS does not penalize the more complex model just because it is more complex. Rather, it weighs a model's complexity relative to what is needed to provide a good fit to the data. The more *excess* complexity a model has, the more it is penalized. If a model's complexity is justified by the data, then the complex model will be preferred over the simpler model (see Pitt et al., in press, for further discussion).

Despite the many shortcomings of RMSD that the present and prior simulations reveal, we do not advocate abandoning it. On the contrary, this work has been quite informative in identifying when and how RMSD can be safely used to guide model selection. In particular, RMSD will probably perform just as well as BMS when the models are similar in complexity. The fact that such similar performance was found with FLMP and WTAV in many simulations suggests that these models are closer in complexity than FLMP and LIM, as Massaro et al. (2001) suggested.

The knowledge gained from this collective body of work suggests a productive way in which to use RMSD in model selection. It can be used if it is first shown to perform well in a model recovery simulation, which will expose biases in the selection method in favoring one model over another. Within the generalizability framework, model selection is statistical inference based on limited data. As the outcomes in Tables 6 and 10 and Figures 1 and 2 show, the quality of the inference depends on the size of the sample and the shape of the response patterns being fitted. Before fitting participants' data, a model recovery simulation should be carried out in which similar response patterns are used and the sample size is the same as that in the actual experiment. If there is considerable variability across individuals, then the simulation should be repeated across a subset of representative response patterns.

If recovery is found to be good with RMSD, the models are probably close enough in complexity that this factor is likely to have a negligible impact on selection. When RMSD fails, then it is an indication that the models are sufficiently different in complexity to require the use of a selection method that takes into account this property of a model. Regardless of the outcome, the results of such a

simulation also serve as an indirect and crude measure of the relative complexities of the two models.

Given the superiority of BMS across testing situations, some readers might find the preceding recommendation odd (one reviewer did). Although BMS is a more robust and general method, technical sophistication is required to use it. In contrast, a model recovery simulation is a simple yet informative diagnostic tool for assessing the accuracy of RMSD given a particular data pattern. If used properly, it should do a good job of guarding against using RMSD in inappropriate situations.

REFERENCES

- GILKS, W. R., RICHARDSON, S., & SPIEGELHALTER, D. J. (1996). *Markov chain Monte Carlo in practice*. New York: Chapman & Hall.
- GRUNWALD, P. (in press). *Minimum description length and maximum probability*. Heidelberg: Springer-Verlag.
- HANSEN, M. H., & YU, B. (2001). Model selection and the principle of minimum description length. *Journal of the American Statistical Association*, **96**, 746-774.
- LINHART, H., & ZUCCHINI, W. (1986). *Model selection*. New York: Wiley.
- MASSARO, D. W. (1988). Some criticisms of connectionist models of human performance. *Journal of Memory & Language*, **27**, 213-234.
- MASSARO, D. W. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, MA: MIT Press.
- MASSARO, D. W., & COHEN, M. M. (1983). Phonological context in speech perception. *Perception & Psychophysics*, **34**, 338-348.
- MASSARO, D. W., COHEN, M. M., CAMPBELL, C. S., & RODRIGUEZ, T. (2001). Bayes factor of model selection validates FLMP. *Psychonomic Bulletin & Review*, **8**, 1-17.
- MYUNG, I. J., BALASUBRAMANIAN, V., & PITT, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, **97**, 11170-11175.
- 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.
- MYUNG, I. J., & PITT, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.
- PITT, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1037-1052.
- PITT, M. A., MYUNG, I. J., & ZHANG, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, **109**, 472-491.
- RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics*, **11**, 416-431.
- VITANYI, P., & LI, M. (2000). Minimum description length, Bayesianism and Kolmogorov complexity. *IEEE Transactions on Information Theory*, **46**, 446-464.

