Mathematical Social Sciences 23 (1992) 199–219
North-Holland

Alternative biased choice models

J.E. Keith Smith

University of Michigan, Ann Arbor, MI 48103, USA

Communicated by A.A.J. Marley
Received 10 August 1990
Revised 27 June 1991

Luce's Biased Choice Model has never had a serious competitor as a model of identification data. Even when it has provided a poor model of such data, other models have done even less well. Two alternative models are presented and the three are fit to a published data set. One alternative model is very much like the Biased Choice Model, differing only in the way it treats response bias. It uses an ordinal assumption about the biases and might be called the Triangular Bias (TB) model. The Guessing Mixture Model (GMM) is quite different, although it too uses the concepts of bias and similarity. It posits that the observed confusion matrix is a probability mixture of two latent matrices, the one involving only similarity, not bias, while the other involves bias, not similarity.

Illustrative data, a confusion matrix based on four stimuli constructed by crossing two binary features, can be naturally described in three hierarchical ways. The most general description ignores the feature structure of the stimuli. The next description, the feature pattern model, assumes that similarity depends only on the *pattern* of feature differences, and the simplest special case assumes that similarity depends only on the product of similarities from each of the features.

For the general description the three models are not strikingly different, with the Biased Choice Model fitting least well, followed by GMM, with TB the winner. For the independent feature form, however, the GMM model fits much better than either of the others. Indeed, the independent feature model cannot be rejected at the 10% level using GMM, even though the sample of data is large.

Key words: Confusion matrices; identification experiments; log-linear models; biased choice; EM algorithm.

1. Introduction

A choice experiment in which an identification function exists, a function mapping each stimulus into a unique response, is called an identification experiment (Luce, 1963). Many models of identification behavior have been put forward and certainly many more will find their way into the literature, but so far none has had more success than Luce's Biased Choice Model (BCM) (1963), also known as the 'similarity choice model' (Townsend and Landon, 1982). It has not always fit identification data, but when it has not neither has any other model.[1]

Special cases of the BCM are quite easy to think of but they seem seldom to be

[1] Comparing the fits of models that are not hierarchically related can be hazardous. Comparisons in this paper will be based on likelihood statistics. Other criteria, such as the unweighted least squares criterion used by Ashby, Lee and Balakrishnan (1991) may order models differently.

adequate to fit the data in question even when the general BCM fits. Indeed, when a special case fits for one subject it often does not fit again, for another subject, for another session, or for another condition.

In this paper I propose two different models which are alternatives to the original Luce formulation, rather than special cases of it. Both are first cousins but are different enough from the BCM to provide considerably different fits.

Before proceeding further we should all understand that the experimental paradigm we are discussing involves presenting to the subject a stimulus from an over-learned stimulus set, but so degraded that it is rather often misidentified as one of the other stimuli in the set. The data are analyzed in the form of a confusion matrix with rows representing the stimuli presented and columns representing response labels of the stimuli the subject saw or thought he saw, or for lack of any other reason, he guessed were presented.

Luce's Biased Choice Model is elegantly simple. It states that the probability of using label j when stimulus i was presented (p_{ij}) is proportional to the similarity of stimulus i to stimulus j (η_{ij}) and to the popularity of, or bias toward, using label j (β_j). Formally,

$$p_{ij} \sim \beta_j \eta_{ij}. \tag{1}$$

Without loss of generality one can define the similarity of a stimulus to itself as 1 (see Appendix A for proof). In addition the model would not be identifiable without the natural restriction that similarity be a symmetric measure,[2] that is that stimulus i is precisely as similar to stimulus j as stimulus j is to stimulus i. Formally,

$$\begin{aligned} &\eta_{ii} = 1, \quad \forall i, \\ &\eta_{ij} = \eta_{ji}. \end{aligned} \tag{2}$$

Finally, the β parameters are necessarily non-negative and homogeneous so that one is free to normalize them to add to one or to pick one as a standard and scale all the other bias parameters relative to it.

This is a log-linear model because the logarithm of each cell probability is a linear function of a set of parameters. Here the parameters are the vectors $\ln \beta_j$ and

[2] Tversky (1977) has convincingly shown that *judgments* of similarity are frequently not symmetric and has provided a mathematical model of such judgments along with descriptions of empirical conditions which should lead to asymmetries. Attempts to generalize this model to cover indirectly measured similarity have not been too successful (Keren and Baggen, 1981; Smith, 1982). Krumhansl (1978) has discussed asymmetries also, but primarily in terms of confusion probabilities. It is clear that conditional probabilities, even in biased choice models, can be asymmetric for two reasons. The simplest is that one response is favored and this is the reason for the β parameters. But even if all β's were equal, the conditional response probabilities might be asymmetric because of the normalization leading to conditional probabilities. If a stimulus is near a large number of other stimuli (in a 'dense' region) many of its η's will be large, causing its normalization constant to be large and therefore all its conditional error probabilities to be small, and in particular, smaller than their partners associated with a relatively isolated stimulus.

$\ln \eta_{ij}$. This is a real advantage for us as data analysts, since it means that there are existing algorithms to calculate maximum likelihood estimates of the parameters, to test various hypotheses and, even better, to calculate confidence intervals on contrasts of various parameters (in logarithmic form). Even more important, if simplifications of the model can be found that retain its multiplicative form, tests of these special cases can be performed as easily and as routinely. All of the models to be discussed here are log-linear models or weighted averages of log-linear models which I call 'aggregated models'.

