Journal of Experimental Psychology:
Learning, Memory, and Cognition
1995, Vol. 21, No. 4, 1053-1064

0278-7393/95/$3.00

OBSERVATIONS

Independence of Lexical Context and Phonological Information in Speech Perception

Dominic W. Massaro
University of California, Santa Cruz

Gregg C. Oden
University of Iowa

M. A. Pitt (1995a) studied the joint influence of phonological information and lexical context in W. F. Ganong's (1980) task. Pitt improved on earlier studies by collecting enough observations to make possible the quantitative analyses of an individual's data. The present article shows that the results of such analyses demonstrate that the integration of phonological information and lexical context is very well accounted for by the fuzzy logical model of perception (FLMP). Although Pitt concluded that the results of his research argued against the FLMP in favor of an interactive feedback system, his conclusion was based on an analysis of *transformed* results. It is argued that this use of a response transformation led to incorrect conclusions and that ultimately, models must be tested directly against observed behavior.

Pitt (1995a) studied the joint influence of phonological information and lexical context in an experimental paradigm developed by Ganong (1980). Pitt improved on earlier studies by collecting enough observations to allow a participant-by-participant evaluation of the ability of models of language processing to account for the results of this task. Previous tests of models using this task have been primarily dependent on group averages which may not be representative of the individuals that make the averages up.

However, the conclusion reached by Pitt (1995a)—that the results of his study revealed violations of the independence hypothesized by the fuzzy logical model of perception (FLMP)—is incorrect. As we show in this article, the results are, in fact, completely consistent with the FLMP and, indeed, provide strong support for it. We also examine a number of critical technical issues pertaining to the study of these questions.

The Fuzzy Logical Model of Perception

Within the framework of the FLMP, perceptual events are processed in accordance with a general algorithm (Massaro, 1987; Oden, 1979, 1984). As shown in Figure 1, the model consists of three operations: feature evaluation, feature integration, and decision. The sensory systems transduce the physical event and make available various sources of information called *features*. These continuously valued features are evaluated and matched against prototype descriptions in memory by a process that integrates individual feature values according to the specifications of the prototypes. An identification decision is then made on the basis of the relative goodness-of-match of the stimulus information with the relevant prototype descriptions. This relative goodness-of-match value thus predicts the proportion of times the stimulus is identified as an instance of the prototype or predicts a rating judgment indicating the degree to which the stimulus matches the category. A strong prediction of the FLMP is that the impact of one source of information on performance increases with increases in the ambiguity of the other available sources of information.

The FLMP provides a natural account of the integration of bottom–up and top–down sources of information in language processing. Indeed, from the beginning (see e.g., Oden & Massaro, 1978), a major attraction of this model has been its ability to account for context dependency in perception while maintaining strict independence in the basic perceptual processes. A very good example of this remarkable fact is provided by Ganong's (1980) results, which established that lexical identity could influence phonetic judgments. In Ganong's article, a continuum of test items was made by varying the voice-onset time (VOT) of the initial stop consonant of consonant–vowel–consonant (CVC) syllables. The vowel–consonant (VC) was also varied and took one of two forms. For example, participants identified the initial consonant as /d/ or /t/ in the context *__ash* (where /d/ makes a word and /t/ does not) or in the context *__ask* (where /t/ makes a word and /d/ does not). A lexical effect was observed because there were more voiced judgments /d/ in the context *__ash* than in the context *__ask*. The form of this effect has been shown to be well accounted for by the FLMP (Massaro & Oden, 1980). Ganong's original results have been replicated by several investigators (Connine & Clifton, 1987; McQueen, 1991; Pitt & Samuel, 1993). However, none of these previous experiments

Dominic W. Massaro, Department of Psychology, University of California, Santa Cruz; Gregg C. Oden, Department of Psychology, University of Iowa.

The research reported in this article and the writing of the article were supported, in part, by Public Health Service Grant PHS R01 NS 20314 and National Science Foundation Grant BNS 8812728. We thank Michael M. Cohen for assistance and Mark A. Pitt for making his data available to us.

Correspondence concerning this article should be addressed to Dominic W. Massaro, Program in Experimental Psychology, University of California, Santa Cruz, California 95064. Electronic mail may be sent via Internet to massaro@fuzzy.ucsc.edu.

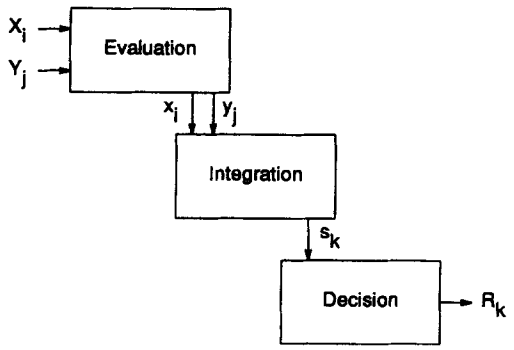

Figure 1. Schematic representation of the three stages involved in perceptual recognition. The three stages are shown to proceed left to right in time to illustrate their necessarily successive but overlapping processing. The sources of information are represented by uppercase letters (indicated by X_i and Y_j). The evaluation process transforms these sources of information into psychological values (indicated by lowercase letters x_i and y_j). These sources are then integrated to give an overall degree of support for a given alternative s_k. The decision operation maps this value into some response, R_k, such as a discrete decision or a rating.

reported results at the individual participant level. Although formal models can be tested against group results, there is always a chance that the average results do not represent the results of the individuals making up the group (Massaro & Cohen, 1993). Pitt is to be applauded for sufficiently testing individuals under each condition in order to have reliable individual results. This is essential for testing competing models and yet has been exceedingly uncommon, even though the interaction of bottom–up and top–down sources of information has been of central interest in the last decades of research on language processing.

There are two sources of information in the Ganong (1980) task: the bottom–up information from the initial speech segment and the top–down context from the following speech segment. In the framework of the FLMP, it is assumed that both of the sources are evaluated and integrated to achieve perceptual identification. Let s_i be the degree of support for the voiced alternative given by the initial segment and c_j be the support for the voiced alternative given by the following context. In this case, the total support for the voiced alternative would be as follows:

$$S(voiced \mid S_i C_j) = s_i \times c_j. \quad (1)$$

The subscript i indexes the ith level along the stimulus continuum of the segmental information and j indexes the level of the context. In the special case of just two response alternatives, the support of one source of information for one alternative can be taken to be one minus the support for the other alternative. (Nothing in what follows actually depends on this simplifying assumption.) In this case, the total support for the voiceless alternative would be as follows:

$$S(voiceless \mid S_i C_j) = (1 - s_i) \times (1 - c_j). \quad (2)$$

Given the relative-goodness rule at the decision stage, the predicted probability of a voiced response, $P(voiced \mid S_i C_j)$, is equal to

$$P(voiced \mid S_i C_j) = \frac{s_i \times c_j}{(s_i \times c_j) + [(1 - s_i) \times (1 - c_j)]}. \quad (3)$$

Test of the Model

This model was applied to the identification results of the 12 individual participants in Pitt's (1995a) Experiment 3a for which 104 observations were obtained for each data point for each participant. The points in Figure 2 give the observed results for each of the 12 participants in the task. For most of the participants, the individual results tended to resemble the average results reported by Pitt and earlier investigators. Of the 12 participants, 10 were influenced by lexical context in the appropriate direction. Participant 1 gave an inverse context effect, and Participant 7 was not influenced by context.

