Psychonomic Bulletin & Review
2001, 8 (1), 1-17

Bayes factor of model selection validates FLMP

DOMINIC W. MASSARO, MICHAEL M. COHEN,
CHRISTOPHER S. CAMPBELL, and TONY RODRIGUEZ
University of California, Santa Cruz, California

The fuzzy logical model of perception (FLMP; Massaro, 1998) has been extremely successful at describing performance across a wide range of ecological domains as well as for a broad spectrum of individuals. An important issue is whether this descriptive ability is theoretically informative or whether it simply reflects the model's ability to describe a wider range of possible outcomes. Previous tests and contrasts of this model with others have been adjudicated on the basis of both a root mean square deviation (RMSD) for goodness-of-fit and an observed RMSD relative to a benchmark RMSD if the model was indeed correct. We extend the model evaluation by another technique called Bayes factor (Kass & Raftery, 1995; Myung & Pitt, 1997). The FLMP maintains its significant descriptive advantage with this new criterion. In a series of simulations, the RMSD also accurately recovers the correct model under actual experimental conditions. When additional variability was added to the results, the models continued to be recoverable. In addition to its descriptive accuracy, RMSD should not be ignored in model testing because it can be justified theoretically and provides a direct and meaningful index of goodness-of-fit. We also make the case for the necessity of free parameters in model testing. Finally, using Newton's law of universal gravitation as an analogy, we argue that it might not be valid to expect a model's fit to be invariant across the whole range of possible parameter values for the model. We advocate that model selection should be analogous to perceptual judgment, which is characterized by the optimal use of multiple sources of information (e.g., the FLMP). Conclusions about models should be based on several selection criteria.

The fuzzy logical model of perception (FLMP; Massaro, 1998) has consistently provided a good description of a variety of results in bimodal speech perception and in many other domains of human performance. The assumptions central to the model are (1) each source of information is evaluated to determine the degree to which that source specifies various alternatives, (2) the sources of information are evaluated independently of one another, (3) the sources are integrated to provide an overall degree of support for each alternative, and (4) perceptual identification and interpretation follows the relative degree of support among the alternatives. In a two-alternative task with /ba/ and /da/ alternatives, the degree of auditory support for /da/ can be represented by a_i, and the support for /ba/ by $(1 - a_i)$. Similarly, the degree of visual support for /da/ can be represented by v_j, and the support for /ba/ by $(1 - v_j)$. The probability of a response to the unimodal stimulus is simply equal to its feature value. For bimodal trials, the predicted probability of a response, $P(/\text{da}/)$ is equal to

$$P(/\text{da}/) = \frac{a_i v_j}{a_i v_j + (1-a_i)(1-v_j)}. \tag{1}$$

In the course of our research, we have found that the FLMP accurately describes human pattern recognition. For example, we have learned that people use many sources of information in perceiving and understanding speech, emotion, and other aspects of the environment. In many cases, these sources of information are ambiguous, and any particular source alone does not usually specify completely the appropriate interpretation. The influence of one modality is greater to the extent that the other is ambiguous, a result well described by the FLMP. The results from many studies are consistent with the FLMP, which describes a universal law of behavior (Massaro, 1998).

In our previous work, we have contrasted the FLMP against several alternative models such as a weighted averaging model (WTAV), which is an inefficient algorithm for combining the auditory and visual sources. For bimodal trials, the predicted probability of a response, $P(/\text{da}/)$ is equal to

$$P(/\text{da}/) = \frac{w_1 a_i + w_2 v_j}{w_1 + w_2} = w a_i + (1-w) v_j. \tag{2}$$

The WTAV predicts that two sources can never be more informative than one. In direct contrasts, the FLMP has consistently and significantly outperformed the WTAV

The research was supported by grants from the National Institute of Deafness and Other Communicative Disorders (PHS R01 DC00236), the National Science Foundation (Challenge Grant CDA-9726363), Intel Corporation, and the University of California Digital Media Innovation Program. D.W.M. is highly appreciative of the encouraging support of Dan Friedman and Bill Rowe. We thank William Batchelder, James Cutting, In Jae Myung, Mark Pitt, and John Wixted for their constructive comments on an earlier version of the paper. Correspondence should be addressed to D. W. Massaro, Department of Psychology, University of California, Santa Cruz, CA 95064 (e-mail: massaro@fuzzy.ucsc.edu).

(Massaro, 1998). Our criterion for model selection has been a goodness-of-fit measure called root mean square deviation (RMSD) between predicted and observed proportions. The RMSD is computed by (1) squaring the difference between each predicted (p) and observed (o) value (this makes all differences positive and also magnifies large deviations relative to small ones), (2) summing the squared differences across all n conditions, (3) taking the mean of these differences, and (4) taking the square root of this mean.

$$\sqrt{\frac{\sum (p-o)^2}{n}}. \tag{3}$$

The RMSD provides an easily understood measure of the agreement between the actual and the theoretical outcomes. An RMSD of .02 means simply that the model's predictions differ by roughly an average of .02 from the observations. Recently, this measure has been called into question given a resurgence of interest in model testing and selection from researchers in various domains of performance and also in the mathematical modeling community (Cutting, Bruno, Brady, & Moore, 1992; Dunn, 2000; Massaro, 1998; Massaro & Cohen, 1993; Myung, Forster, & Browne, 2000; Myung & Pitt, 1997, 1998).

Cutting et al. (1992) claimed that the good fit of the FLMP did not necessarily reflect a good psychological theory but rather some type of selectivity and scope (flexibility) to fit any type of data, even random results. Massaro and Cohen (1993) countered these claims by demonstrating that the FLMP can be proven false and does not have a superpower to predict a plethora of functions or to absorb random variability (see also Massaro, 1998). More recently, Dunn (2000) provided an innovative assessment of a model's propensity to fit arbitrary sets of data. These analyses measure the extent to which the prediction range of a model extends into the potential outcome space. Using this criterion, he found that the FLMP and an alternative, the linear model of perception (LIM) did not differ. The linear model is an additive model analogous to the WTAV given in Equation 2 except that no weight parameter is included and no averaging occurs. Dunn concluded from his analyses "that if the FLMP enjoys any advantage relative to the LIM in being able to fit arbitrary data points, the size of this advantage is very small" (p. 21).

Myung and Pitt (1997) explored the identifiability of three extant models by simulating hypothetical data from a 2 × 8 factorial design, with 20 observations at each of the 16 experimental conditions. They began with three sets of hypothetical parameter values and simulated results from three different models for 100 subjects for each parameter set. The models used to simulate the results were (1) a linear model (LIM) in which the values from the two independent variables were simply averaged (as in Equation 2 but without the weight parameter w), (2) the FLMP (Massaro, 1998), and (3) a model based on signal detection theory (TSD, Massaro & Friedman, 1990). In our earlier work (Massaro, 1987; Massaro & Friedman, 1990), we found that these models made different predictions from one another and that one model's predictions could not mimic another model's predictions when there was no variability in the data. However, it was pointed out that FLMP and TSD made very similar predictions and were probably indistinguishable in practice (when sampling variability and other sources of noise are a factor).

Myung and Pitt (1997, 1998) found that the RMSD measure of goodness-of-fit was not always sufficient to recover the model that actually generated the original data. They found that the FLMP and TSD were more flexible than the LIM in that they gave a better account of the simulated results, even when the LIM was used to generate the data. When the FLMP or TSD was used to generate the hypothetical results, the LIM never provided a better fit than the other two models. When the LIM was used to generate the hypothetical results, it provided a better fit than the other two models only about 28% of the time. Li, Lewandowsky, and DeBrunner (1996) found similar results by evaluating a model's flexibility in terms of its parameter sensitivity, defined as the change in a model's predictions due to variations in the model's parameters, and parameter interdependence, defined as the amount of covariation among parameters during their estimation. They evaluated the FLMP and LIM in the context of the four-factor experiment of Bruno and Cutting (1988) and found that the FLMP had significantly more parameter sensitivity and only somewhat greater parameter interdependence than the LIM.

Proponents of additive or averaging models should have been pleased with these results because they indicate that the LIM might have been erroneously rejected because the competing models were too flexible (Cutting et al., 1992). This is obviously an undesirable state of affairs for advocates of the FLMP, and it challenges our previous work in this arena. On the basis of recent techniques of model testing developed by Jefferys and Berger (1992) and Kass and Raftery (1995), Myung and Pitt (1997) proposed a method of model selection that incorporates both functional form and model flexibility (we prefer this term rather than the term complexity, used by Myung and Pitt) as criteria for selecting the best model (DiCiccio, Kass, Raftery, & Wasserman, 1997). When applied to the simulated results, this method provided a recovery of the "correct" model about 88% of the time. Thus, the previously rejected LIM based on RMSD might actually have been the correct model for previous experiments. Myung and Pitt (1997, 1998), however, did not pursue this possibility and did not present any analyses of actual empirical results (although there are many relevant data sets in the literature) in order to address whether previous outcomes using RMSD were invalid.

The method of model selection is called Bayes factor (Kass & Raftery, 1995) and is defined as a ratio of two marginal likelihoods

$$\frac{P(D\,|\,M_1)}{P(D\,|\,M_2)}, \tag{4}$$

where $P(D\,|\,M_i)$ is the probability of the observed data across all possible parameter values. The term corresponds to an average of likelihoods under a prior distribution of the parameters. Myung and Pitt (1997, 1998) gave an integral form for the marginal likelihood:

$$P(D\,|\,M_i) = \int P(D\,|\,\theta, M_i)P(\theta\,|\,M_i)d\theta \qquad (i = 1,2), \tag{5}$$

where θ is a parameter vector under model i, $P(D\,|\,\theta, M_i)$ is the likelihood function, and $P(\theta, M_i)$ is the prior density of θ for model i.