I will make all my points using a single example. The data have been published (Pachella, Smith and Stanovich, 1978), and are fairly simple, but still manage to illustrate a number of techniques for exploring the implications of a data set for Luce's BCM and for the alternative choice models referred to in the title. The purpose of this exposition is not to say anything definitive about how perception or speeded classification or any other psychological task really works, but rather to contribute to the toolbox available for studying such data sets. No doubt if this example had not yielded results differing among the alternative models I wanted to present, I would have hunted up a more favorable example.

The data table under discussion was filled as a small part of an investigation of the temporal course of stimulus identification. A major manipulation was to require the subject to make his/her identification before or as soon as possible after a deadline. Deadlines were varied to give a picture of how identification changed from early in the process until late. The stimuli were four stylized letters: B, C, D, and E. The letters were formed from the factorial crossing of two features, a vertical bar on the right side, distinguishing between B and D on the one hand and C and E on the other, and a small central horizontal bar characteristic of B and E but not of C and D.

The data presented here were produced by subject J.H. working with a deadline of 360 ms. Each stimulus was presented 75 times in each of four sessions spread for the most part over four days, so the confusion matrix was four by four with roughly 300 observations per stimulus. Although there is some evidence that there were minor differences from one day to the next they did not seem systematic so they

Table 1
Subject J.H. with a 360 ms deadline

Stimulus	Response				Total
	B	C	D	E	
B	201	23	59	10	293
C	37	203	20	50	310
D	86	41	162	9	298
E	37	59	23	181	300
Total	361	326	264	250	1201

have been ignored. Ignoring these differences would normally, although not necessarily, inflate the measures of goodness of fit I will report later. The data are presented in Table 1.

The data are rather typical. The entries in diagonal cells containing correct responses are rather larger than the others since about 62% of the responses are correct. This is no surprise because the deadline was chosen to provide about 60% correct responses. It seems clear that at this level of data processing the small central bar is not too valuable a clue, both the B–D and the C–E confusions being quite numerous. The right vertical bar, on the other hand, even in the first few hundred milliseconds has already contributed heavily to the identification task. B and E seem to be quite clearly distinguished, as do C and D, although perhaps not quite so separated as B and E. Finally, 'label' biases are another obvious feature of the data, 'E' seemingly being rather an unpopular response. Indeed, even casual inspection suggests a kind of alphabetic preference.

So far our analysis has been at the 'eyeball' level, but still richer than merely testing for independence, finding that it does not obtain, but with no clue as to how it is wrong. The classes of models to be taken up in this paper all provide ways of going much beyond simple independence. All of them are either generalizations or specializations of the biased choice model, and all are examples of multinomial models as defined in Riefer and Batchelder (1988). The generalizations are what I advertise in the title, the specializations are presented because I feel they are not considered often enough.

First we fit Luce's BCM using maximum likelihood. The cell expectations using this model are presented in Table 2 along with the deviations from the data for comparison.

The index G^2 for these data is 12.39 which, with three degrees of freedom, is significant at the 1% level. Quite clearly the model does not fit. At the same time the general features of the expected values seem appropriate. Even with the hundreds of observations made here the largest error is less than ten. A more useful index for data analysts and for the rest of this paper is G^2/df, which has many of the

Table 2
Subject J.H. expected values (deviation from observed)

Stimulus	Response				Total
	B	C	D	E	
B	201	24(−1)	54(5)	14(−4)	293
C	36(1)	203	29(−9)	42(8)	310
D	91(−5)	32(9)	162	13(−4)	298
E	33(4)	67(−8)	19(4)	181	300
Total	361	26	264	250	1201

Table 3
Subject J.H. similarities and biases

Stimulus	Similarities (η)			
	B	C	D	E
B	1.000	0.145	0.389	0.112
C	0.145	1.000	0.168	0.277
D	0.389	0.168	1.000	0.092
E	0.112	0.277	0.092	1.000
Biases	0.317	0.260	0.229	0.194

statistical characteristics of an F-value with df and infinite degrees of freedom. Here G^2/df (call it F^*) is 4.13, clearly significant but, as we will soon see, not really large.

Even though the model does not fit, for comparative purposes we can go ahead and estimate parameters, presented in Table 3.

The parameter estimates seem to reflect the same characteristics we noticed in the raw data but we still do not have a very good idea how simple the model might be. Do we know, for example, whether any of the off-diagonal similarities are different? Why are B and C, which differ on both features, estimated as being more similar than are B and E, which differ only on one? B and E differ with respect to the vertical bar just as do C and D, yet the similarities are not the same. Is the difference reliable? The bias parameters seem to be alphabetically ordered, but are the differences really reliable?

The most conservative data analyst might say: 'Why bother? These questions do not arise since the most general form of the BCM has been shown not to fit.' If only to illustrate the procedures let us try to address some of these questions. Are the six similarities different from one another? When we fit a model which assumes that they are not we get a G^2 of 122.33 with eight degrees of freedom. The additional degrees of freedom, five of them, were bought at a cost of increasing G^2 by 109.94 so $F^*=22.0$. The similarities are vastly different from one another. Along the same line, is there evidence of differential response bias? Fitting a model having no differential response bias yields a G^2 of 41.66 with six degrees of freedom. The increase is 29.27 with three degrees of freedom, or $F^*=9.76$, not quite so large but impressive nevertheless. If one were to accept for the moment the Luce BCM for this data one would feel compelled to accept also differential similarities and differential biases.

Somewhat more deeply, if the features are to be relevant to the identification process, confusions that correspond to the same feature differences should have the same similarity. The letter pairs CD and BE, the pairs BC and DE, and the pairs BD and CE all differ by the same feature combination, so by the simplest model should also have the same similarity indices. We will refer to this model as the

feature pattern model. If we subscript similarities with v when the stimuli differ only in the vertical feature, with h when they differ only in the horizontal feature and with hv when they differ in both features, we obtain the pattern of similarities shown in Table 4.