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 either /Ift/ or /Is/. Thus, there were six levels of bottom–up information and two contexts. A free parameter was necessary for each level of bottom–up information and for each level of contextual support. Thus, eight free parameters were estimated to predict the 12 independent data points: six values of s_i and two values of c_j. These parameters were estimated using the program STEPIT (Chandler, 1969). The parameter values in the prediction equations of the FLMP were iteratively adjusted by minimizing the squared deviations between the observed and predicted values. The program determines the set of parameter values that come closest to predicting the observed results. The goodness-of-fit of the model was given by the root-mean-square deviation (RMSD)—the square root of the average squared deviation between the predicted and observed values.

The lines in Figure 2 give the predictions of the FLMP. As can be seen in the figure, the model generally provided a good description of the results of this study. The RMSD between predicted and obtained was .017 on the average across all 12 independent individual participant fits. For the 10 participants showing appropriate context effects, the RMSD ranged from .003 to .045, with a median of .007. Thus, for each of these individuals, the model captured 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 yielded a pattern of curves in the shape of an American football, which is a trademark of the FLMP.

The only participant whose data reflected an effect of context not following this pattern was Participant 1, whose context effect went in the direction opposite to the context effect for the other participants (and opposite to reasonable expectation). The FLMP gave a very poor description of this participant's results, yielding an RMSD of .066. Although one would ordinarily not dwell on the anomalous data for a single participant, it may be worthwhile to observe that the fact that the FLMP did not provide a good fit to this participant's data is

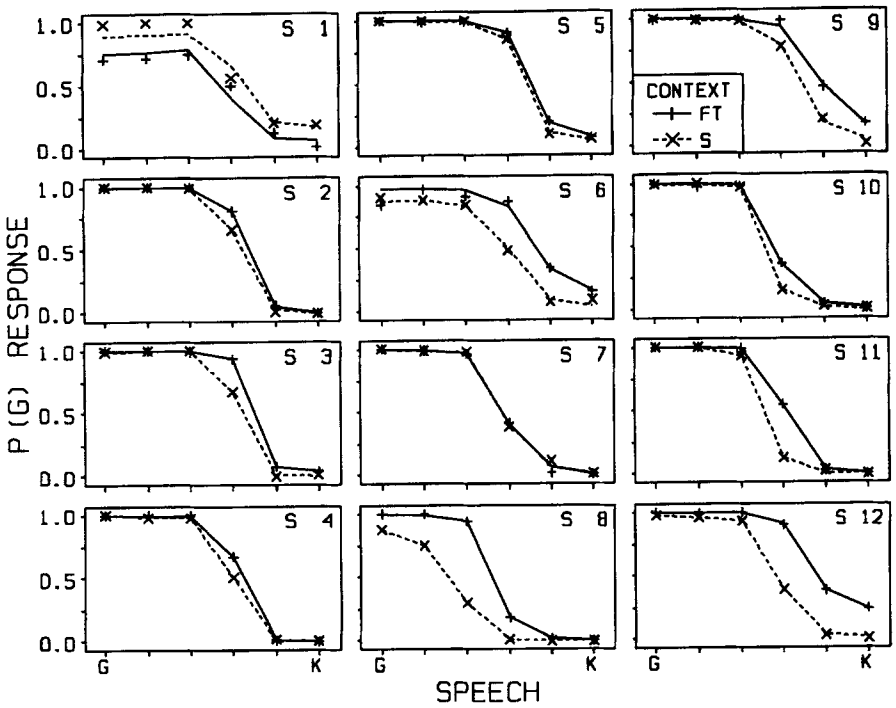

Figure 2. Observed (points) and predicted (lines) proportion of /g/ identifications for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995a) Experiment 3a. Predictions of the fuzzy logical model of perception.

evidence that the model is not so powerful that it can fit anything. The repeated success of the model appears to have led some researchers to suspect that it has excessive parameters or some other "unfair" advantage that somehow makes it effectively unfalsifiable. Such suspicions are wholly unfounded (see Massaro & Cohen, 1993). In fact, the evaluation of independence of process by means of the FLMP uses just the same number of free parameters as in the test for interaction within the analysis of variance (ANOVA). As the data for Participant 1 demonstrated concretely, it is entirely possible for the FLMP to fail to fit data; indeed, it is conceptually just as possible as it is for an interaction to be obtained with an ANOVA.

Figure 3 shows the parameter values in terms of the support for /g/ for the stimulus and context sources of information. As can be seen in the figure, the parameter values varied in a sensible way with changes in stimulus and context. The support for /g/ fell as the segment level was changed from /g/ to /k/. The support for /g/ was larger given the /Ift/ than the /Is/ context. This is important because one would not conclude that the model gave a good description of the data if it required unreasonable parameter values to do so.

Although the FLMP provided a good description of the results, it is worthwhile to know how good is good. Even if a model is perfectly correct, we cannot expect it to fit observed results perfectly. Reasonably accurate models must be stochastic or have built-in variability, as do observed results. As it is stated, the FLMP is deterministic (has no variability) at the feature evaluation and integration processes and becomes stochastic only at the decision process. The variability at the decision process is due to the relative-goodness rule in which the probability of a response is equal to the merit of that alternative relative to the sum of the merits of all relevant alternatives. For example, given a relative-goodness value of .8, that alternative is randomly chosen .8 of the time. This is analogous to flipping a coin that is biased to give a certain outcome 80% of the time. With a finite number of observations, we cannot expect that the actual probability of responding with the alternative will be exactly .8, even if the model is correct. A strong prediction of the model is that the observed variability should be equal to that expected on the basis of simple binomial variability.

The standard deviation of a binomial distribution (with two outcomes) is equal to

$$\sigma = \sqrt{\frac{pq}{N}}, \tag{4}$$

where p is the probability of one outcome, q is the probability of the other ($q = 1 - p$), and N is the number of observations. Applying this equation to the present task, p is equal to $P(/g/)$, $q = P(/k/)$, and N is 104, the number of observations at a given experimental condition for a given participant.