This Bayes factor method of model selection seeks to handicap models to the extent that they can predict a large range of outcomes with changes in their parameter values. If a model predicts a large range of outcomes with changes in parameter values, then its ability to predict a single data set with all possible parameter values will be very poor. On the other hand, a model that predicts only a small range of outcomes across changes in its parameter values will do much better if the data to be predicted are within that small range of outcomes predicted by the model. This is the logic of handicapping models based on their flexibility.

Viewed from a slightly different perspective, the Bayes factor should not be an entirely new concept to mathematical modelers. One of the traditions in the field is to value what might be dubbed the invariance of parameters in a model. For example, Atkinson, Bower, and Crothers (1965) advocated the value of parameter invariance across different experimental situations. Models might be considered to have additional value when they are able to give a fairly good description of a given set of results across a fairly broad range of parameter values. The Bayes factor formalizes this principle by determining the probability of the observed data given a model across all possible parameter values of the model. Thus, the Bayes factor is concerned with goodness-of-fit across *all* possible parameter values. In contrast, a model is tested using RMSD by finding a *single* set of estimated parameter values to maximize goodness-of-fit to the data. According to the RMSD measure, the investigator is satisfied if at least one set of parameter values gives a good fit. According to the Bayes factor, the investigator is satisfied only if the model does a respectable job of prediction across all parameter values.

The Bayes factor adjusts a model's goodness-of-fit index by the model's ability to describe a large range of different data configurations. A model capable of fitting a broader range of data configurations than another is not necessarily the better model. We desire a model to have good taste and to predict only a constrained set of data outcomes—if any configuration can be predicted, it is not falsifiable. In a series of simulations, for example, it was demonstrated that neural networks with hidden units could mimic a variety of different information processing models (Massaro, 1988). The Bayes factor handicaps a model to the extent that it can predict a broad range of data configurations other than the observed data, by simply different parameter values. According to the assumptions underlying Bayes factor, a better model is one that predicts only data close to the data actually observed, regardless of the parameter values.

This important analysis and potential solution provided by the Bayes factor alerted us to the possibility that our previous model tests may have led us to incorrect conclusions. In many experiments, the FLMP has been found to provide a significantly better fit than alternative models. The demonstration of Myung and Pitt (1997) reveals that our conclusions might have been invalid given the potentially more flexibility of the FLMP to fit results, even results that were not generated by that model. There were several aspects of the Myung and Pitt simulation, however, that did not mirror our prototypical experimental situations. First, the authors simulated data from an *unweighted* averaging model (LIM) rather than a *weighted* averaging model (WTAV) that we have tested in all of our research (Massaro, 1998; Massaro & Cohen, 1976). The FLMP always gave a significantly better fit than the WTAV even though the WTAV also had one additional free parameter than the FLMP, and we did not adjust the RMSD measures to reflect this difference in number of parameter values. We expect the WTAV to be more flexible than the LIM, which would influence the outcome of model selection using Bayes factor.

Weighted averaging is more psychologically realistic than unweighted averaging in that it is unlikely that each influential factor contributes equally to performance in pattern recognition tasks. A weighting parameter allows that a .7 scale value from one factor might make a different contribution than a .7 value from another factor. Differential weighting in the FLMP and TSD descriptions emerges from the nonlinear combination of the two sources of information corresponding to the two factors. Second, the authors simulated data from a highly *asymmetrical factorial* design, whereas we usually carry out *symmetrical expanded* factorial designs. The latter are much more efficient than the former in discriminating among different models. A symmetrical design has the highest ratio of independent observations relative to free parameters, and the expanded design provides an additional set of data points whose expected values are predicted by the same parameter values. Third, the authors used only three hypothetical sets of parameter values to generate hypothetical data, whereas we have contrasted the models in literally dozens of independent tests.

To explore these differences, we replicated their analyses with additional data sets. Given the similar predictions of the FLMP and TSD, however, we eliminated the latter from our horse races. The major concern in the field is also the nature of the additive (linear) and nonadditive models (e.g., Dunn, 2000), which is addressed by comparisons of LIM, WTAV, and FLMP. We also reanalyzed two data sets from our laboratory and a recent experiment carried out by Pitt (1995; Massaro & Oden, 1995).

In addition to analyzing the actual results, we simulated data from these experimental situations in order to assess the validity of the RMSD and Bayes factor measures of model selection.

REPLICATION AND EXTENSION OF MYUNG AND PITT

In replicating and extending Myung and Pitt's analysis, data sets were generated using parameters that simulated five different response patterns in a 2 × 8 factorial design experiment, with two response alternatives. In the context of speech perception this would correspond to choosing between the outcomes /ba/ or /da/, for example. The first three parameter sets are taken directly from Myung and Pitt's (1997) simulation. The fourth and fifth parameter sets were chosen to produce data sets that would exhibit a pattern with either a small or large effect of the two-level factor (Figure 1). Each parameter set was composed of 10 parameter values (2 + 8) with the exception of WTAV, which has a weight (.2426) as an additional parameter (Massaro, 1998, pp. 59–60). Figure 1 shows the

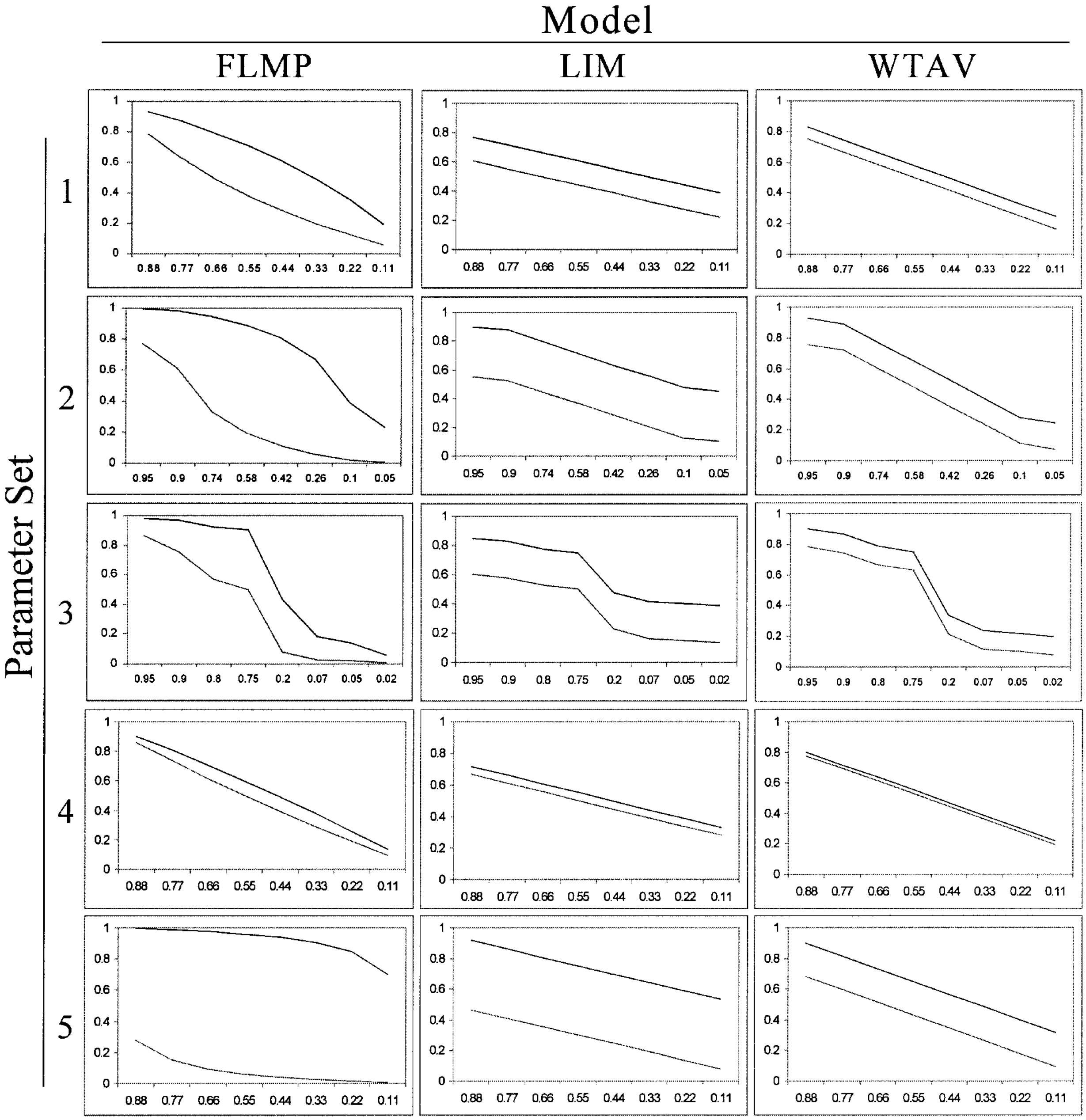

Figure 1. Response patterns generated from five different sets of parameters with three models: FLMP, LIM, and WTAV (without sampling variability or additional noise).

five sets of response patterns for the three models. Probabilities of choice outcomes were generated for each of the 16 conditions by applying a given set of parameter values to a given model. The probability of a subject selecting one of the two responses (e.g., a /ba/) was calculated for each of the 16 experimental conditions as predicted by a particular model.