Note that this model has only three similarity parameters and not six as the general BCM has. It says only that any pair of stimuli that differ on a set of features will have the same similarity index as any other pair that differ on the same set. It is an interesting special case of the BCM precisely because it takes note of the structure of the stimuli. Nothing is implied about a relation between any of the parameters in Table 4, although one might prefer not to find that η_{hv} is larger than the other two, since stimuli that differ by two features should not be more similar than stimuli which differ by only one.

An even more parsimonious model, which still falls in the log-linear class, is the model which has it that different similarity indices are needed only for each feature, not for every combination of features. The similarity of two stimuli is the product of the similarity indices of the features on which they differ. In our example that would mean that B and C would have to be less similar than B and E because B and C differ on both the horizontal and the vertical features, while B and E differ only on the vertical feature. It means more than that. This could be called an independent features model or, in line with log-linear terminology, it is a model of 'additive features' (in terms of the logarithms of the feature ηs). Data that satisfy the feature pattern model but not the additive feature model can be either sub-additive or super-additive depending upon whether:

$$\eta_h \eta_v < \eta_{hv} \quad \text{sub-additive}$$

or

$$\eta_h \eta_v > \eta_{hv} \quad \text{super-additive.}$$

When this model is fitted we find a strongly sub-additive fit with η_{hv} more than 2.5 times as large as the product of η_h and η_v. The fit of the model in terms of F^* is nearly the same as that of the BCM model, $G^2 = 25.1$, df $= 6$, $F^* = 4.2$. The model does not fit, but it does give a reasonable feel for the data.

When we look at the fit of the additive model we begin to see just how discrepant a real mismatch is. The maximum likelihood fit of an additive feature model yields

Table 4
Similarity pattern for feature pattern model

	B	C	D	E
B	1	η_{hv}	η_h	η_v
C	η_{hv}	1	η_v	η_h
D	η_h	η_v	1	η_{hv}
E	η_v	η_h	η_{hv}	1

a G^2 of 59.7, an increase of 34.6 with the gain of only one degree of freedom. This is roughly equivalent to a normal deviate of $\sqrt{34.6}$ or 5.9.

Summing up our analysis of the efforts of J.H. with a 360 ms deadline, we see that his error pattern is highly heterogeneous, that is his pattern of confusions is related to the structure of the stimuli. In addition he uses the different possible responses with different frequencies, the alphabetically earlier responses being more 'popular' than the later ones. The BCM model can be rejected as a complete model of J.H.'s data, but it is close enough to be usefully descriptive.

The feature pattern model describes the data fairly well. A very strong sub-additive effect is observed in that confusions when two features differ between stimuli are hardly any less frequent than when only the vertical stroke is the discriminative feature. The small horizontal stroke adds little to the discriminative powers of J.H. when the stimuli already differ with respect to the vertical bar.

2. An ordered bias alternative

Rob Nosofsky (personal communication, 1988; see also Nosofsky, 1991) suggested that to call the β parameters 'response bias' was a rather presumptuous way to treat the BCM parameters. It is quite easy to think of characteristics of experimental situations which would lead to data not satisfying the equal β situation but to attribute this as necessarily due purely to response processes would be premature.

The BCM as it is usually described involves a set of 'response biases' parameterized as β_j and adding to one. The additivity is really more a technical convenience than it is a feature of the model, however. What is determined by the data is really the ratio of any pair of β's, and these differ from η's in that the η's are symmetric in their index arguments, whereas the β ratios, or rather the logarithms of these ratios, are anti-symmetric. If one defines

$$\gamma_{ij} = \beta_i / \beta_j, \tag{3}$$

then the BCM requires that

$$\gamma_{ij}\gamma_{jk}\gamma_{ki} = 1, \quad \text{for all } i,j,k. \tag{4}$$

Values of γ_{ij} greater than one are interpretable as a preference for response i over response j, or a bias.

Indeed, deviations from this condition alone determine the goodness of fit of the BCM to a set of data. As is shown in Appendix B, if the set of consistent estimators of γ's, g_{ij}, defined as

$$g_{ij} = \sqrt{\frac{n_{ji}n_{ii}}{n_{ij}n_{jj}}},$$

where n_{ij} is the frequency of response j to stimulus i, satisfies equation (4), the data

satisfy Luce's BCM. At first glance it may seem counter-intuitive that the fit of the model depends only on the antisymmetric functions and not on the η's (note that η cancels from this function). The reason is that the model contains parameters for all possible symmetric functions of two subscripts. The model is, so to speak, saturated with similarities. Anything symmetric is permitted to happen; but the biases must have a form that allows separation into individual response effects. The argument is technically the same as that for the Bradley-Terry-Luce model of preferential choice (Luce, 1959).

Another kind of anti-symmetric function can be used if there is a natural ordering on the stimuli. For instance, our example has the ordering B, C, D, and E, the alphabetic ordering. Using this ordering we can define another antisymmetric function:

$$\gamma_{ij}^{*} = \begin{cases} b & \text{if } j \text{ precedes } i, \\ 1 & \text{if } i = j, \text{ and} \\ 1/b & \text{if } i \text{ precedes } j. \end{cases} \tag{5}$$

This function, if it replaces the function of equation (3), represents a much less constrained model since it proposes only one bias parameter, not $I-1$ different ones. If it is used, the resulting biased choice model stands between the biased choice model which corresponds to the case with $b=1$ (the 'unbiased biased choice model'?) and the classic BCM.