The benchmark RMSD may be determined by computing the binomial variance (pq/N) at each of the 12 experimental conditions, averaging these 12 values, and taking the square root. This benchmark RMSD was determined for each of the

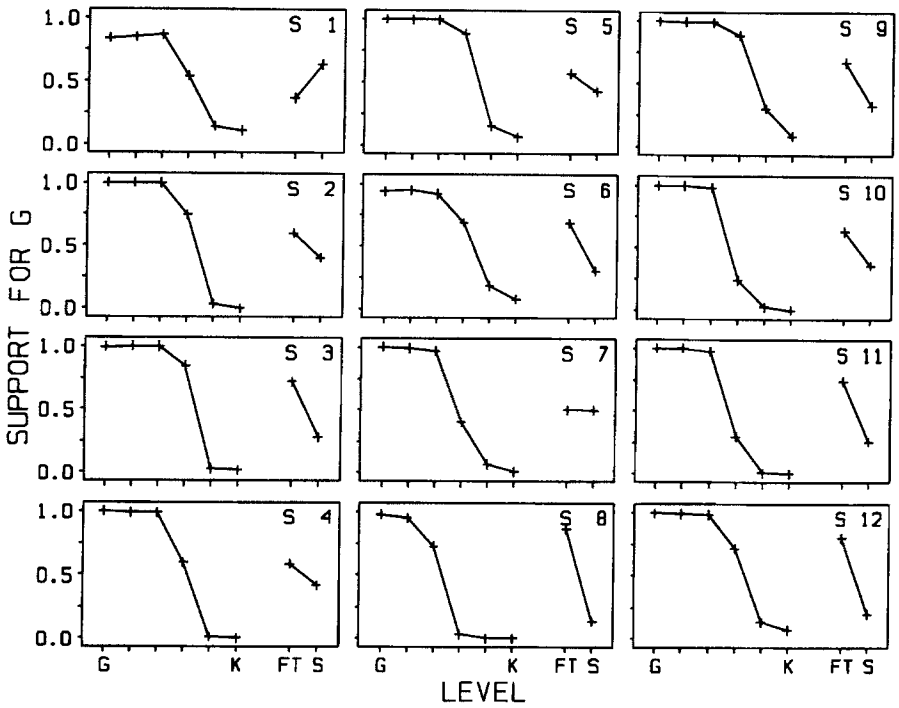

Figure 3. Parameter values for the fuzzy logical model of perception in terms of the support for /g/ for the stimulus and context sources of information.

participants and compared with the RMSD values from the fit of the FLMP to the observed results shown in Figure 2. The average of these 12 individual participant benchmarks was .024, on the same order as the observed RMSDs, which averaged .017. An ANOVA comparing the observed and benchmark RMSDs was not statistically significant, $F(1, 11) = 2.97, p = .11$. Thus, we concluded that the FLMP described the results as accurately as could be expected from a correct model.

Accounting for Other Response Forms

Although the focus in Pitt's (1995a) article, as with most other studies in this area, was on the identification response data, there were several other types of data obtained across the various experiments as well. Rating and response latency data are valuable both because they may yield converging support for conclusions and because each provides finer grained information regarding psychological processes than does the binary forced-choice response.

Ratings

In addition to the forced-choice response, participants in Pitt's (1995a) Experiment 3a were also required to rate their confidence in their decision. These confidence judgments can be treated as ratings on an interval scale to provide another test of the FLMP (see Oden, 1979). To provide this test, the six confidence categories were linearly transformed into values between zero and one, with the endpoint categories given the values 0 and 1. In this case, the ratings varied between *very sure* /k/ (0) and *very sure* /g/ (1). A mean rating was computed for each participant at each of the 12 experimental conditions. Figure 4 gives the mean ratings for the 12 participants in Experiment 3a. These mean ratings were fit by the FLMP, the predictions of which are also shown in Figure 4. The FLMP gave a good description of the results with an average RMSD across the 12 individual fits of .026. An average participant was also created by averaging the results across all 12 participants. The RMSD for the fit of the model to the average participant was .011. Thus, the FLMP is also able to describe exactly the results used by Pitt to compute the A_g (sensitivity) measures that were argued to provide evidence against the model.

Absolute Identification

Pitt's (1995a) Experiment 3b required participants to identify absolutely the test stimulus as one of four responses. We analyzed the results in two ways. We computed the proportion of /g/ responses by calling the two most /g/-like responses /g/ and the other two responses /k/. We also computed a /g/-ness rating by transforming the absolute judgments into a value between zero and one, as we did in our analysis of the results for Experiment 3a. The individual results of the 3 participants were fit by the FLMP. Figure 5 gives the observed proportion of /g/ along with the predictions of the FLMP. For the proportion measure, the average RMSD for the 3 participants was .007. Figure 6 gives the rating of /g/ along with the predictions of the FLMP. For the rating measure, the average RMSD was .042. Thus, the FLMP also gives a fairly good description of individual results recorded in an absolute identification task.

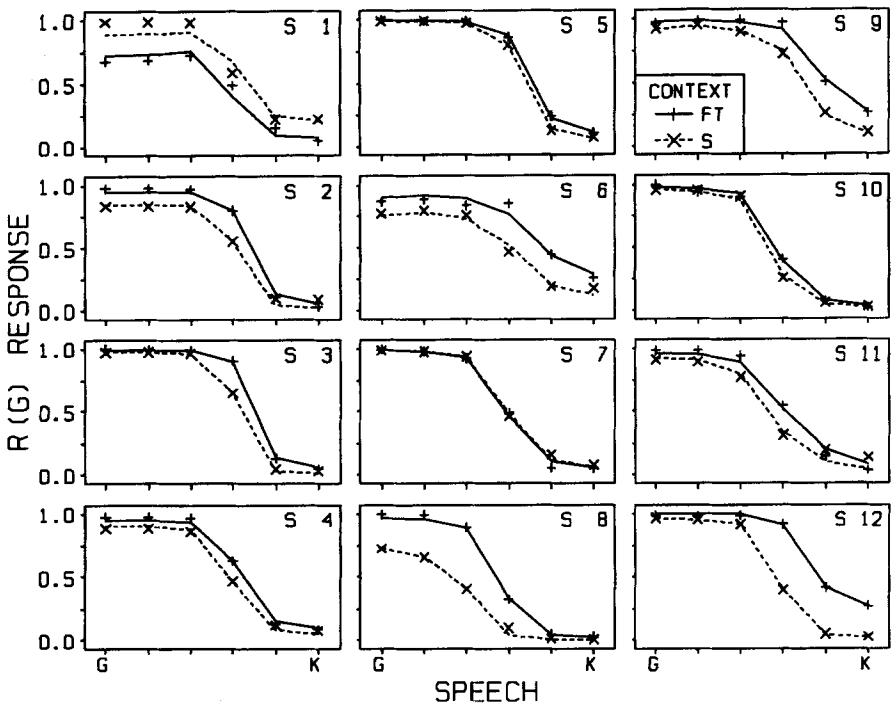

Figure 4. Observed (points) and predicted (lines) rating of /g/ for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995a) Experiment 3a. Predictions of the fuzzy logical model of perception.