To generate data from these probabilities that simulate actual human responses, sampling variability was introduced using the binomial probability distribution. For each of the 16 response probabilities, 20 simulated observations (uniformly random between 0 and 1) were generated that were compared with the probability for each condition. If the simulated observation was less than the probability for that condition, then that observation was taken as a subject choosing /ba/; otherwise it would be taken as a /da/ response. For example, in the FLMP case mentioned earlier, the probability of choosing /ba/ in the first condition was .934. If the generated random number was less than .934, then a /ba/ response was recorded for that trial; otherwise a /da/ response was recorded. After this was completed for all 20 observations, the proportion of /da/ responses was taken for each condition to produce the response characteristics of an individual subject. There were 100 simulated subjects using this method for each of the three models for all five parameter sets.

The FLMP, LIM, and WTAV were each fit to the five simulated data sets using both the Bayes factor and the RMSD. The Bayes factor was evaluated using a Monte Carlo simulation of the integral specified in Equation 5 (see also Myung & Pitt, 1997) and was numerically evaluated with 500,000 iterations. Software to perform this calculation was kindly provided by Myung and Pitt. The RMSD method used for model comparison relies on minimizing the RMSD value across the 16 conditions in the experimental design.

Replicating the Myung and Pitt (1997) findings, Table 1 shows that the RMSD measures were biased in favor of the FLMP. When the FLMP generated the results, the RMSD selected the FLMP 93% of the time. When the LIM generated the results, the RMSD selected the LIM only about 12% of the time. Table 2 shows that, using the Bayes factor, these values were 75% and 89%, respectively. However, as we expected, the RMSD was less biased in favor of the FLMP relative to the WTAV. When the WTAV generated the results, the RMSD selected the WTAV 42% of the time. This still fell short of the Bayes factor, which selected the WTAV 74% of the time. However, Data Sets 4 and 5 demonstrate that the Bayes factor is not infallible. The Bayes factor selected the WTAV as the best model 92% of the time for Data Set 4 generated by the FLMP. For Data Set 5, the WTAV failed to recover itself by a narrow margin, being beaten out by the FLMP 52% of the time. Although there is no obvious explanation for these outcomes, they suggest some caution in the use of the Bayes factor.

To summarize, the results, shown in Tables 1 and 2, agree with Myung and Pitt (1997) by showing better model

Table 1
Summary of RMSD and Percentage of Model Wins

	Data Generated From Parameters																	
	Set 1			Set 2			Set 3			Set 4			Set 5			*M*		
Model Fitted	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV
FLMP	0.0546*	0.0546	0.0658	0.0314*	0.0576	0.0601	0.0317*	0.0618	0.0523	0.0601	0.0612	0.0528	0.0246*	0.0456*	0.0522*	0.0405	0.0562	0.0566
% win	88.0	53.0	51.0	100.0	64.0	52.5	100.0	58.0	52.0	77.5	62.0	65.5	100.0	74.0	66.0	93.1	62.2	57.4
LIM	0.0864	0.0636	0.0703	0.1195	0.0622	0.0792	0.1338	0.0659	0.0857	0.0910	0.0636	0.0876	0.1520	0.0571	0.0872	0.1165	0.0625	0.0820
% win	0.0	20.0	4.5	0.0	12.5	0.0	0.0	10.0	0.0	0.0	11.5	0.0	0.0	5.5	0.0	0.0	11.9	0.9
WTAV	0.0683	0.0634	0.0661	0.1114	0.0614	0.0615	0.0755	0.0640	0.0540	0.0629	0.0617	0.0539	0.0554	0.0559	0.0572	0.0747	0.0613	0.0585
% win	12.0	27.0	44.5	0.0	23.5	47.5	0.0	32.0	48.0	22.5	26.5	34.5	0.0	20.5	34.0	6.9	25.9	41.7

*$p < .05$.

recovery using Bayes factor than RMSD for the 2 × 8 factorial design (Bayes factor achieving 10 hits and 1 false alarm vs. RMSD with 4 hits and 2 false alarms). However, the results were highly dependent on the parameter values chosen. The Bayes factor favored the WTAV for Set 4 data and favored the FLMP for Set 5 data.

A comparison of the curves in Figure 1 reveals a potential limitation in simulating data based on equivalent parameter values for the different models. As can be seen in the figure, equivalent parameters for the different models can lead to very different hypothetical data sets, and different parameters for the different models can lead to fairly similar hypothetical data sets. For example, the FLMP simulated data for Parameter Set 1 is most similar to the WTAV simulated data for Parameter Set 2, whereas the WTAV simulated data for Parameter Set 1 is most similar to the FLMP simulated data for Parameter Set 4. When equivalent parameters for the different models are used, any observed differences in a model's ability to fit another model's data set might be (at least partially) due to the data set itself rather than the model. We suggest that the best comparison of simulated data from different models should be driven by the parameter values of a model found in the fit to real data. We carry out this technique in the next section.

EXTENSION TO PROTOTYPICAL DESIGNS AND REAL DATA

In addition to simulation tests with hypothetical parameter values, we used parameter values from real data. Conveniently, we had already established two different databases used for model testing in a bimodal speech perception task.[1] A typical manipulation is to vary the ambiguity of each modality of information by systematically making a continuum between two different syllables; that is, how much it resembles each syllable. Synthetic speech (or at least a sophisticated modification of natural speech) is necessary to implement this manipulation. We used synthetic speech to cross five levels of audible speech varying between /ba/ and /da/ with five levels of visible speech varying between the same alternatives. We also included the unimodal test stimuli to implement the expanded factorial design.

The properties of the auditory stimulus were varied to give an auditory continuum between the syllables /ba/ and /da/. In analogous fashion, properties of our animated face were varied to give a continuum between visual /ba/ and /da/. Five levels of audible speech varying between /ba/ and /da/ were crossed with five levels of visible speech varying between the same alternatives. In addition, the audible and visible speech also were presented alone for a total of 25 + 5 + 5 = 35 independent stimulus conditions. Six random sequences were determined by sampling the 35 conditions without replacement giving six different blocks of 35 trials. An experimental session consisted of these six blocks preceded by six practice trials and with a short break between sessions. There were four sessions of testing for a total of 840 test trials (35 × 6 × 4).

Table 2
Summary of Log Marginal Likelihood Approximations of the Bayes Factor and Percentage of Model Wins

	Data Generated From Parameters																	
	Set 1			Set 2			Set 3			Set 4			Set 5			*M*		
Model Fitted	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV	FLMP	LIM	WTAV
FLMP	−15.48	−17.34	−16.61	−14.20*	−16.15	−16.05	−14.27*	−16.94	−16.20	−15.75	−17.14	−16.09	−12.64*	−15.53	−15.85	−14.46	−16.61	−16.15
% win	71	0	1	100	0	14	100	1	19	8	0	2	94	12	52	74.6	2.6	17.6
LIM	−16.25	−15.31*	−15.59	−17.17	−14.97*	−16.43	−21.60	−15.75*	−18.20	−16.88	−15.85*	−18.06	−21.93	−15.28*	−17.74	−18.76	−15.43	−17.20
% win	7	99	39	0	98	3	0	90	1	0	68	0	0	88	0	1.4	88.6	8.6
WTAV	−15.95	−15.72	−15.51	−17.83	−15.58	−15.63*	−17.13	−16.10	−15.92*	−15.29*	−15.87	−15.53*	−13.69	−16.01	−16.25	−15.98	−15.85	−15.76
% win	22	1	60	0	2	83	0	9	80	92	32	98	6	0	48	24	8.8	73.8

*$p < .05$.

Thus there were 24 observations at each of the 35 unique experimental conditions. Subjects were instructed to listen and to watch the speaker and to identify the syllable as /ba/ or /da/. This experimental design was used with 82 subjects (Massaro, Cohen, Gesi, & Heredia, 1993; Massaro, Cohen, & Smeele, 1995) and these results have also served as a database for testing models of pattern recognition (Massaro, 1998, chaps. 2 and 10).

The mean observed proportion of /da/ identifications was computed for each subject for the 35 unimodal and bimodal conditions. The points in Figure 2 give the observed proportion of /da/ responses for a subject who can be considered typical, for the auditory alone (left plot), the bimodal (middle plot), and the visual alone (right plot) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. For the unimodal plots, the degree of influence of a modality is indicated by how much the response function changes across the continuum. By this criterion, both the auditory and the visual sources of information had a strong impact on the identification judgments. As illustrated in the left and right plots, the identification judgments changed systematically with changes in the audible and visible sources of information. The likelihood of a /da/ identification increased as the auditory speech changed from /ba/ to /da/, and analogously for the visible speech.

For the bimodal results in the middle plot, the degree of influence is again indexed by the changes in the functions across the variable plotted on the x-axis, and by the spread among the curves for the variable described in the key or legend. By these criteria, both sources had a large influence in the bimodal conditions. The curves across changes in the auditory variable are relatively steep and also spread out from one another with changes in the visual variable. Finally, the auditory and visual effects were *not* additive in the bimodal condition, as demonstrated by a significant auditory–visual interaction. The interaction is indexed by the change in the spread among the curves across changes in the auditory variable. This result is consistently obtained in this type of experiment. It means that the influence of one source of information is greatest when the other source is neutral or ambiguous.