In our example a value of γ^{*} greater than one would imply that the response corresponding to the earlier stimulus in the sequence would be the more popular but that the preference of B over E would be no greater than that of C over D. The form of the model is then

$$p_{ij} \propto \gamma_{ij}^{*} \eta_{ij}. \tag{6}$$

A completely post hoc justification for this model in this experiment might be that for the most part the stimulus will have reduced the choice to one of two possible responses very early. The extreme emphasis on quick responses, coupled with a *sequential* search of the possibilities, might lead the subject to respond with the first response consistent with what he/she has seen so far, and thus the earlier response in the natural order would be favored. It is not rare that stimuli in an identification experiment have one or more natural orders and the justification here is meant to suggest that future experimenters make provision for building the possible effects of such orders into the data analysis.

Regardless of why this model might work, we proceed with fitting it to the data. We note that we are fitting two fewer parameters, and that our model 'explains' why the biases we noted when fitting the BCM might have arisen. The earlier in the sequence a stimulus is, the more frequently it will obtain the bias advantage. When the model is fit, the obtained G^2 is 10.75 with five degrees of freedom, $F^{*}=2.25$, and a nominal attained significance level of 0.057. Not only is G^2 smaller by 2, but we have two more degrees of freedom. It is evident that this model is an alternative,

Table 5
Subject J.H. fit of alphabetic bias

Stimuli		Responses			
		B	C	D	E
B	Obs	201.0	23.0	59.0	10.0
	Exp	198.1	20.5	57.1	17.4
C	Obs	37.0	203.0	20.0	50.0
	Exp	39.5	204.8	24.5	41.1
D	Obs	86.0	41.0	162.0	9.0
	Exp	87.9	36.5	163.2	10.4
E	Obs	37.0	59.0	23.0	181.0
	Exp	29.6	67.9	21.6	180.9

not a special case, since a special case will always yield a G^2 at least as large as its more general relative. The observed and expected cell entries are displayed in Table 5.

To achieve this nearly acceptable fit the earlier response had an advantage of 1.87 : 1 over the later one. This number is the maximum likelihood estimate of b^2 in the model and can be calculated directly from the expected values in Table 5 with the formula

$$\hat{b}^2 = \frac{m_{kj} m_{ji} m_{ik}}{m_{ij} m_{jk} m_{ki}}, \quad i<j<\mathrm{k}, \tag{7}$$

where m_{ij} are cell expected values (slightly more accurate expected values were used in the calculation above). Returning to Table 5 we see that the preference of B over E, looking at the cell expected values, should have been even larger. More than half of the remaining G^2 can be accounted for by the two BE cells. Indeed, we can quite easily extend this alternative model by allowing the early–late feature to have different strengths depending on the difference between i and j. This costs two degrees of freedom and effectively assigns one whole parameter to the BE pair which saturates those two cells. The results of that analysis are shown in Table 6.

This is easily the best fit we can report in this paper. There are three degrees of freedom, G^2 is 2.82, and $F^* = 0.94$. Remember that the basic BCM also had three degrees of freedom but a G^2 of 12.39.

Despite this good fit, however, there are still three disturbing facts. The first is that the alphabetic response asymmetries are not ordered in an intuitively satisfying manner. The BE, three-step asymmetry is 4.17, the two-step asymmetry is 1.55, and the one-step asymmetry is 2.06. The lack of a monotone relation is not serious since the inversion in the order is not large. Indeed, we cannot reject a model in which

the three-step asymmetry is large, larger than the one- and two-step asymmetries.

The second is that the similarity indices do not make much sense either. B and E are estimated to be less similar than B and C and also D and E are estimated to be more similar than B and E. In both cases two stimuli that differ on only one feature seem *less* similar than a pair that differ on both features.

The final disturbing fact is that when we try to fit a feature pattern model along with the alphabetic distance model, the fit is very poor indeed: G^2 is 18.00, an increase of 15.1 with three degrees of freedom, an F^* of 5.0. If we accept this marginal fit and look again at additivity we find an even more subadditive set of parameters than before. So far in all the models we have considered the two-feature similarity index has ranged from 3 to 5 *times* the product of the two one-feature indices.

Indeed, in scaling endeavors sub-additivity has often cropped up. In an interesting chapter Kendall (1971) defined a 'horse-shoe' effect that seemed to show up repeatedly in archaeological seriation, a kind of multidimensional scaling problem in which archaeologists use frequencies and co-occurrences of artifacts in graves and other ancient sites to deduce the historic times when these deposits were laid down. Using various non-metric scaling techniques investigators found they needed two or more dimensions to represent the similarities as distances, even when the time order of deposits was well established. The plot of sites represented the time dimension as a curve, a 'horse-shoe' or hook rather than a straight line. This is of course another kind of sub-additivity, a case when the triangular inequality should be a real equality and is not.

The archaeological example and the confusion matrix example are very similar. Co-occurrences on the one hand and confusions on the other are taken to represent 'nearness' and items which should, given linearity, be far apart are actually too close together. One way this can be represented is to twist the representation back on itself, forming a hook or 'horse shoe'.