Lexical Influence as a Function of Reaction Time (RT)

The FLMP is also consistent with the important finding that the size of the lexical context effect was positively correlated with RT. As shown in Pitt's (1995a) Experiment 1 and his Figure 2, the contribution of lexical context increased as the RT to make the judgment increased. In the framework of the FLMP, bottom–up phonological information and top–down lexical information are processed over time. The FLMP predicts that support for /g/ from the bottom–up and top–down sources grows with processing time. However, given test items such as /gIft/ or /kIs/, the lexical information occurs later in the test word than the phonological information in the initial consonant. Therefore, there is necessarily a delay in the arrival of the lexical information relative to the stimulus information from the initial consonant. Thus, there is a delay of the top–down influence relative to the bottom–up influence, but both functions grow with increases in processing time (see Massaro & Cohen, 1991; Massaro & Cohen, 1994).

To provide a quantitative test of this interpretation, the FLMP was fit to the data for the 14 individual participants and to the average results from Pitt's (1995a) Experiment 1. The model was fit to the three identification conditions in the slow, medium, and fast RT conditions (Pitt, 1995a, Figure 2). It was assumed that the available processing time differed in the three conditions. The available processing time was taken to be the mean RT for each participant under each of the three conditions. The mean RTs across the 14 participants were 345, 445, and 633 ms. The mean RTs for individual participants were taken to be the available processing time for the

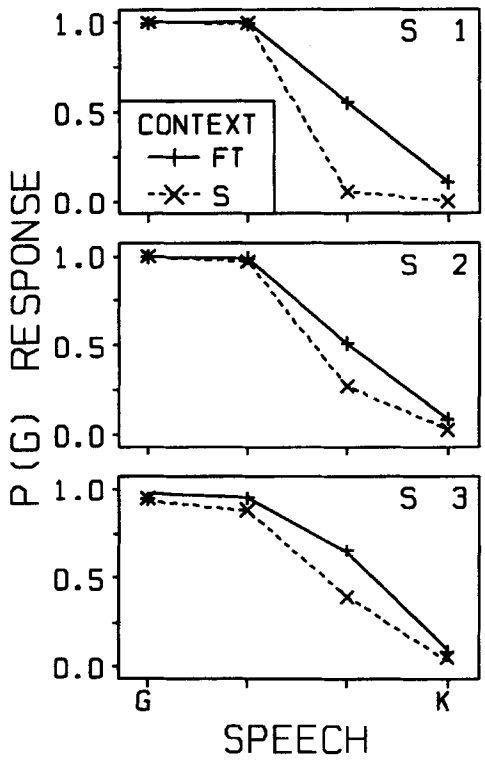

Figure 5. Observed (points) and predicted (lines) identifications of /g/ for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995a) Experiment 3b. Predictions of the fuzzy logical model of perception.

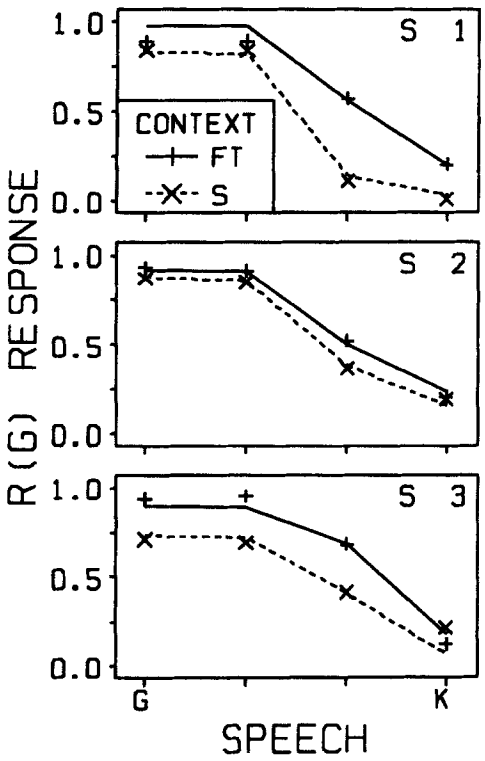

Figure 6. Observed (points) and predicted (lines) ratings of /g/ for FT and S contexts as a function of the speech information of the initial consonant. Results from Pitt's (1995a) Experiment 3b. Predictions of the fuzzy logical model of perception.

phonological information. These same values minus a constant delay c_l for the arrival of the lexical information were taken to be the available processing times for the lexical information.

We assumed that evaluation of the speech information from the initial segment followed the function given by Equation 5, which gives the amount of support, s, for /g/, defined as $S(s)$:

$$S(s) = \alpha(1 - e^{-\theta t}) + .5(e^{-\theta t}). \quad (5)$$

Equation 5 describes the evaluation process as a negatively accelerated growth function of processing time t. It is assumed that the information provided by this source of information can be represented by α; θ, the rate of processing this information, is independent of the information value. The value α represents the asymptotic support for the alternative /g/. These α values are equal to the parameter values indicating the degree of support in the typical fit of the model when processing time is not a factor. If α corresponds to the amount of support for the voiced alternative /g/, then changing the speech syllable from a somewhat ambiguous /g/ to a less ambiguous /g/ would yield a larger α value but would be processed with a fixed θ. As noted above, different processing times would be available for the three RT conditions.

The same type of equation would describe the growth of the lexical context information c, except that the constant delay for its arrival would be subtracted from the available processing time.

Integration of the outputs of evaluation is assumed to occur continuously, producing an overall goodness-of-match value for each of the test alternatives. Given two sources of information, the output of integration, $S(s, c)$, is given by Equation 6:

$$S(s, c) = [\alpha_s(1 - e^{-\theta t}) + .5(e^{-\theta t})] \times [\alpha_c(1 - e^{-\theta(t-c_l)}) + .5(e^{-\theta(t-c_l)})], \quad (6)$$

where α_s is the asymptotic support from the initial consonant stimulus, and α_c is the asymptotic support from the lexical context. The value of θ is the same for both the stimulus information and the lexical context, and c_l is the delay between the onset of the processing and the onset of the lexical information. The relative-goodness rule is simply instantiated when the feature evaluation and integration are completed, as constrained by the RT condition. The same operations occur at all three RT conditions; only the available processing time differs across the conditions.

To provide a baseline for the fit of this dynamic FLMP, the standard (static) FLMP was first fit to each of the three RT conditions separately, with a unique set of 10 parameters for each RT condition. The average RMSD for the fit of the data for the 14 individual participants was .017, whereas the RMSD for the fit of the data for the average participant was .010. Thus, the static FLMP provides an excellent description of the results, consistent with the good fit of the data for the participants in Pitt's (1995a) Experiments 3a and 3b.