The FLMP gave a better description than the WTAV model for 94% of these 82 subjects. To analyze the validity of the RMSD measure, we created a set of hypothetical subjects who behaved according to either one model or the other. This Monte Carlo simulation involved creating 20 simulated subjects for each model for each real subject. The parameter values that optimized the model fit (using RMSD) for a given subject were used to simulate hypothetical results conforming to the model's outcome. Thus, the simulated subject had some probability of response for each experimental condition. For example, the predicted probability of a /da/ response for the real subject might be .75 for a given condition. In our two-alternative task, the probability of a /ba/ response would consequently be .25. For each simulated subject, a uniform random number between 0 and 1 is drawn. If the number was less than or equal to .75, then the simulated response would be a /da/. If the number was greater

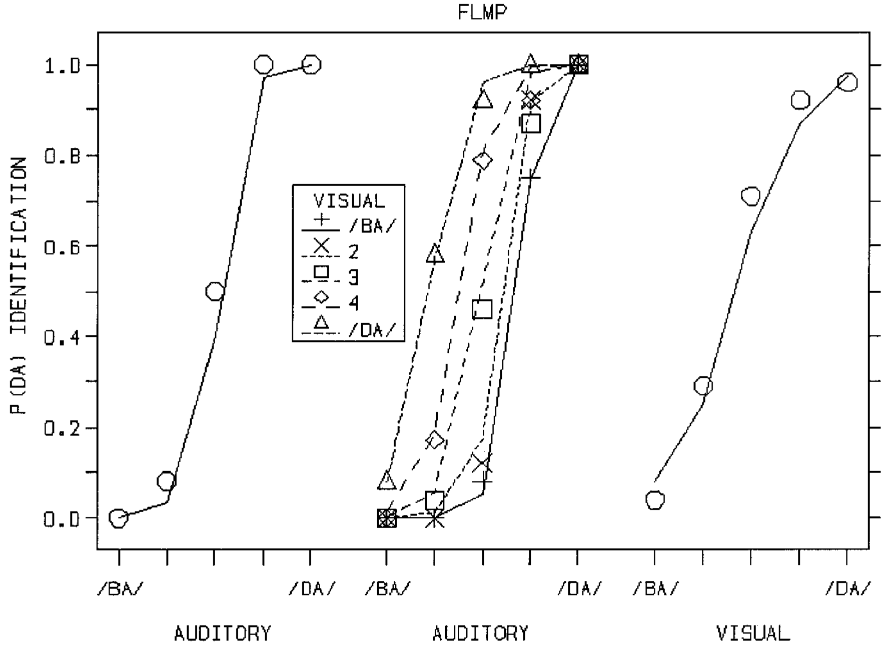

Figure 2. The points give the observed proportion of /da/ identifications for a typical observer in the auditory-alone (left panel), the factorial auditory–visual (center panel), and the visual-alone (right panel) conditions as a function of the five levels of the synthetic auditory and visual speech varying between /ba/ and /da/. The lines give the predictions of the FLMP.

than .75, then the simulated response would be a /ba/. This computation was carried out 24 times to simulate the 24 observations in the experiment. Because the simulation uses the same number of trials, it should have the same sampling variability as was present in the data set being modeled. The same procedure was carried out for each of the 35 conditions of the experiment, resulting in a set of results corresponding to 1 simulated subject. Because any 1 simulated subject does not provide a good estimate of the variability, the procedure was used to create 20 simulated subjects for each real subject. For these simulated subjects, with the same sampling variability, the RMSD measure was sufficient to recover the original model that generated the data. For both data sets, the incorrect model was recovered only 1% of the time.

To further test the validity of the RMSD measure, we compared it against the Bayes factor for our prototypical design. Similar to Myung and Pitt (1997), we generated 100 simulated subjects with the Monte Carlo simulation procedure described above. Rather than choosing arbitrary parameter values, the mean parameter values for the fit of the FLMP averaged across the 82 real subjects were used to generate 100 FLMP subjects. Similarly, the mean parameters for the fit of the WTAV were used to generate 100 WTAV subjects. These two groups of 100 data sets each were then fit by both models with either the RMSD measure or the Bayes factor.

Replicating the earlier simulation, the results show that the RMSD measure is quite sufficient for our prototypical design with 99% correct recovery for the FLMP and WTAV models. The Bayes factor also did well by correctly recovering the FLMP and WTAV models for 98% of the cases. These results indicate that both Bayes factor and RMSD were adequate for selecting among these competing models under these conditions.

To directly insure that the Bayes factor does not revise our conclusions in past work, we tested the FLMP against the WTAV model using the Bayes factor for our prototypical design. These two models were fit to the observed data of the 82 subjects. Similar to the RMSD results, the Bayes factor showed that the FLMP gave a better fit of the 82 subjects with an average marginal log likelihood of −49.0 in contrast to a value of −57.2 for the WTAV. Using the Bayes factor selection method, the FLMP fit better than the WTAV for 80% of the subjects. Although this difference was statistically significant, 80% wins is still somewhat short of the 94% wins using the RMSD selection method. Because of this discrepancy, we repeated the Bayes factor with 5,000,000 rather than 500,000 iterations in the computation of the marginal likelihoods. The FLMP now fit better than the WTAV for 94% of the subjects. These results support the idea that the RMSD measure yields similar conclusions to the Bayes factor for the conditions of our prototypical design.

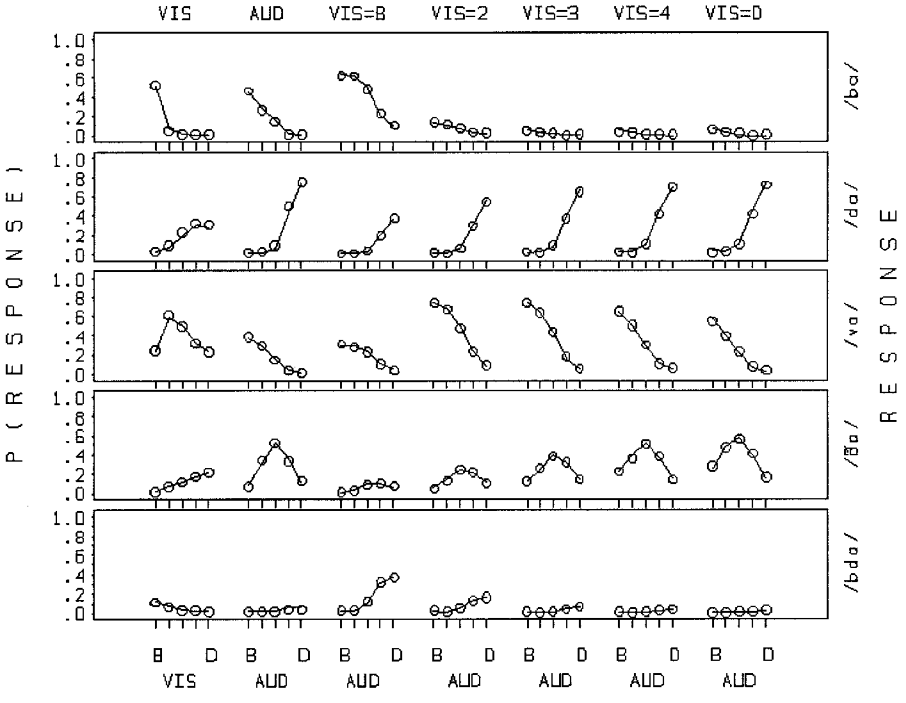

Figure 3. Observed (points) and predicted (lines) proportion of /ba/, /da/, /va/, /tha/, and /bda/ identifications for the visual-alone (leftmost plot), auditory-alone (second plot), and bimodal (remaining plots) conditions as a function of the five levels of the synthetic auditory (AUD) and visual (VIS) speech varying between /ba/ (B) and /da/ (D). The lines give the predictions of the FLMP for the task with a fixed set of eight response alternatives.

EIGHT-ALTERNATIVE TASK

We explored the same question when the number of response alternatives was eight. We replicated our basic 5 × 5 expanded factorial design with eight rather than just two response alternatives. This basic task was carried out in four different experiments to give a total of 36 subjects in this data set (Massaro, 1998, chap. 10). Each of the 35 stimulus conditions was tested 12 times. It should be noted that this experiment has only half the number of observations per condition as the previous one, which is important because goodness-of-fit is highly dependent on sampling variability. Subjects were instructed to listen to and watch the talker, and to identify the syllable as /ba/, /da/, /bda/, /dba/, /tha/, /va/, /ga/, or "other." The category "other" was to be used by the subject whenever none of the other seven responses seem suitable. These eight response alternatives were determined from pilot studies in which the responses were not constrained.

Figure 3 gives the average proportion of responses across the 36 participants. Although the test continua were between /ba/ and /da/, we obtained several other response alternatives. The most frequent responses were /ba/, /da/, /va/, /tha/, and /bda/. The alternatives /dba/, /ga/, and "other" were seldom used. There was a strong contribution of both audible and visible speech. The number of /ba/ judgments increased toward the /ba/ end of the visible continuum. The /bda/ judgments occurred primarily when a visible /ba/ was paired with an auditory syllable from the /da/ end of the continuum. Visible /da/ articulations increased the likelihood of /da/, /tha/, and /va/ responses. The visual information influenced the likelihood of a /va/ judgment primarily at the /ba/ end of the auditory continuum. The visual /ba/ endpoint stimulus decreased the number of /va/ responses, whereas the other four visual levels increased the number of /va/ judgments at the /ba/ end of the auditory continuum. These judgments reflect the contribution of both auditory and visual speech, even when subjects are permitted a larger permissible set of response alternatives.