Table 6
Subject J.H. alphabetic distance bias

Stimuli		Responses			
		B	C	D	E
B	Obs	201.0	23.0	59.0	10.0
	Exp	200.4	19.5	63.1	10.0
C	Obs	37.0	203.0	20.0	50.0
	Exp	40.5	201.2	22.4	45.9
D	Obs	86.0	41.0	162.0	9.0
	Exp	81.9	38.6	167.5	10.0
E	Obs	37.0	59.0	23.0	181.0
	Exp	37.0	63.1	22.0	177.9

3. The Guessing Mixture Model (GMM)

A very early model for confusion data was called the 'all-or-none model' (Townsend, 1971) or the 'simple guessing model' (Broadbent, 1967). The scenario for this model is that the subject is presented with the stimulus and she either 'detects' the stimulus, in which case she correctly names it, or she misses the stimulus altogether, in which case she guesses, according to some response bias. The experimenter cannot tell which state the subject was in on any trial so his data have been aggregated for him. The mathematical model for this scenario is

$$p_{ij} = \delta_{ij} P_i + \pi_j (1 - P_i), \tag{8}$$

where δ_{ij} is the Kronecker delta, with a value of 1 when $i = j$ and 0 otherwise, P_i is the probability of being in the detect state when stimulus i is presented, and π_j is the response bias toward response j and $\sum \pi = 1$. An interesting special case of this model would have it that all the stimuli were equally detectable, i.e. $P_i = P$ for all i.

We must be quite careful in relating this pair of models to the BCM, because the relation is a complicated one, as has not always been clear to me (see Smith, 1973). If we just consider the off-diagonal cells, the 'errors', for the 'all-or-none model' and for the BCM, we see that the conditional all-or-none model is indeed a special case of the conditional BCM. The all-or-none model is the BCM with all similarity parameters equal to one, or actually any other constant value, when we consider the error cells by themselves.

The models differ fundamentally in how they account for the correct responses. In the all-or-none account the diagonal entries are big (or at least bigger) because they consist not only of lucky guesses but also of trials on which the presented stimulus was correctly perceived. In the BCM the diagonal cells are large (if they *are* large) because the similarity of a stimulus to itself is large, relative to the other similarities. As the number of cases observed in the diagonals decreases, the BCM increases the estimates of all off-diagonal similarities to maintain the best possible fit, while the all-or-none model reduces estimated 'true' perceptions as long as there are any to reduce, then sets estimated true perceptions to zero, and finally has to accept fewer than expected 'by chance' observations in the diagonal cells. The maximum likelihood fit of the BCM to a set of data will *necessarily* fit the diagonal (correct) cells exactly. So will the simple guessing model, *if the number of observations in the diagonal cells is large enough*. If the subject is performing consistently above chance the expected values for simple guessing will be a BCM. It will not be the best fitting BCM of course because simple guessing is only a special case of biased choice. Both models fit the diagonal because both must agree with the data there mathematically. The simple guessing model, is, however, always a valid special case of the following quite different alternative choice model, the Guessing Mixture Model (GMM).

Like the simple guessing model, GMM has it that on a certain fraction of the trials, $(1 - \alpha)$, the subject is either not attentive, slow, asleep, or, for some reasons, has not a clue as to what stimulus was presented. On those trials the subject guesses

according to some response set. The important thing is that on those trials responses and stimuli are independent. On the other trials the subject is attentive, quick, and alert. Unlike the subject described in the simple guessing model, however, the alert GMM subject cannot always report the correct stimulus. The conditional probability of response j when stimulus i was presented is proportional to η_{ij}, where the η_{ij} have the same properties as did the similarity parameters of the BCM. As a formula we have

$$p_{ij} = (1-\alpha)\pi_j + \frac{\alpha\eta_{ij}}{\sum_j \eta_{ij}}. \tag{9}$$

The special form of simple guessing when stimuli are equally detectable is obtained from equation (9) by setting all similarity parameters to 0 if i is not equal to j, and keeping η_{ii} at one as before. The parameter α takes the place of the detection parameter P. Variants on equation (9) spring readily to mind. There might be different α's for different stimuli, or there might be an additional set of biases on 'attention' trials as well as on the 'inattention' trials. A similar model was proposed in Pachella et al., but equation (9) will define for our purposes the GMM. The GMM can be called an aggregated log-linear model. If there were some way we could tell which correct responses were real detections and which were merely lucky guesses we would have the information to fit a log-linear model to the guessing data, estimate bias parameters, then fit another log-linear model to the attentive data, estimate similarities, and then put the estimates back into equation (9) to see how well the whole thing fits. As it is we need new techniques to overcome the fact that the nature of our data aggregates these cells. This is not the place to go into the promise and problems of aggregated log-linear models in general, but a little background will illuminate the application to confusion matrix models.

The approach is to distinguish between latent tables and a manifest table. The manifest table is the table actually observed. In our example Table 1 is the manifest table. We know all its entries. If we try to fit that data with (say) the simple guessing model it is useful to think of a latent table as well. Again in our example that table has four more cells, an extra one for each diagonal cell. Each diagonal cell is *really* two cells that have been aggregated. If we only knew how many of the entries in the diagonal cells were true detections and how many were lucky guesses, we could fit one or another log-linear model easily.

What is needed is a way to estimate not only the parameters of the model, but to estimate the 'observed' cell frequencies in the aggregated, manifest cells. The procedure used is based on the so-called EM algorithm (Dempster, Laird and Rubin, 1977). The name comes from first Estimating the missing cell frequencies, and then Maximizing the likelihood, conditional upon those estimates. Then, almost always, a better estimate of the latent cell frequencies can be obtained and the cycle continues. In our case we improve the latent cell frequencies by making them proportional to the recently computed expected values in those cells. Perhaps a small numerical example will clarify this step. Suppose the number of correct identifica-

tions of the letter B is 201, as above, and that the current estimate of the number of detections of B and the estimated number of lucky 'guesses' of B in the fitted model are 75 and 25 and the 'pseudo-observed' numbers are 180 and 21. Note that the pseudo-observed numbers *must* add to 201. The new values of the pseudo-observed are made proportional to the currently estimated expected values, in this case a ratio of 3 : 1 or 150.75 and 50.25.