The fit of the dynamic FLMP reduced the total number of free parameters required from 30 to 12. In this case, the FLMP was being tested against three times as much data (16 vs. 48 observations), with only 2 additional parameter values (θ corresponding to the rate of processing and c_l corresponding to context delay). Figure 7 gives the predictions of this dynamic FLMP to the average results. As can be seen in the figure, the dynamic FLMP nicely described the increase in the influence of lexical context with increases in processing time. The average RMSD for the fits of the data for the 14 individual participants was .046, whereas the RMSD for the fit of the data for the average participant was .020. Considering the savings in the number of free parameters, the dynamic FLMP did a respectable job of describing the results. The average value of θ was 8.59, and the average value of context delay c_l was 297 ms. It is encouraging that the value of c_l was a reasonable estimate of the time between the onset of the syllable and the onset of the lexical context. The good description given by the dynamic FLMP reveals that the influence of lexical context as a function of RT was parsimoniously described as being due to processing time. Postulating different strategies or some other qualitative influence was not necessary.

The FLMP might also explain the identification results across the RT partitioning in Pitt's monetary bias conditions. Because the monetary bias was known before the test stimulus was presented, its effect should have been greatest at the shortest RTs. With longer RTs, the stimulus was processed more completely and produced a bigger influence on performance. This prediction is consistent with the observed results.

Pitfalls in Model Testing

The success of the FLMP in accounting for the variety of Pitt's (1995a) data is no small accomplishment, especially because we have shown it to capture the precise quantitative

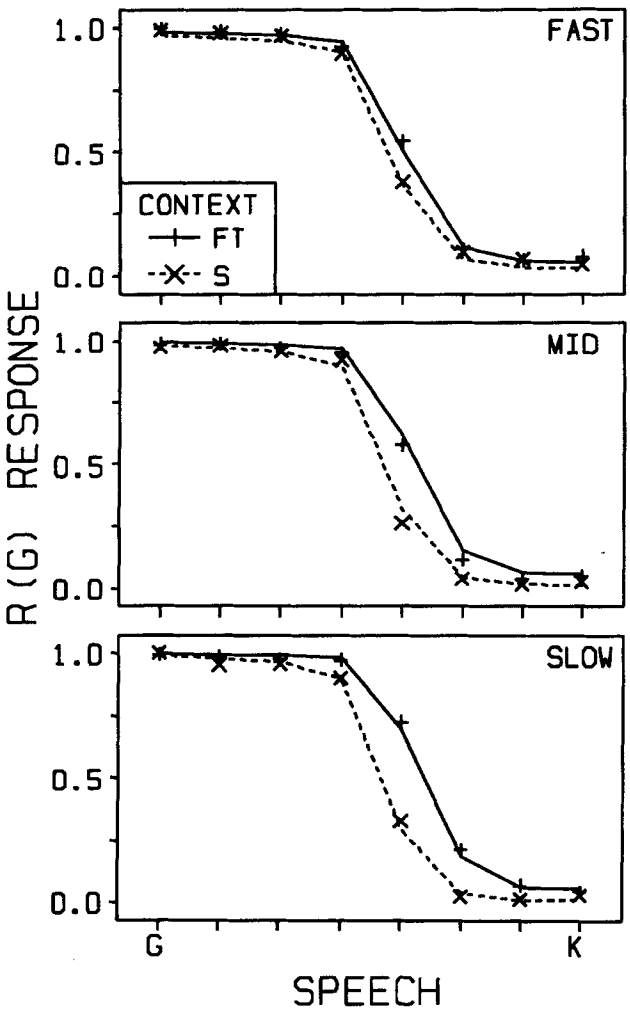

Figure 7. Observed (points) and predicted (lines) proportion of /g/ identifications for FT and S contexts as a function of the speech information of the initial consonant. The three panels give the results for the fast, medium (mid), and slow identifications.

pattern of individual participants' responses. How then, one might ask, was Pitt led to conclude that the results of his study falsified the independence prediction of this same model? Indeed, the reader of that article may believe that he or she has seen substantial signs of interactivity in the data Pitt presented, such as that in his Figure 3. However, as we have seen, this apparent evidence of violations of independence is illusory. Furthermore, as we now show, this illusion appears to be the consequence of at least three problems: (a) focusing on differences between differences, (b) using a suboptimal parameter estimation procedure, and (c) using an inexact stand-in for the model. Each of these problems contributes to producing the spurious appearance of substantial violations of independence. We now consider each of these culprits in turn.

Differences of Differences

The critical figures in Pitt's (1995a) article plot differences of z scores (or A_g scores) for adjacent stimuli under each respective context condition. The crucial comparisons are thus given in terms of differences in these differences as a function of context. This is a valid way to address whether there are any Context × Stimulus Level interactions, but it tends to amplify the appearance of any such interactions because the overall range is greatly reduced. To illustrate this point, we plot in the left panel of Figure 8 some stepwise differences of z scores based on the data[1] from Pitt's Experiment 3a that show the crossover interaction that is characteristic of the data that he believes support interactivity. In the right panel of Figure 8, the same information has been replotted in the more usual fashion for considering interactions, that is, in terms of the z scores directly. This change by itself does not make the interaction go away, of course, but it shows it not to be as monumental as one might have been led to suppose from the original figure. The curves are mostly parallel, which means that there is really only a small interaction for which we must still account.

Suboptimal Parameter Estimation Procedure

The parameter estimation procedure implicit in the d' (and A_g) analysis used by Pitt (1995a) was suboptimal in the sense that it did not provide the best possible fit to the data of the model on which the signal detection analysis was based (TSD—the theory of signal detectability), let alone of the FLMP. It therefore does not really constitute a fair test of the question of independence. Discrepancies from the model (apparent violations of independence) may not be the fault of the model but rather of the estimation procedure. The problem is that these analyses do not constitute least squares estimation procedures in the sense that really counts; that is, they are not least squares estimates in terms of the actual data being evaluated (i.e., the response proportions).

This may perhaps be more easily understood by direct example. The top section of Table 1 provides the proportions corresponding to the z scores plotted in Figure 8. The next three sections of the table detail Pitt's (1995a) z-score analysis in terms of the equivalent additive model fit to the z scores. First, we provide the z scores (standard normal deviates) corresponding to the original proportions along with the row and column means of these z scores. The row means have been adjusted by subtracting out the grand mean to express the main effect for rows as the comparable amount above and below the grand mean. For each cell of the matrix, the sum of the respective column mean and the respective adjusted row mean provides the predicted z score for the linear model as fit to the z-score transformed data.[2] These values are shown in the next two rows in this section of Table 1. Finally, these values are converted back to proportions using the inverse of the cumulative standard normal function (i.e., the values used to assess independence when the z-score analysis is employed). The

[1] Just how these values are based on the data is explained in a subsequent section. For now, it does not matter what these values are; what matters is that they display the general pattern that Pitt obtained in his studies.

[2] This is simpler than but equivalent to the more usual form of the linear model of the ANOVA, which is (row mean − grand mean) + (column mean − grand mean) + grand mean.