NOTES

1. One small caveat is that the results obtained using BMS hold if MCMC is used in place of simple Monte Carlo.
2. RMSD defined here and used in Massaro et al. (2001) differs from the RMSD of Myung and Pitt (1997), where it was defined as

$$\text{RMSD} = \sqrt{\sum (prd_i - obs_i)^2 / (N - k)}.$$

See also Massaro (1998).

3. All BMS results reported in the present paper were obtained using MCMC methods.

4. For completeness, the data in Figure 2 are presented in Appendix B, broken down by the types of errors made in each of the six bins. For the most part, the data add little to what was learned from Table 8, so they will not be discussed further.

APPENDIX A

Parameter Set 4 from Tables 1 and 2 of Massaro et al. (2001):

θ_1	θ_2	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
0.45	0.55	0.11	0.22	0.33	0.44	0.55	0.66	0.77	0.88

Four parameter sets that caused BMS to misrecover the correct model. The first two are the most similar to those used by Massaro et al.

0.3971	0.4523	0.1157	0.2402	0.397	0.5114	0.5982	0.6399	0.8038	0.9166
0.4416	0.4831	0.1451	0.3053	0.3731	0.4296	0.5549	0.7351	0.8252	0.8600
0.7986	0.8093	0.1904	0.2707	0.3175	0.5356	0.5755	0.5942	0.7419	0.7928
0.5133	0.6254	0.3599	0.4845	0.4943	0.6659	0.6704	0.7655	0.8063	0.8201

Parameter set from Table 6 of Massaro et al. (2001):

θ_1	θ_2	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
0.97	0.98	0.01	0.15	0.3	0.4	0.6	0.7	0.85	0.99

Four parameter sets that caused BMS to misrecover the correct model. The first two are the most similar to those used by Massaro et al.

0.9525	0.9642	0.1083	0.2472	0.3498	0.4403	0.5835	0.6462	0.8228	0.9651
0.9616	0.9858	0.03096	0.09027	0.2251	0.3899	0.5833	0.6806	0.8251	0.9022
0.7109	0.8290	0.4783	0.4840	0.5315	0.6371	0.6620	0.7617	0.8098	0.8921
0.8165	0.8278	0.2478	0.2513	0.2929	0.6194	0.7119	0.7276	0.8841	0.9251

APPENDIX B

Mean Model Recovery Performance of RMSD and BMS in Each of the Six Bins

Rank-Ordered Parameter Set	Model Fitted	Selection Method			
		RMSD Data Source		BMS Data Source	
		NMP	WNAV	NMP	WNAV
Bin 1 (1–500)	NMP	0.0538	0.0611	–14.70	–18.00
	% win	66	50	72	8
	WNAV	0.0565	0.0609	–15.41	–16.34
	% win	34	50	28	92
Bin 2 (501–2,000)	NMP	0.0509	0.0608	–14.36	–17.51
	% win	72	39	73	11
	WNAV	0.0559	0.0602	–15.26	–16.22
	% win	28	61	27	89
Bin 3 (2,001–4,000)	NMP	0.0477	0.0657	–14.08	–17.80
	% win	77	49	86	10
	WNAV	0.0583	0.0660	–15.43	–16.54
	% win	23	51	14	90
Bin 4 (4,001–6,000)	NMP	0.0485	0.0635	–14.34	–17.90
	% win	90	37	92	3
	WNAV	0.0686	0.0614	–16.18	–16.08
	% win	10	63	8	97
Bin 5 (6,001–8,000)	NMP	0.0442	0.0648	–14.00	–17.99
	% win	99	32	98	8
	WNAV	0.0852	0.0623	–17.02	–16.22
	% win	1	68	2	92
Bin 6 (8,001–10,000)	NMP	0.0315	0.0649	–13.02	–17.80
	% win	100	32	100	7
	WNAV	0.1210	0.0594	–19.82	–15.94
	% win	0	68	0	93

Note—Parameter sets were rank ordered as a function of how WNAV-like the response function was that NMP produced using each parameter set.