This can easily be shown to improve the goodness of fit, so that the iteration at every step decreases the G^2 measure. Since G^2 is necessarily positive it will approach a limit point, at which point parameter values and latent cell entries are maximum likelihood estimates.

The algorithm is very similar to those used in ordinary maximum likelihood solutions of log-linear problems. A major difference, however, is that the EM procedure does not necessarily reach a unique maximum. Fitting the simple guessing model will be an example of a problem with a unique solution; fitting the general GMM will be an example of a model with a manifold of (LR) equivalent solutions. We begin by displaying in Table 7 the J.H. data when the first E (estimating) step has been carried out.

The estimates in the final column are not to be taken seriously. They merely get the process started. The single detectability, simple guessing model for this table is the same as an independence model for an ordinary table. The column parameter for the last column is the detectability parameter P. The other four column parameters, when normalized, are estimators of the bias vector. For such a table the maximization step is very simple, indeed right out of an elementary text. The maximum likelihood estimate for the 'lucky guess' of B is $(261)(293)/1201 = 63.7$ and the same value for the B detection cell is $(400)(293)/1201 = 97.6$, about half again more. The new E stage is to change the observed table to reflect the same proportionality. Instead of 101 and 100, we insert 79.4 and 121.6 in the latent cell table. The new latent cell frequencies now add to 201, as they must, and are proportional to 63.7 and 97.6, which is as close to the model as one can come with the current constraints. Completing this iterative process, taking six cycles for this data set, we obtain Table 8 as the final latent cell 'data' table.

Table 7
Subject J.H. latent data table

Stimulus	Guesses				Correct
	B	C	D	E	
B	101	23	59	10	100
C	37	103	20	50	100
D	86	41	62	9	100
E	37	59	23	81	100
Totals	261	226	164	150	400

Table 8
Subject J.H. final latent data table

Stimulus	Guesses				Correct
	B	C	D	E	
B	52.9	23	59	10	148.1
C	37	44.6	20	50	158.4
D	86	41	29.2	9	132.8
E	37	59	23	24.6	156.4
Totals	212.9	167.6	131.2	93.6	595.7

The next cycle would use these to compute a new cycle of expected values, which would now be nearly the same as the preceding cycle. The G^2 with eight degrees of freedom would be 117.5 or $F^* = 14.7$. This model does not fit at all! Neither does the model allowing different detectabilities. In fact, it fits even worse, $F^* = 22.0$. It is not that detectabilities are different; it is that there are unique confusions for specific pairs of stimuli. In particular, stimuli that differ *only* in whether or not there is a short horizontal feature half way up are much more frequently confused than any other pair of stimuli. These data cannot be adequately described without some notion of similarity. That by itself is interesting when we remember that the subject was supposed to examine the stimuli and make his response in 360 ms.

4. The full GMM model

Having ruled out simple guessing we return now to the general GMM of equation (9). Now instead of having one additional cell for each row containing the 'real' detections, we will have to accommodate an entire additional table to hold data from the attentive state as well as the one holding guesses. The procedure is just as before except that now after every cycle of approaching the maximum likelihood solution we go back and adjust the latent cell data to match proportionately the current expected values in each manifest cell. Convergence can be excruciatingly slow, although for our data 50 ms on the main frame or 50 s on my micro is adequate.

Much more important is the fact that the final result will usually depend upon the starting point. Fitting our model leads to one *manifest* table of estimated cell entries and thus to one G^2 value, but to a whole range of parameter values. Many different tables of latent expected values will be obtained, depending on the particular starting point of the iteration, but each will aggregate to the same manifest set of expected values. Parameters are estimated from latent cell expected values, and these will differ from one starting configuration to another as a function of the estimated time in the attentive state. Each solution will attribute different amounts

Table 9
Subject J.H. manifest GMM

Stimuli		Responses			
		B	C	D	E
B	Obs	201.0	23.0	59.0	10.0
	Exp	198.4	24.0	61.5	9.2
C	Obs	37.0	203.0	20.0	50.0
	Exp	35.7	205.3	23.3	45.7
D	Obs	86.0	41.0	162.0	9.0
	Exp	82.6	35.3	168.3	11.8
E	Obs	37.0	59.0	23.0	181.0
	Exp	40.6	65.5	18.8	175.1

of a manifest cell's contents to attentive responses and to guesses, even though the manifest cell total stays the same.

The manifest table with observed and expected values is shown in Table 9. This model fits with a G^2 of 5.27. The degrees of freedom should be two because we are now fitting an α as well, but there may be some question about this because of the lack of uniqueness. One set of parameters is given in Table 10 and another in Table 11. Both of these parameter sets will lead to the same manifest expected values given in Table 9.

Notice that Table 10 has a zero for a similarity, BC, and Table 11 has a zero for a response bias, E. This is no accident. I chose these two tables because they are end points on a manifold from a null similarity to a null bias. One can plot a continuous path through the parameter space from the model of Table 10 to that of Table 11. For every model on the path G^2 will be 5.27. At one end of the path the maximum number of guesses will be estimated and at the other the maximum

Table 10
Subject J.H. similarities and biases, GMM solution 1 ($\hat{\alpha}=0.765$)

Stimulus	Similarities (η)			
	B	C	D	E
B	1.000	0.000	0.318	0.037
C	0.000	1.000	0.076	0.236
D	0.318	0.076	1.000	0.056
E	0.037	0.236	0.056	1.000
Biases	0.478	0.348	0.129	0.044

Table 11
Subject J.H. similarities and biases, GMM solution 2 ($\hat{\alpha}=0.809$)

Stimulus	Similarities (η)			
	B	C	D	E
B	1.000	0.018	0.330	0.054
C	0.018	1.000	0.093	0.250
D	0.330	0.093	1.000	0.073
E	0.054	0.250	0.073	1.000
Biases	0.522	0.374	0.104	0.000

number of 'attentive' responses will be estimated. Only if the data table is best fit by a parameter set with one similarity parameter *and* one bias parameter equal to zero will the maximum likelihood solution for parameters be unique. The difference here may be small, but if one is looking to obtain distance indicators from the logarithms of similarity indices, Table 10 will be quite a shock.