The FLMP is tested against results with multiple response alternatives in the same manner as with just two response alternatives (Massaro, 1998, pp. 184–186). With more than two alternatives, it is necessary to estimate a unique parameter to represent the degree to which each source of information supports each alternative. The degree to which the auditory speech supports an alternative such as /ba/ would be a_{Bi}, whereas v_{Gj} would correspond to the visual support for the alternative /g/. The total support for the alternative /ba/ would be

$$S(/\mathrm{ba}/) = a_{Bi} v_{Bj}\,. \tag{6}$$

The support is computed in a similar manner for the other alternatives. The probability of a particular response such as /ba/ would be

$$P(/\mathrm{ba}/) = \frac{S(/\mathrm{ba}/)}{\sum S_r}, \tag{7}$$

where $\sum S_r$ is the sum of the total support values over all r possible alternatives.

The fit of this model requires five a_i and five v_j parameters for each of the eight response alternatives, for a total of 80 free parameters. This might seem like a large amount, but we have increased the number of data points to be predicted by the same factor. We are now predicting 35 × 8 = 280 data points. The fit of this model to each of the 36 subjects produced an average RMSD of .0507. Figure 3 also gives the average of the predicted results. To assess whether the FLMP maintains its advantage with multiple response alternatives, we compared this fit with that of a single-channel model (or equivalently, a weighted averaging model). The fit of this competing model was about two times poorer, giving an RMSD of .1049.

For these 36 participants, the FLMP gave a better description than the WTAV model for 97% of the real subjects. To analyze the validity of the RMSD measure, we repeated the Monte Carlo simulations by creating simulated subjects for each real subject as in the two-alternative task. Each simulated subject had some probability of response for each experimental condition that is set equal to the predicted proportion of response in a real subject. In this case, there are eight predicted probabilities that sum to 1 for each experimental condition. This computation was carried out 12 times to simulate the 12 observations in the experiment. By using the same number of trials, the simulation has the same sampling variability as was present in the data set being modeled. The same procedure was carried out for each of the 35 conditions of the experiment, resulting in a set of results corresponding to 1 simulated subject. As in the two-alternative task, we created 20 simulated subjects. For these simulated participants, the RMSD measure was sufficient to recover the original model that generated the data. For data sets generated from both the FLMP and WTAV, the incorrect model was recovered only about 1% of the time.

We also evaluated the FLMP against the WTAV model using the Bayes factor. For the two-alternative task, the Bayes factor is performed by calculating the binomial response probabilities for each set of parameter values given some N number of observations. Since this experiment allowed for eight response alternatives, we calculated multinomial response probabilities instead of binomial using the following equation:

$$p = \frac{n!}{\prod_{i=1}^{r} (x_i)!} \left(\prod_{i=1}^{r} p_i^{x_i} \right), \tag{8}$$

where n is the number of observations, r is the number of response alternatives, x_i is the number of choices for response i, and p_i is the observed probability of response i.

The observed responses of the 36 subjects served as the data for the Bayes factor. The FLMP fit with a log likelihood of −163.5, while the WTAV model performed worse, with a log likelihood of −180.2. Using the Bayes factor, the FLMP fit 97% of the subjects better than the

WTAV model. These results are consistent with those obtained with the RMSD measure and further support our claim that the RMSD is an accurate measure of model performance for our prototypical design. These results also extend this conclusion to designs with more than two response alternatives.

PITT (1995) DATA

Pitt (1995) studied the joint influence of phonological information and lexical context in an experimental paradigm developed by Ganong (1980). In this task, a speech continuum is made between two alternatives, and the contextual information supports one alternative or the other. The initial consonant of the CVC syllable was varied in six steps between /g/ and /k/. The following context was either /Ift/ or /Is/. The context /Ift/ favors or supports initial /g/ because *gift* is a word whereas *kift* is not. Similarly, the context /Is/ favors or supports initial /k/ because *kiss* is a word whereas *giss* is not. He improved on earlier studies by collecting enough observations to allow us to perform a subject-by-subject evaluation of the ability of specific models of language processing to account for the results. Previous tests of models using this task have been primarily dependent on group averages which may not be representative of the individuals that make up the averages. Each model was applied to the identification results of the 12 individual subjects in Pitt's Experiment 3a, for which the greatest number of observations (104) were obtained for each data point for each subject. The points in Figure 4 give the observed results for each of the 12 subjects in the task. For most of the subjects, the individual results tend to resemble the average results reported by Pitt and earlier investigators. Ten of the 12 subjects were influenced by lexical context in the appropriate direction. Subject 1 gave an inverse context effect and Subject 7 was not influenced by context.

According to the FLMP, both the bottom-up information from the initial speech segment and the top-down context are evaluated and integrated. If s_i is the degree of support for the voiced alternative given by the initial segment and c_j is the support given by the following context, the total support for the voiced alternative is

$$S(voiced\,|\,S_iC_j) = s_i \times c_j, \tag{9}$$

The support for the voiceless alternative would be

$$S(voiceless\,|\,S_iC_j) = (1 - s_i) \times (1 - c_j). \tag{10}$$

The predicted probability of a voiced response is simply

$$P(voiced\,|\,S_iC_j) = \frac{S(voiced\,|\,S_iC_j)}{S(voiced\,|\,S_iC_j) + S(voiceless\,|\,S_iC_j)}. \tag{11}$$

In producing predictions for the FLMP, it is necessary to estimate parameter values for each level of each experimental factor. The initial consonant was varied along six steps between /g/ and /k/, and the following context was

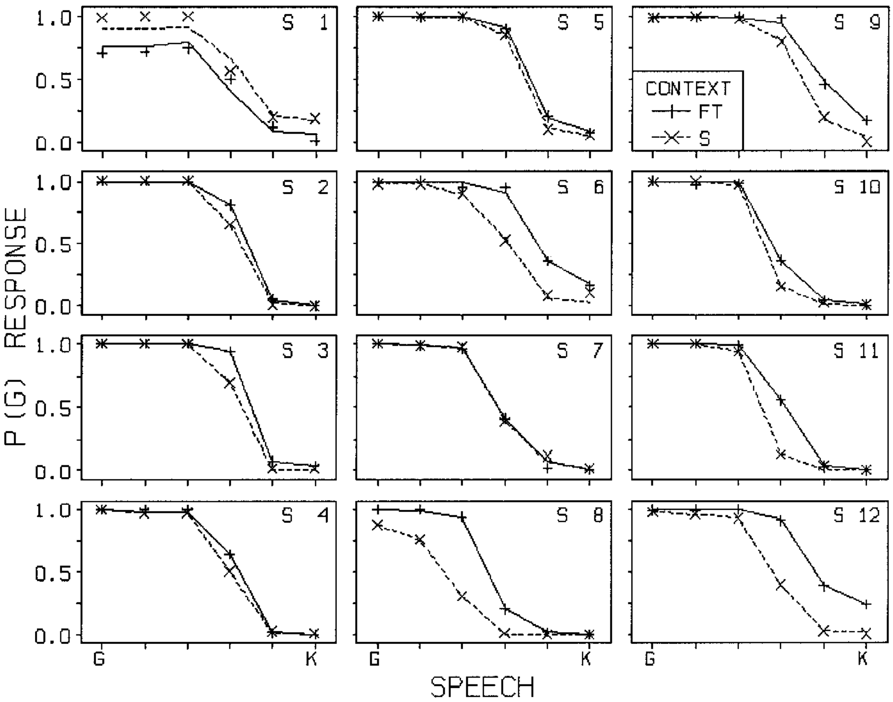

Figure 4. Observed (points) and FLMP's predictions (lines) of /g/ identifications for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995) Experiment 3a.

either /Ift/ or /Is/. Thus, there were six levels of bottom-up phonological information s_i and two contexts c_j. A free parameter is necessary for each level of bottom-up information, but it is reasonable to assume that the contextual support given by /Is/ is one minus the lexical support given by /Ift/, so that only one value of c_j needs to be estimated. Thus, seven free parameters are used to predict the 12 independent data points: six values of s_i and 1 value of c_j.

The lines in Figure 4 also give the predictions of the FLMP. As can be seen in the figure, the model generally provides a good description of the results of this study. The RMSD between predicted and observed results is .017 on the average across all 12 independent fits. For the 10 subjects showing appropriate context effects, the RMSD ranges from .003 to .045 with a median of .007. Thus, for each of these individuals, the model captures the observed interaction between phonological information and lexical context: The effect of context was greater to the extent that the phonological information was ambiguous. This yields a pattern of curves in the shape of an American football, which is a trademark of the FLMP.

To further confirm that the FLMP was not overfitting data only due to its putative flexibility, an analysis of data was performed using the Bayes factor. Replicating the RMSD analysis, the Bayes factor decided in favor of the FLMP for 11 of the 12 subjects, a highly statistically significant result (Table 3). The FLMP gave a better fit of the observed results with a marginal log likelihood of −19.9, whereas the marginal log likelihood of the WTAV was −31.6.

Subject 1 was the only subject whose context effect was in the opposite direction relative to the other subjects (and opposite to reasonable expectation). The FLMP gave a very poor description of this subject's results, yielding an RMSD of .066. The Bayes factor also selected the WTAV over the FLMP as the better fitting model. The fact that the FLMP did not provide a good fit to this subject's data is evidence that the model is not so flexible that it can fit anything. The failure of the FLMP in describing these anomalous results supports our argument that the FLMP does not have an excessive flexibility or some other "unfair" advantage.