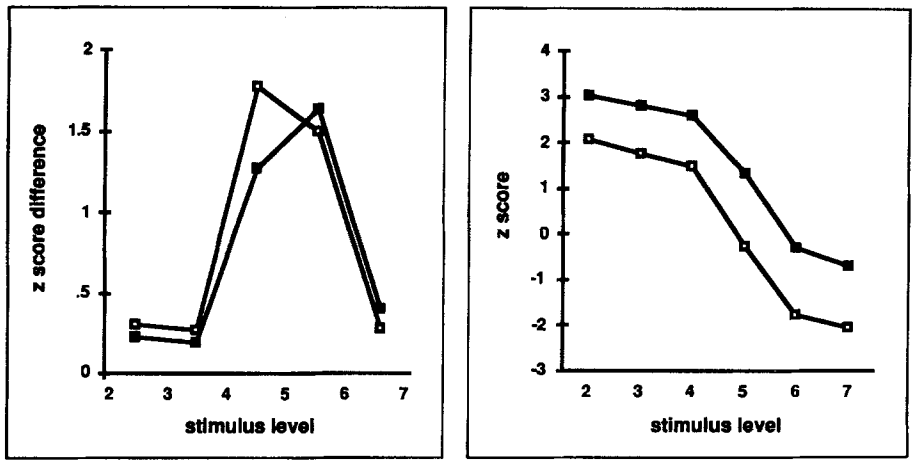

Figure 8. The left panel shows a set of *z* scores plotted as differences between adjacent stimulus levels, as in the fashion of Pitt (1995a). The right panel shows the same *z* scores plotted directly in the more usual way. The filled squares correspond to the FT context, and the empty squares to the S context. The stimulus level goes from /g/ to /k/. See text for explanation.

RMSD between these predictions and the original proportion values was .026. Even with the suboptimal estimation procedure, this RMSD was reasonably small, which indicates that we were correct in concluding on the basis of visual inspection of the right panel of Figure 8 that the interaction there (and, hence, the corresponding interaction in the left panel of Figure 8) was really not very substantial.

However, is the outcome of this analysis really as well as the TSD version of the independence model can do? Would it be possible to find parameter values for the additive *z*-score model that would fit the data better (i.e., yield a smaller RMSD value) than do the marginal means of the *z* scores? The answer is most definitely yes: The point we are presently making is that this analysis is not a least squares procedure in the data and that it thereby still underestimates the degree to which the data can be accounted for by the hypothesis of independence. This is illustrated in the fifth and sixth sections of Table 1, which contain the predicted *z* scores we obtained by estimating row and column parameters corresponding to those used above through an iterative procedure (in this case, the Solver routine in Excel 4.0) that minimized the resulting RMSD when the predicted *z* scores were used to produce the predicted proportions. The parameter values obtained by these means are shown along with the resulting predicted *z*-score and proportion values. These values yielded an RMSD of .004, which constitutes a dramatic improvement over that obtained by the suboptimal estimation implicit in Pitt's (1995a) *z*-score analysis.

Comparing the predicted proportions obtained by these two procedures readily revealed the superiority of fitting the data themselves rather than some nonlinearly transformed scores. Note particularly that the data for the critical fifth stimulus level (.909 and .390 for the two contexts, respectively) is now well accounted for (.905 and .392), whereas it was not at all well accounted for using the suboptimal procedure (.877 and .460). Of course, the fit has gotten slightly worse in a few cells, namely for some of those in the tails of the normal distribution. This illustrates the problem of doing least squares estimation in nonlinearly transformed data: The suboptimal procedure was willing to trade off a much worse fit in the middle of the proportion scale for piddling gains in the tails because extreme proportions are overemphasized when transformed to *z* scores (a small difference in *p* values in the tails leads to big differences in the corresponding *z* values).

Note that up to this point, we are still using the TSD model. Thus, there is nothing in what has just been demonstrated that depends on the FLMP. Rather, simply applying a more appropriate parameter estimation procedure with the TSD model has caused the apparent violations of independence to all but evaporate.

The Normal Instead of the Logistic

It may seem to be overkill, but there is still one more important matter to consider: the exact technical consequences of using the TSD to stand in for the FLMP. For present purposes, the only real functional difference between TSD and FLMP is in the transformation that converts the model to linear form. Indeed, if one were to apply the TSD model using the logistic instead of the normal probability distribution, then it would become mathematically isomorphic to the FLMP under the present conditions. It is well-known that the logistic is closely related to the normal in form; for many purposes, they can be treated as being practically interchangeable. This was the basis of Massaro's (1989b) use of the *z* transform with the FLMP.

Nevertheless, the difference between the normal and the logistic may sometimes be of real importance. A detailed analysis of Pitt's (1995a) data has revealed that this situation is just such a case. The simplest concrete demonstration of this is given by comparing the fit to the data of the TSD model to that of the FLMP using the optimal estimation procedure in both instances. For example, in the last section of Table 1, we

Table 1
Original Proportions and z Scores Used to Illustrate the Pitfalls in Pitt's (1995a) Analyses

Proportions and z scores	Stimulus level					
	2	3	4	5	6	7
Original proportions						
/Ift/context	.999	.997	.995	.909	.383	.242
/Is/context	.981	.961	.933	.390	.038	.020
z scores corresponding to original proportions						
/Ift/context[a]	3.026	2.799	2.608	1.338	−.297	−.701
/Is/context[b]	2.072	1.764	1.498	−.280	−1.774	−2.056
M	2.549	2.281	2.053	.529	−1.035	−1.379
Predicted *z* scores from LSF of linear model to *z* scores[c]						
/Ift/context	3.178	2.910	2.682	1.158	−.406	−.750
/Is/context	1.920	1.652	1.424	−.100	−1.665	−2.008
Predicted proportions						
/Ift/context	.999	.998	.996	.877	.342	.227
/Is/context[d]	.973	.951	.923	.460	.048	.022
Predicted *z* scores from direct LSF of TSD model to proportions						
/Ift/context[e]	3.655	3.347	3.081	1.309	−.293	−.698
/Is/context[f]	2.071	1.763	1.497	−.274	−1.877	−2.282
Column parameters	2.863	2.555	2.289	.517	−1.085	−1.490
Predicted proportions						
/Ift/context	1.000	1.000	.999	.905	.385	.243
/Is/context[g]	.981	.961	.933	.392	.030	.011
Predicted proportions from direct LSF of FLMP to proportions						
/Ift/context[h]	.999	.997	.995	.909	.383	.242
/Is/context[i]	.981	.961	.933	.390	.038	.020
Column parameters[j]	.995	.990	.982	.717	.136	.074

Note. LSF = least squares fit.
[a]Adjusted row mean = .629. [b]Adjusted row mean = −.629. [c]Predicted *z* scores = column mean + adjusted row mean. [d]Root-mean-square deviations (RMSD) = .026. [e]Row parameter = .792. [f]Row parameter = −.792. [g]RMSD = .004. [h]Row parameter = .799. [i]Row parameter = .201. [j]RMSD = .000.

present the results of fitting the FLMP to the original proportions using the same iterative estimation procedure as was used with the TSD model (presented in the fifth and sixth sections of the table). That is, the procedure searched for parameter values for the FLMP that would minimize the RMSD between predicted and original proportions. The predicted proportions are listed in the table along with the parameter values on which they are based. (Note that in this case, as described earlier, the two row parameters sum to one rather than zero.) As can readily be seen, the FLMP fits the original values even better that the TSD did. In fact, the predictions are identical to the original values listed in the table and, so, the RMSD is exactly zero! Thus, here is a case where the fit of the TSD model is only an approximation of the degree to which the FLMP can account for the data. That is, the difference between the normal and the logistic really does make a difference.