5. Feature patterns in GMM

In both of the earlier families we examined the fit of the feature pattern model. That is the model, remember, in which the similarity of two stimuli depends only on which collection of feature differences the stimuli had. I think I have implied that this would be a minimally rewarding model to find fitting. If so, I should back down slightly. B and D differ with respect to the tiny horizontal bar, as do C and E. But the BD pair both have the long vertical bar in common, so they might be more similar in terms of *proportional* differences than are C and E, which have fewer features in toto. Surely as one keeps adding identical features to each of two visual objects one will arrive at two composite objects for which the one distinguishing feature is too insignificant to notice. Tversky (1977), in his work on judged similarity, cites empirical evidence of this tendency.

Nevertheless it is quite easy to study additivity if the feature pattern model is at least to some degree supported. The feature pattern model in GMM also has a maximum likelihood ridge. On the ridge G^2 is 10.552 with five degrees of freedom, not significant at the 5% level. As was noted above, the additive feature model is a special case of the feature pattern model for which $\eta_{hv}=\eta_h\eta_v$. When we fit that model, much to our surprise we find that G^2 is also 10.552, now with six degrees of freedom, and suddenly insignificant at the 10% level.

This should not have been too surprising. At one end of the ridge one pair of stimuli will have a similarity index of zero, which will almost certainly be the similarity between pairs of stimuli that differ on both features. This will be a table with extreme super-additivity unless one or the other feature is totally effective by

itself. At the other end of the ridge the similarity when both features differ will be as large as possible, and frequently large enough to make the system sub-additive. If this is the case, by continuity some parameter set in the middle will exhibit additivity, and that is what has happened in the J.H. data. Indeed, by very similar reasoning we see that the additivity solution is unique, because of the monotonicity in the similarities along the ridge. Only if the entire one-dimensional manifold which maximizes the likelihood exhibits super-additivity will there not be an additive solution.

In the J.H. data there is additivity. The similarity parameter estimates are

$$\hat{\eta}_v = 0.0218 \quad \text{and} \quad \hat{\eta}_h = 0.2478,$$

showing the almost complete effectiveness of the vertical cue and somewhat incomplete use of the horizontal cue. We can calculate likelihood-based 95% confidence intervals for both parameters. These are

$$\{0.192 < \eta_h < 0.310\}_L \quad \text{and} \quad \{0.000 < \eta_v < 0.074\}_L.$$

These values are obtained by inverting the maximum likelihood test. If we were to fix the value 0.192 for η_h and fit the rest of the model we would obtain a G^2 just 3.841 larger than the minimum, which is the 5% significance level for G^2 with one degree of freedom. The lower limit for η_v is interesting. It means that no positive value for η_v is small enough to increase G^2 by a significant amount. Such an estimate would be impossible in the BCM unless all cells sharing that parameter had zero data entries. The GMM, however, can handle the fact that there are no attentive responses in such cells because it can attribute such responses to inattentive responses or guesses. It is even more constraining in the additive model, since it implies that the *only* attentive confusions are those between B and D and those between C and E.

The question of degrees of freedom is still a problem, but not puzzling. If one specifies a point on the ridge, like either end point or the additive point, the program takes that to mean that one parameter, a bias or similarity, or one contrast, additivity, was specified a priori, and hence not estimated from the data. More work is needed on the statistical questions when we do not have a unique maximum likelihood.

Whatever the niceties of statistical evaluation the difference between the additive feature solution in GMM and the classic BCM is impressive. With Biased Choice and the J.H. data the additive feature model produced a G^2 of 59.709 with seven degrees of freedom. GMM on the other hand with six degrees of freedom yielded a G^2 of 10.552. Indeed even the most general form of the BCM produced a G^2 of 12.39 with only three degrees of freedom.

6. Caveats, qualifications, and comments

It is tempting to draw conclusions about how line segment letters are processed

in deadline conditions, but that would be foolhardy on the basis of this paper. As an illustration of this foolhardiness notice that both the 'triangular bias' model and the GMM fit the data quite well, even though their accounts are not similar. The triangular bias model sounds rather like a satisfying model. The subject searches the alphabet serially until an acceptable response appears, or time is running out. If time runs out she emits the response corresponding to the most similar stimulus so far. The fit is good, but the similarity parameter estimates are puzzling.

On the other hand, GMM would have it that on a fourth to a fifth of the trials J.H. was unprepared, emitting a stimulus independent response just to meet the deadline. On the rest of the trials he seemed to be using an additive feature model with the vertical feature almost entirely processed during the 360 ms period, but with only a beginning use of the horizontal feature. We remain puzzled with this model why his biases are as they are.

The conditions of this experiment seem very suggestive of those labeled 'state-limited' by Garner and Hahn (1978). Would these models be useful in studying 'process-limited conditions' as well? Of course we do not know any of those things from the data of one long-gone subject, chosen who knows how.