Table 3
RMSD and Log Marginal Likelihood Approximations of the Bayes Factor for Pitt (1995) Data

	RMSD		Log Marginal Likelihood	
Subject	FLMP	WTAV	FLMP	WTAV
1	0.0662	0.0445	−31.85	−28.08
2	0.0047	0.0286	−16.62	−30.74
3	0.0058	0.0494	−20.92	−29.64
4	0.0057	0.0311	−14.46	−26.90
5	0.0083	0.0198	−15.66	−19.84
6	0.0450	0.0745	−24.78	−30.10
7	0.0210	0.0196	−17.73	−26.26
8	0.0029	0.1182	−21.06	−53.85
9	0.0183	0.0593	−17.59	−33.51
10	0.0093	0.0402	−19.77	−24.29
11	0.0073	0.0808	−19.16	−32.14
12	0.0051	0.1025	−19.71	−43.53
Average	0.0166	0.0557	−19.94	−31.57

BENCHMARK CRITERION FOR GOODNESS-OF-FIT

We cannot expect our models to predict the proportion of judgments exactly. The reason is sampling variability: A finite sample cannot be expected to match the actual probability of an event. Even if we knew the true predicted probability for some experimental condition, we could not expect to observe this probability in actual practice. Thus, it is necessary to know the sampling variability in order to evaluate goodness-of-fit. There are methods to determine how accurate the prediction has to be to be considered correct. These involve a computation of the variability in our predictions, in terms of a benchmark RMSD. This makes transparent the pivotal role played by RMSD. We have seen how the RMSD provides a measure of the deviation of a model's predictions from a set of observed results. The benchmark RMSD gives an analogous measure of the deviation between a model's predictions and data that were actually generated by the model.

With just two response outcomes, there is a computation that allows us to estimate sampling variability directly. This direct estimation is given by a closed-form equation based on binomial variance. Binomial variance is simply the expected variability for two-outcome experiments and is a function of the probability of each outcome and the number of repeated observations. The probability of each outcome is estimated from the observed response proportions, and the number of observations is equal to the number of trials. The binomial variance must be averaged over all conditions of the experiment, and its square root is the benchmark RMSD (see Massaro, 1998, chap. 10). The benchmark RMSD can be computed by either a closed-form expression when there are two response alternatives or by Monte Carlo simulation for three or more response alternatives.

In earlier simulations of the two-alternative task (Massaro, 1998), we found that the fit of the FLMP fell slightly shy of the benchmark. However, the addition of decision noise (noise added at the response selection stage) with a standard deviation of .1 brought the RMSD for the FLMP in line with what would be expected from the data being generated by this model. It is therefore important to determine whether our conclusions about model selection hold up when additional noise is added to the model fits. When this amount of noise was added to simulated results, the WTAV model gave the best fit to data simulated by this model for 93% of the simulated subjects. The FLMP gave the best fit to data simulated by this model for 96% of the simulated subjects. Thus, the RMSD measure for model selection doesn't seem to be problematic for our prototypical experimental situations.

The procedure of computing benchmarks and adding noise in the model fits revealed a severe limitation in the WTAV. In the two-alternative task, no amount of noise could be added that could bring the WTAV in line with that expected if indeed the WTAV was the correct model that generated the results (Massaro, 1998, chap. 10).

When additional variability was added to the results in the eight-alternative task, the models continued to be recoverable. In earlier simulations (Massaro, 1998), we found that the addition of decision noise (noise added at the response selection stage) with a standard deviation of .012 brought the RMSD for the FLMP in line with what would be expected from the data being generated by this model. When this amount of noise was added to simulated results, the FLMP gave the best fit to data simulated by FLMP for 99.6% of the simulated subjects. For the WTAV, .118, or over nine times as much noise as required for the FLMP, was necessary to bring the RMSD in line with the observed RMSD. When this amount of noise was added to simulated results, the WTAV model gave the best fit to data simulated by the WTAV for 95.6% of the simulated subjects. Clearly, the RMSD measure is capable of recovering the "correct" model in our prototypical experimental situations. Thus, the RMSD measure for model selection doesn't seem to be problematic for our prototypical experimental situations.

For a final comparison of the RMSD and Bayes factor, we evaluated model recovery with increasingly noisy data. We expect that with increasing noise, model recovery will decline monotonically for both the FLMP and WTAV models. Eighty-two simulated subjects were created by Monte Carlo simulation. Rather than using the average parameter values, we used the parameter values of each real subject to generate each simulated subject. Gaussian distributed noise was added during simulation by adding a noise value to the sampled probability. The noise value was created by randomly selecting a number from a Gaussian distribution with a standard deviation at some given noise level. There were seven noise levels given in standard deviations: 0.00, 0.05, 0.10, 0.15, 0.20, 0.40, and 0.80. Seven sets of FLMP data and seven sets of WTAV data were generated using this method and then were cross-fit by each model using RMSD and Bayes factor.

Table 4 shows the average RMSD values for the RMSD method and the average marginal log likelihoods using Bayes factor. The results for the RMSDs indicate, as expected, model recovery declines for both the FLMP and WTAV models as the data become more noisy. The results for the Bayes factor show that model recovery declined for the FLMP, but that model recovery for the WTAV model actually levels off at a large value of 88%. If this was the result of the superior model recovery ability of Bayes factor, then why isn't the FLMP also recovered at high levels of noise? Our interpretation is that Bayes factor is biased for less flexible models. Models are less flexible when their predicted results do not change much with changes in their parameter values. Given the lack of meaningful data in the noisy conditions the goodness-of-fit (actually, poorness-of-fit as indexed by the large RMSDs) for both models is about the same. However, Bayes factor penalizes the FLMP for being more flexible. Therefore, care should be taken to ensure that the data are not too noisy when using Bayes factor because an unfair advantage will be given to simple (less flexible) models.

Table 4
Summary of Average RMSD Values, Log Marginal Likelihood Approximations, and Percentage of Wins for Each Model for RMSD and Log Marginal Likelihood (Bayes Factor) as a Function of Noise Level

		Data Generated With Decision Noise			
		RMSD		Log Marginal Likelihood	
Noise (*SD*)	Model	FLMP	WTAV	FLMP	WTAV
0.00	FLMP	0.0367	0.0864	−35.9	−41.9
	% win	99	1	96	3
	WTAV	0.1085	0.0530	−41.0	−37.5
	% win	1	99	4	97
0.05	FLMP	0.0524	0.0920	−38.1	−43.4
	% win	98	2	95	4
	WTAV	0.1127	0.0650	−43.1	−38.1
	% win	2	98	5	96
0.10	FLMP	0.0767	0.1071	−41.9	−46.0
	% win	96	6	89	6
	WTAV	0.1233	0.0881	−45.6	−40.4
	% win	4	94	11	94
0.15	FLMP	0.1021	0.1267	−44.8	−49.3
	% win	94	13	78	11
	WTAV	0.1384	0.1137	−48.9	−43.2
	% win	6	87	22	89
0.20	FLMP	0.1273	0.1478	−48.9	−52.6
	% win	88	22	78	11
	WTAV	0.1558	0.1390	−52.8	−46.5
	% win	12	78	22	89
0.40	FLMP	0.2150	0.2278	−66.7	−70.1
	% win	78	47	62	12
	WTAV	0.2292	0.2281	−70.9	−63.0
	% win	22	53	38	88
0.80	FLMP	0.3089	0.3141	−91.9	−95.2
	% win	66	42	61	12
	WTAV	0.3173	0.3192	−96.9	−86.4
	% win	34	58	39	88

To clarify the results and our interpretation, we computed sensitivity and criterion values using the theory of signal detectability, even though we are not making any strong claims about the nature of the underlying distributions. This analysis is simply being used to measure how well a model selection procedure distinguishes the two models and whether there is a bias to favor one of the models over the other. Table 5 gives the d' values computed from the hit and false alarm rates. As can be seen in the table, the d's decrease for both types of model selection as the data become more noisy. Furthermore, the overall sensitivity of both methods of model selection is about the same. However, RMSD is better able to discriminate between the two different models with low noise and less able to discriminate the two models with high noise. The Bayes factor is better able to discriminate the two models when the data are very noisy.

The beta values are also shown in Table 5 and indicate the degree to which each model selection procedure favors one model over the other. Beta values less than

Table 5
Summary of *d′* Values for RMSD and Log Marginal Likelihood (Bayes Factor)

	RMSD		Bayes Factor	
Noise Level	*d′*	beta	*d′*	beta
0.00	4.65	1.00	3.63	1.27
0.05	4.11	1.00	3.40	1.20
0.10	3.31	0.72	2.78	1.58
0.15	2.68	0.56	2.00	1.58
0.20	1.95	0.68	2.00	1.5
0.40	0.85	0.74	1.48	1.90
0.80	0.61	0.94	1.45	1.92
M	2.59	0.81	2.39	1.56

1 favor the FLMP and beta values greater than 1 favor the WTAV. As shown in Table 4, as the noise increases and the *d′*s decrease, the RMSD method tends to favor the more flexible FLMP and the Bayes factor tends to favor the less flexible model.

One argument against the use of RMSD is that its measure of discrepancy between the observed and predicted probability values is necessarily linear even though the expected variance across probability is nonlinear. In a two-alternative task, for example, the expected variance is greatest for probabilities of .5 and decreases monotonically as the probabilities move to either 0 or 1. The RMSD (see Equation 3) simply adds the squared discrepancies without taking into account this nonlinearity. The computed RMSD is necessarily influenced to a greater extent by the conditions with the more ambiguous probabilities than the unambiguous probabilities. This is not a major problem if the goal is simply to contrast different models against the same observed probabilities. Because the different models are being tested against the same observed probabilities, the bias in the RMSD measure is the same across the different models. For the absolute assessment of model accuracy, we have the benchmark RMSD. Conveniently, the benchmark RMSD provides a standard that overcomes this limitation in the computation of the observed RMSD. That is, the benchmark RMSD is based on the same observed probabilities and therefore provides a direct and valid comparison for the observed RMSD.