As the reader has undoubtably by now surmised, the values plotted in Figure 8 (and the corresponding proportions listed in the first section of Table 1) were computed from the predictions of the FLMP when fit to the data for one of Pitt's (1995a) participants (Participant 12—chosen for having healthy effects of both factors). Thus, throughout this section, we have been dealing with values that followed the form of the FLMP perfectly. This has been done to emphasize the finding that of the procedures for evaluating this model, only a direct fitting of the model itself to the data itself provided a real test. In the course of doing so, we have also demonstrated a rather startling result: Applying the *z*-score analysis to values that perfectly follow the FLMP can yield results of exactly the form that Pitt takes as evidence against the model. The pattern of the plot in the left panel of Figure 8 is entirely spurious as an indication of interactivity—it results from applying the *z*-score analysis to values that perfectly follow the FLMP and, thus, are purely noninteractive.

It should be stressed that it is not just that Participant 12's values show that this state of affairs is *possible.* This pattern generally holds for the other individual participants as well. In fact, if we take the separate predictions of the FLMP when fit to the individual data for each of the 12 participants in Experiment 3a and apply the *z*-score analysis to each, followed by the *t* test that Pitt (1995a) computed, we find that the differences for the 4–5 and 5–6 steps are significant, just as Pitt did. Furthermore, in Monte Carlo studies that involved adding appropriate binomial noise to the predictions of the FLMP for each of the 12 participants, we nearly always found the same pattern of the interaction, and we obtained significant *t* tests for Steps 4–5 and 5–6 as often as not.

All of this section has focused on the *z*-score analysis based on TSD. As it turns out, exactly the same considerations apply

to the A_g analysis, even though it is taken to be a nonparametric measure. Figure 9 plots the pattern that results when the A_g values were computed on the basis of the predictions of the FLMP for each participant and then averaged across participants.[3] Again, the resulting pattern was highly similar to that actually obtained by Pitt, and the relevant comparisons were significant across participants both for the specific case when the data followed the FLMP exactly and also in the majority of Monte Carlo trials when binomial noise was added.

Signal Detection Paradigm

So what, then, does all of this say about the use of the signal detection model as a means for separating perceptual from postperceptual processes? Although the signal detection paradigm has been an extremely useful tool for psychologists, it has also on occasion been misleading in terms of how dependent measures are interpreted. Contrary to intuition, we showed that dependent measures are not so easily equated with perceptual and nonperceptual processes. Pitt (1995a) also committed this error by equating the sensitivity measure d' with perception and the decision bias measure with postperceptual guessing. It is now clear that the bias measure can also reflect perceptual processing and that it is necessary to distinguish between two types of bias (Massaro & Cowan, 1993). The first, called a *belief bias,* refers to the perceptual interpretation of the stimulus. The second, called a *decision bias,* refers to the participant's inclination to respond, given the payoff contingencies. In these terms, we hypothesize that lexical context and monetary payoff are necessarily different in their influence on performance. Lexical context may reasonably be viewed as belief bias and monetary payoff as decision bias.

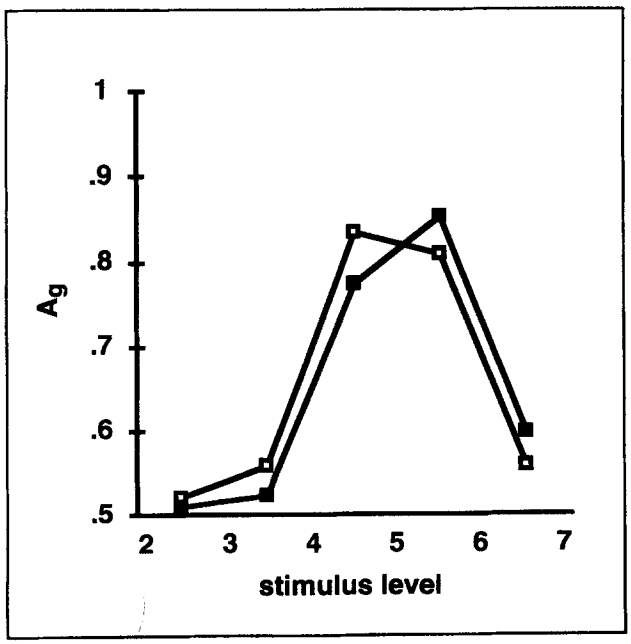

Figure 9. Mean A_g values between adjacent stimulus levels computed from predictions of the fuzzy logical model of perception for individual participants. The filled squares correspond to the FT context and the empty squares to the S context. The stimulus level goes from /g/ to /k/. *See text for explanation.*

Pitt's (1995a) mistake in this regard is that he believes only a change in sensitivity can reflect perceptual processing. As emphasized in Massaro's (1989b) use of signal detection, however, the FLMP predicts an effect of lexical context on bias, not sensitivity. As emphasized even more strongly, this bias effect is interpreted as a true perceptual effect. To demonstrate that a true perceptual effect can be reflected in a belief bias and produce results consistent with the independence assumption of the FLMP, we refer the reader to the results of a series of experiments on speech perception by eye and ear. As has been well documented (Massaro, 1987, 1989a), a speaker's face has a robust effect on the perception of the corresponding auditory speech. As an example, the auditory syllable /ba/ paired with the mouth movement corresponding to /da/ is heard as the syllable /va/ or /ða/ (Massaro & Cohen, 1990; McGurk & MacDonald, 1976). This is a truly perceptual effect, as everyone who has experienced the effect first-hand knows. In fact, it is not possible to filter out consciously the influence of the visible speech. Participants instructed to look at the face but to report only the sound showed about the same influence (Massaro, 1987, chapter 3, Figures 9 and 12). Given the strong evidence that the influence of visible speech is indeed perceptual, it is important to note that this influence has been consistently well described by the FLMP (Massaro, 1987; Massaro & Cohen, 1990).