A TEST OF NEWTON'S LAW

We provide one further illustration of our hypothesis that one model selection procedure is not always better than another. As we have argued, the Bayes factor might unduly favor simple but incorrect models. A data set was generated by assuming the gravitational force in Newtons (N) was measured between two large objects (.98 and .97 kg) and eight objects of varying masses (.99, .85, .70, .60, .40, .30, .15, and .01 kg) for a total of 2 × 8 = 16 measurements. Each of these masses was a parameter value resulting in 2 + 8 = 10 parameters for each model. For the weighted additive version of Newton's law (WNAV), an additional weight value was included that had a value of 1.0. The 16 measurements were sampled 40 times each to produce a unique data set. This sampling was repeated 100 times to create 100 data sets. Finally, for each model the universal gravitational constant (G) was

$$6.672\times10^{-11}\frac{m^3}{(kg\cdot\sec^2)},$$

and the distance between objects (d) was 8.17×10^{-6} m or .008 mm, resulting in a measured force (F) always between 0 and 1.0. Newton's law was calculated according to the following:

$$F=\frac{GMm}{d^2}, \tag{12}$$

where M is the mass of the two large objects and m is the mass of the eight objects. Newton's law was compared with a simple weighted additive version of Newton's law:

$$F=\frac{G(wM+(1-w)m)}{d^2}, \tag{13}$$

where w is a weight between 0 and 1.0.

Newton's law (NMP) and a simple weighted additive version of Newton's law (WNAV) were each fit to the simulated data set using both the Bayesian factor and the RMSD. The Bayes factor was evaluated using a Monte Carlo simulation of the integral specified in Equation 5 of Myung and Pitt (1997) and was numerically evaluated with 1,000,000 iterations, in the same manner as our previous tests. The RMSD method used for model comparison relies on minimizing the RMSD value across the 16 conditions. Table 6 gives the outcomes of the Bayes factor and RMSD selection procedures. The Bayes factor erroneously favored the WNAV over the true NMP 78% of the time ($p < .05$). The RMSD criterion, on the other hand, always recovered the true NMP. The RMSD erred in the opposite direction, significantly favoring the NMP for data generated by the WNAV. Thus, this example further reinforces our conclusion that RMSD is biased toward the more flexible model (which we already knew) and that the Bayes factor may be biased toward the less flexible model (which was not previously known). Although an unbiased model selection method is the goal and may yet be devised, multiple methods of selection should be used in theory testing.

Table 6
Summary of Log Marginal Likelihood Approximations of the Bayes Factor and RMSD for Newton's Law (NMP) and a Simple Weighted Additive Version of Newton's Law (WNAV)

	Data Generated From Parameters			
	Bayes Factor		RMSD	
Model Fitted	NMP	WNAV	NMP	WNAV
NMP	−23.19	−30.69	.0517*	.0154*
% win	22	0	100	90
WNAV	−13.81*	−13.81*	.0764	.0167
% win	78	100	0	10

*$p < .05$.

ISSUE OF FREE PARAMETERS

Some modeling traditions place great weight on deriving a priori constraints on a model's free parameters. Earlier advocates of the localist connectionist modeling had great faith in specifying the parameter values before any data were observed. According to Grainger and Jacobs (1998), the state of the art today would involve large-scale parameter-fitting procedures to test among different models. Frauenfelder and Peters (1998) adopted a parameter estimation routine to obtain a better match to existing data. In defense of this action, they said that the new parameter set was chosen so that it did not change the basic behavioral pattern of the TRACE model. However, it is important to determine what one means by the basic behavioral pattern of a given model. In principle, if a model has a set of free parameters, then any parameter values within that set should represent the "basic behavioral pattern" of the model. In FLMP, for example, there are no parameter values that would produce a different set of behavioral patterns than those that have been assumed.

We cannot expect a model's predictions of behavior to be exact or even very accurate without first taking into account what results are being predicted. As an example, we cannot know exactly how often a given person will identify one of the visible speech syllables as a particular alternative. Individual subjects give similar but not identical results for the same experiment. We can know that one syllable might be more likely to be identified as a /ba/ than another, but we do not know how much more. This uncertainty would preclude the quantitative test of models if we were not able to determine the values of (estimate) free parameters.

The idea of free parameters has received a steady stream of bad press. Miller, Galanter, and Pribram (1960, p. 182) remarked that "a good scientist can draw an elephant with three parameters, and with four he can tie a knot in its tail." Although we can grant that too many free parameters elevate a model beyond falsifiability, they are still necessary for accurate prediction. To convince the reader that prediction is not possible without free parameters, we carried out the following exercise. The argument is that we cannot make a priori predictions of how a given person will categorize a given source of information. To illustrate this, we computed the average probability of a /da/ response, *P*(/da/), for each subject to each of our test conditions. Consider the third level of the unimodal auditory stimulus, A_3, and the third level of the unimodal visual stimulus, V_3. Figure 5 plots *P*(/da/) for each of our 82 subjects to the unimodal V_3 as a function of *P*(/da/) to the unimodal A_3.

As can be seen in the figure, the response probabilities are distributed across the complete range of possible values for both A_3 and V_3. Figure 5 also shows that there is very little correlation between the two modalities. This

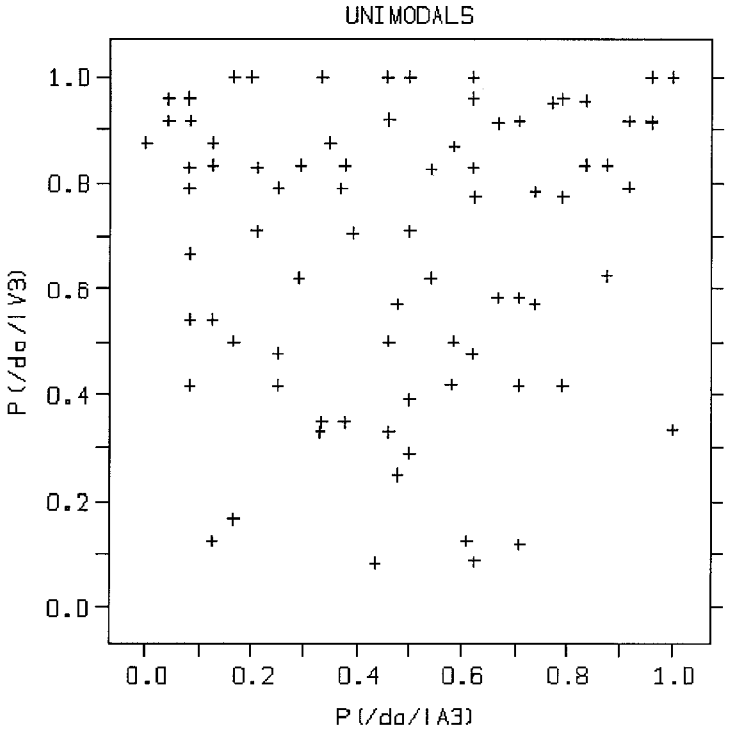

Figure 5. The probability of a /da/ response for each of 82 subjects to V_3 as a function of their probability of a /da/ response to A_3.

result replicates in another way the earlier finding of independence between auditory and visual speech recognition tasks (Raney, Dancer, & Bradley, 1984). Knowing a person's performance in one modality will not help predict his/her performance in the other modality. The conclusion from this exercise is that we cannot have a parameter-free model of speech performance. Given this variability across subjects, we cannot hope to predict a person's judgments without some type of parameter estimation based on that person's actual results. Even traditional psychophysical theories require free parameters. Similarly, speech scientists cannot be expected to predict categorization of a speech stimulus even if all of its physical properties are known. This outcome also is consistent with our proposal (Massaro, 1998, chap. 11) that average results can be meaningless and that theories should be aimed at individual performance. Most importantly, it might seem unreasonable that a model should be handicapped for flexibility when this flexibility is needed to account for individual variability.

Some researchers are uncomfortable with any model the predictions of which require free parameters (Grant, Walden, & Seitz, 1998). As a solution for the variability in Figure 5, they might propose that a subject could be given two independent tests. Parameters could be estimated from the first test and used to predict the results of the second (Dijkstra & de Smedt, 1996). This is not an unreasonable suggestion as long as it is realized that the parameter estimates will not be as accurate as they would be when estimated from all of the data being predicted. This method of testing a model against new results based on parameter estimates from old results must necessarily give a poorer description of performance than the case in which all of the observations being predicted are used to estimate the free parameters. Physical theories are sometimes used as ideal examples of prediction: Edmund Halley was able to predict the location of his comet over 100 years later, for example. What is forgotten, however, is that he had a very reliable measure of the comet's location at an earlier time in a highly deterministic physical system. The recent decades of Chaos have left even physical scientists less sanguine about chance and prediction (Casti, 1994; Waldrop, 1992).

AN ANALOGY WITH NEWTON'S LAW

As an attempt to clarify the role of predictability and the use of free parameters, we draw an analogy between the FLMP and Newton's Law of Universal Gravitation. His law states that the gravitational force (F) between any two bodies of mass m and M, separated by a distance d, is directly proportional to the product of the masses and inversely with the square of their distance.

$$F = \frac{GMm}{d^2}, \tag{14}$$

where G is the universal gravitational constant.

In the FLMP, the bimodal performance is some function of the product of the information of the auditory and visual sources presented separately. For two choice alternatives, the response information on bimodal trials, b_{ij}, is given by

$$b_{ij} = \frac{a_i v_j}{a_i v_j + (1 - a_i)(1 - v_j)}, \tag{15}$$

where i and j index the levels of auditory and visual information, respectively.