Pitt (1995a) appeared to find similar effects of both lexical context and monetary payoffs on response probability. Thus, it is important to find some other measure that could distinguish between these two influences. Connine and Clifton (1987) studied the effects of both lexical context and monetary payoff in the Ganong (1980) task. The lexical contribution occurred only within the ambiguous range of the segmental information. A monetary payoff scheme was imposed only on nonwords to bias the participants to respond with one alternative or the other. These results replicated those found with lexical context. Given just the response probabilities, there was no evidence that these manipulations reflected different types of bias. However, the pattern of RTs for the two tasks differed even though the response probabilities did not. Given a lexical context, the RTs of word judgments were faster only for speech stimuli that gave a lexical context effect on response probability. The monetary payoff produced RTs that were always faster for the bias-consistent alternative even for speech stimuli that

[3]To compare A_g, it was necessary to extend the FLMP so that it would generate predicted confidence ratings. We simply presumed that the ratings corresponded to response bins separated by five equally spaced (on the average) cutoffs, each of which varied from occasion to occasion according to a normal distribution with constant variance. We further presumed that the middle cutoff corresponded to the point that separated the /k/ responses from the /g/ responses in the identification task and that the spacing between the cutoffs was a half of a standard deviation. Informal exploration of variations on these presumptions indicates that the specifics do not matter much; of course, a more systematic exploration might yield an even more impressive demonstration, but the present one seems to make the point sufficiently well.

gave no effect of payoff on response probability. Thus, the RTs were consistent with the hypothesis that the lexical context induced a belief bias and the monetary payoff a decision bias.

Pitt (1995a) replicated the Connine and Clifton (1987) study and looked to the d' analysis for differences between monetary bias and lexical bias. The differences he observed were taken as critical evidence against models that do not assume interactive activation. Although both types of bias appeared to produce similar shifts in the identification functions, they putatively produced different patterns in the d' analyses.

However, once again this difference appears to be illusory. When we applied the FLMP to the data for this study, we obtained fits that were comparably good to those that were obtained with the lexical context manipulation. The RMSD between predicted and obtained was .013 on the average across all 12 independent individual participant fits. The RMSD ranged from .001 to .028 across the 12 fits. In terms of the FLMP, both lexical context and monetary payoff reflected independent influences on performance. However, we viewed the lexical influence as perceptual (belief bias) and the monetary influence as decision bias. Some evidence for this distinction comes from the RT analysis in both the Connine and Clifton (1987) and Pitt (1995a) studies. In Pitt's study, for example, the influence of lexical context increased systematically with increases in RT (see Figure 7), whereas this was not the case for the influence of monetary payoff. This difference in the effects of monetary payoff and lexical context supports our view that they represent two different types of influence, even though the FLMP gives an equally good description of both.

Pitt's Reply

Although Pitt (1995b) will have the final word in this issue, we have been allowed to respond briefly to his reply to allow the reader to be better informed. Pitt implies that only the poor fit to Participant 1's data shows that the FLMP cannot fit any set of data. However, he does not reference Massaro and Cohen's (1993) article, in which we substantiate the falsifiability of the FLMP. On the basis of an article by Crowther et al. (1995), Pitt claims that the "FLMP's power may lie in its ability to generate equally good fits ... with different parameter settings" (p. 1,037) and "This property of the model may blur its usefulness" (p. 1,037). However, neither of these claims were made by Crowther et al. One of our main points is that the d' transformation is the incorrect one for testing the FLMP. Therefore, we believe that his claim that the plots of d' in Figures 1 and 2 of his reply somehow test the FLMP is categorically false.

Pitt (1995b) stated that "Massaro and Oden seem to place more faith in the identification data than in the detection data" (p. 1,038). However, the "detection data" are simply transformations of the identification data, but his statement implies they are different things. He also stated, "This seems to be a change from past practices." (p. 1,038) and refers to Massaro (1989). This completely misrepresents Massaro's 1989 paper. First, Massaro (1989) did test the FLMP against the identification data. Second, the detection transformation was used primarily to test the TRACE model, not the FLMP. In the follow-up paper (Massaro & Cohen, 1991), the FLMP and TRACE were tested against only identification data. Our past practices are consistent with our recommendation that models should be tested against observed behavior, if possible, and not simply against transformed results.

Conclusion

We have shown that, contrary to his conclusions, Pitt (1995a) has presented some of the most convincing evidence to date for the independence of stimulus information and context in speech perception. The FLMP gave a good description of the joint contribution of stimulus information and lexical context on the judgments of individual participants. In addition, the model provided reasonable accounts for the results involving a variety of response measures and manipulations. Once again, careful examination of data has found context-dependent perception to be well explained in terms of stimulus information and contextual information making independent but joint contributions to word recognition. Top–down effects on sensitivity have yet to be convincingly demonstrated; there continues to be no reason to believe that top–down activation of lower level representations plays any role in perception.

References

Chandler, J. P. (1969). Subroutine STEPIT—Finds local minima of a smooth function of several parameters. *Behavioral Science, 14,* 81–82.

Connine, C. M., & Clifton, C. (1987). Interactive use of lexical information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 13,* 291–299.

Crowther, C. S., Batchelder, W. H., & Hu, X. (1995). A measurement-theoretic analysis of the fuzzy logical model of perception. *Psychological Review, 102,* 396–408.

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

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

Massaro, D. W. (1989a). Multiple Book Review of *Speech perception by ear and eye: A paradigm for psychological inquiry. Behavioral and Brain Sciences, 12,* 741–794.

Massaro, D. W. (1989b). Testing between the TRACE model and the fuzzy logical model of perception. *Cognitive Psychology, 21,* 398–421.

Massaro, D. W., & Cohen, M. M. (1990). Perception of synthesized audible and visible speech. *Psychological Science, 1,* 55–63.

Massaro, D. W., & Cohen, M. M. (1991). Integration versus interactive activation: The joint influence of stimulus and context in perception. *Cognitive Psychology, 23,* 558–614.

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. (1994). Visual, orthographic, and lexical influences in reading. *Journal of Experimental Psychology: Human Perception and Performance, 20,* 1107–1128.

Massaro, D. W., & Cowan, N. (1993). Information processing models: Microscopes of the mind. *Annual Review of Psychology, 44,* 383–425.

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 3, pp. 129–165). New York: Academic Press.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264,* 746–748.

McQueen, J. M. (1991). The influence of the lexicon on phonetic categorization: Stimulus quality in word-final ambiguity. *Journal of Experimental Psychology: Human Perception and Performance, 17,* 433–443.

Oden, G. C. (1979). A fuzzy logical model of letter identification. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 336–352.

Oden, G. C. (1984). Integration of fuzzy linguistic information in language comprehension. *Fuzzy Sets and Systems, 14,* 29–41.

Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review, 85,* 172–191.

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

Pitt, M. A. (1995b). Data fitting and detection theory: A reply to Massaro and Oden (1995). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21,* 1065–1067.

Pitt, M. A., Samuel, A. G. (1993). An empirical and meta-analytic evaluation of the phoneme identificaton task. *Journal of Experimental Psychology: Human Perception and Performance, 19,* 699–725.

Received September 12, 1994
Revision received November 21, 1994
Accepted November 23, 1994 ■