In the same way that Newton requires information about the masses (and their distance) in order to predict the gravitational attraction, the FLMP requires information about the information values of the auditory and visual sources in order to predict the bimodal outcome. One cannot determine both of these types of values without some form of prior measurement. In physics, the masses are measured, whereas in psychophysics, some psychophysical or psychological test can be carried out to determine information values.

We agree with Myung and Pitt (1997) and others that model flexibility is an important dimension of predictive models and one that must be assessed in choosing among competing models. The Bayes factor method of model selection is one specific technique for assessing flexibility and handicapping models accordingly. To gain further insight into this method, we can extend our analogy in physics even further to consider various techniques of model testing and selection. In the same manner that the LIM has been proposed as an alternative to the FLMP, we can compose an alternative to Newton's law in which the two masses are added rather than multiplied.

$$F = \frac{G(M + m)}{d^2}, \tag{16}$$

where G is the universal gravitational constant.

If we generalize from the Bayesian model selection results of Myung and Pitt (1997) for the FLMP and the WTAV, we might propose that Newton's law is more complex than its alternative. If we adhered to the Bayes factor method of model selection, goodness-of-fit would be measured across the entire range of possible parameter values for m and M. For illustrative purposes, we could even assume that the two masses will always be between 0 and 1 (which extends the parallel even further). We can impose binomial sampling variability on the measurement of the masses and replicate the model selection simulations. We expect that the Bayes factor would find that Newton's law is more complex than the linear alternative. On the other hand, using a goodness-of-fit across the entire range of possible parameter values for m and M is not necessarily justified.

An alternative method of assessing model flexibility has been proposed by Dunn (2000), which measures the extent to which the prediction range of a model extends into the potential outcome space. Although this analysis seems to run counter to the results obtained by the Bayes factor method of model selection, it is not necessarily so. The Bayes factor measures the prediction range across the parameter space of each model, whereas Dunn's method selects the data points in the outcome space completely at random. In our view, Dunn's work not only supports our earlier evidence that the FLMP is not inherently capable of fitting random data, but it also adds to the arsenal of means for evaluating model flexibility.

Not surprisingly, we observed that the Bayes factor is not a panacea for model selection. As described, it is grounded in the unsubstantiated assumption that an appropriate criterion for model selection is what we have called parameter invariance—the ability to predict an observed data set across changes in the model's space of parameter values. In speech perception, we have seen that individual differences in information require the full range of parameter values and that it might not be reasonable to demand that a model give a good fit to a data set regardless of the actual parameter values that are used.

SUMMARY AND CONCLUSIONS

It should be noted that LIM refers to different models across the different assessments of flexibility. Li et al. (1996) and Dunn (2000) evaluated the FLMP and LIM in the context of the models defined by Massaro and Cohen (1993) for the four-factor experiment of Bruno and Cutting (1988). In these two cases, the LIM was a true additive four-factor model with a background source of information. The LIM used by Myung and Pitt (1997) and in the current simulations and tests was an unweighted averaging model.

Generalizing across all of the various analyses, a working hypothesis is that the FLMP is somewhat more flexible than LIM when evaluated in terms the change in a model's prediction due to variations in the model's parameters. This flexibility is taken into account in the Bayes factor method of model selection. On the other hand, The FLMP does not appear to have a meaningful advantage over the LIM in predicting random or arbitrary data. For future practice, investigators should recognize that the FLMP does have greater flexibility than the LIM and this should be accounted for in model comparisons. Either simulations such as those carried out in the present paper or the Bayes factor are reasonable methods for evaluation of this flexibility. Our analyses also revealed that the FLMP is not much more flexible than the WTAV. The extra parameter of the WTAV relative to the LIM appears to bring the linear model in the range of flexibility of the FLMP.

The adequacy of the Bayes factor is in retrospect easily understood. If two models give an equally good description of a data set, then the simpler (less flexible) one will be preferred. This outcome is reasonable. If the data set is highly variable, however, the same two models will give an equally poor description of a data set. Using the Bayes factor, the simpler (less flexible) one will be preferred. However, this conclusion could be wrong because the data themselves cannot distinguish among the models, and handicapping for flexibility could work against the correct model. Of course, the bias we have shown for the Bayes factor should not be generalized beyond the conditions we have investigated.

As to be expected from the scientific process, there are no easy answers. In a recent paper, Reinhard Selton (1998), a 1994 Nobel laureate in Economics, formulized and justified a quadratic scoring rule over a logarithmic scoring rule. The former is essentially equivalent to RMSD whereas the latter is closely related to maximum likelihood (MLE). The logarithmic scoring rule is too sensitive to small differences between very small probabilities. This reference is not meant to argue for RMSD over MLE or over Bayes factor, but to simply highlight, as in all things, there is no holy grail of model evaluation for scientific inquiry. As elegantly concluded by Myung and Pitt (1997), the use of judgment is central to model selection. Extending their advice, we propose that investigators should make use of as many techniques as feasible to provide converging evidence for the selection of one model over another. More specifically, both RMSD and the Bayes factor can be used as independent metrics of model selection. Inconsistent outcomes should provide a strong caveat for the validity of selecting one model over another.

REFERENCES

Atkinson, R. C., Bower, G. H., & Crothers, E. J. (1965). *An introduction to mathematical learning theory*. New York: Wiley.

Bruno, N., & Cutting, J. E. (1988). Minimodularity and the perception of layout. *Journal of Experimental Psychology: General*, **117**, 161-170.

Casti, J. L. (1994). *Complexification: Explaining a paradoxical world through the science of surprise*. New York: HarperCollins.

Cutting, J. E., Bruno, N., Brady, N. P., & Moore, C. (1992). Selectivity, scope, and simplicity of models: A lesson from fitting judgments of perceived depth. *Journal of Experimental Psychology: General*, **121**, 364-381.

DiCiccio, T. J., Kass, R. E., Raftery, A., & Wasserman, L. (1997). Computing Bayes factors by combining simulation and asymptotic approximations. *Journal of the American Statistical Association*, **92**, 903-915.

Dijkstra, T., & de Smedt, K. (1996). Computer models in psycholinguistics: An introduction. In T. Dijkstra & K. de Smedt (Eds.), *Computational psycholinguistics* (pp. 3-23). London: Taylor & Francis.

Dunn, J. C. (2000). Model complexity: The fit to random data reconsidered. *Psychological Research,* **63**, 174-182.

Frauenfelder, U. H., & Peters, G. (1998). Simulating the time course of spoken word recognition: An analysis of lexical competition in TRACE. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 101-146). Mahwah, NJ: Erlbaum.

Ganong, W. F., III (1980). Phonetic categorization in auditory word recognition. *Journal of Experimental Psychology: Human Perception & Performance*, **6**, 110-125.

Grainger, J., & Jacobs, A. M. (1998). On localist connectionism and psychological science. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 1-38). Mahwah, NJ: Erlbaum.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory–visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory–visual integration. *Journal of the Acoustical Society of America*, **103**, 2677-2690.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and Bayesian analysis. *American Scientist*, **80**, 64-72.

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

Li, S.-C., Lewandowsky, S., & DeBrunner, V. E. (1996). Using parameter sensitivity and interdependence to predict model scope and falsifiability. *Journal of Experimental Psychology: General*, **125**, 360-369.

Massaro, D. W. (1987). *Speech perception by ear and eye: A paradigm for psychological inquiry*. Hillsdale, NJ: Erlbaum.

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. (1976). The contribution of fundamental frequency and voice onset times to the /zi/–/si/ distinction. *Journal of the Acoustical Society of America*, **60**, 704-717.

Massaro, D. W., & Cohen, M. M. (1993). The paradigm and the fuzzy logical model of perception are alive and well. *Journal of Experimental Psychology: General*, **122**, 115-124.

Massaro, D. W., Cohen, M. M., Gesi, A., & Heredia, R. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics*, **21**, 445-478.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. T. (1995). Crosslinguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition*, **23**, 113-131.

Massaro, D. W., & Friedman, D. (1990). Models of integration given multiple sources of information. *Psychological Review*, **97**, 225-252.

Massaro, D. W., & Oden, G. C. (1995). Independence of lexical context and phonological information in speech perception. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1053-1064.

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure or behavior*. New York: Holt, Rinehart & Winston.

Myung, I. J., Forster, M. R., & Browne, M. W. (2000). [Special issue on Model Selection]. *Journal of Mathematical Psychology*, **44**(1).

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychonomic Bulletin & Review*, **4**, 79-95.

Myung, I. J., & Pitt, M. A. (1998). Issues in selecting mathematical models of cognition. In J. Grainger & A. M. Jacobs (Eds.), *Localist connectionist approaches to human cognition* (pp. 327-355). Mahwah, NJ: Erlbaum.

Pitt, M. A. (1995). The locus of the lexical shift in phoneme identification. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **21**, 1037-1052.

Raney, L., Dancer, J. E., & Bradley, R. (1984). Correlation between auditory and visual performance on two speech reception tests. *The Volta Review*, **86**, 134-141.

Selton, R. (1998). Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, **1**, 43-62.

Waldrop, M. M. (1992). *Complexity: The emerging science at the edge of order and chaos*. New York: Simon & Schuster.

NOTE

[1] These data, corresponding model fits, parameter values, and RMSDs are available at http://mambo.ucsc.edu/psl/8236/.

(Manuscript received August 25, 1999;
revision accepted for publication April 11, 2000.)