One final comment about the GMM. Measurement models for a long time have included a 'noise component'. Frequently this component is assumed to be unbiased on average, which leads to a strategy of using large samples to wash out the effects of the noise. But large samples in confusion matrices do not wash out the effects of stimulus independent responses. The stability due to averaging measurement data may be attained with categorical data by using mixture models like the GMM.

7. Summary

The objectives of this paper were to propose some alternatives to the classical BCM and to suggest some situations in which they might be useful. When general, non-structural forms of the three models are used there will not often be a lot to choose between them. This may not necessarily extend to studies of their special cases or to the meaning of the parameters they use.

In the data set we studied here so extensively we observed that additivity in the GMM actually fit better than the general form of Biased Choice with twice the number of degrees of freedom left for error.

We saw that, like many other aggregative log-linear models, GMM often may not provide unique solutions. Sometimes, as in the current example, the ambiguity may be turned to advantage.

Appendix A

Any model satisfying 'quasi-symmetry' (Bishop et al., 1975) can be reparameterized as a Luce Biased Choice Model (BCM).

Bishop et al. define quasi-symmetry as

$$m_{ij} = a_i b_j d_{ij}, \quad \text{with } d_{ij} = d_{ji}, \quad \forall i,j. \tag{A1}$$

We define

$$\eta_{ij} = \frac{d_{ij}}{\sqrt{(d_{ii} d_{jj})}}, \tag{A2}$$

$$\beta_j = b_j / \sqrt{d_{jj}} \tag{A3}$$

and

$$\alpha_i = a_i / \sqrt{d_{ii}}. \tag{A4}$$

In terms of these parameters then

$$p_{ij} = \frac{m_{ij}}{m_{i^+}} = \frac{\beta_j \eta_{ij}}{\sum_k \beta_k \eta_{ik}}. \tag{A5}$$

Here p_{ij} is the conditional probability of response j given the presentation of stimulus i and m_{i^+} is the total number of responses to stimulus i. By construction η_{ii} is one for all i, as was required. It should, however, be noted that sometimes the Luce BCM is taken to be restricted to that part of the parameter space in which $\eta_{ij} \leq 1$ for all i,j. If that definition is accepted, then the BCM is only a subset of the quasi-symmetry models, since nothing in the preceding development precludes some η_{ij}, $i \neq j$, from being larger than 1.

Appendix B

Because the numbers $p_{ij} = m_{ij}/m_{i^+}$ are maximum likelihood estimators of the population of conditional probabilities they are necessarily consistent estimators. We show here that a square data table $[m_{ij}]$ satisfies the BCM if and only if

$$m_{ij} m_{jk} m_{ki} = m_{ji} m_{ik} m_{kj}, \quad \forall i,j,k. \tag{B1}$$

The necessity is obvious on substitution of the BCM into equation (B1). Sufficiency requires more effort. We assume at the beginning that all m's are strictly positive or, rather, include this constraint in what we will refer to as a BCM. Certain patterns of zero expected values could be allowed in an extended version of the BCM. In particular we can allow m_{ij} to be zero if we insist that m_{ji} also be zero and that the system of equations (B1) not include equations involving those cells, or other such empty pairs.

Notice also that equation (B1) puts no constraint on m_{ii}. We will assume that all diagonal cells have strictly positive entries. Real applications almost always have large correct identification frequencies. Models that propose no *probability* of a correct response are not considered here. We will assume a strictly positive set of m_{ii} which might as well be the observed observations n_{ii} in any application.

We now define a set of column construct parameters γ_{ij}:

$$\gamma_{ij} = \sqrt{\frac{m_{ji} m_{ii}}{m_{ij} m_{jj}}}. \tag{B2}$$

When we multiply two of the parameters, γ_{ij} and γ_{jk}, we obtain:

$$\gamma_{ij}\gamma_{jk} = \sqrt{\frac{m_{kj} m_{ji} m_{ii}}{m_{ij} m_{jk} m_{kk}}} = \gamma_{ik}, \tag{B3}$$

using condition (B1). Summing equation (B3) over k we obtain:

$$\gamma_{ij}\gamma_{j^+} = \gamma_{i^+}$$

or

$$\gamma_{ij} = \beta_i/\beta_j. \tag{B4}$$

Mutatis mutandis we can define $\delta_{ij} = \alpha_i/\alpha_j$ for rows. Using equations (B3) and (B4) we find that

$$\alpha_i \beta_j = \frac{m_{ij}}{m_{ji}} \alpha_j \beta_i \tag{B5}$$

or

$$\frac{m_{ij}}{\alpha_i \beta_j} = \frac{m_{ji}}{\alpha_j \beta_i} = (\text{say})\ d_{iu}.$$

Thus, d_{ij} is symmetric and the conditions of (B1) are sufficient to have a quasi-symmetric model and, by Appendix A, a BCM.

References

F.G. Ashby, W.W. Lee and J.D. Balakrishnan, Comparing the biased choice model and multidimensional decision bound models of identification, Math. Soc. Sci. 23(2) (1992) this issue.

Y.M.M. Bishop, S.E. Fienberg and P. Holland, Discrete Multivariate Analysis (MIT Press, Cambridge, Mass, 1975).

D.E. Broadbent, Word-frequency effect and response bias, Psychol. Rev. 74 (1967) 1–15.

A.P. Dempster, N.M. Laird and D.B. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), J. Roy. Statist. Soc., Ser. B 39 (1977) 1–38.

W.R. Garner and F. Hahn, Letter identification as a function of type of perceptual limitation and type of attribute, J. Exp. Psychol.: HPP 4 (1978) 199–209.

G. Keren and S. Baggen, Recognition models of alphanumeric characters, Perception and Psychophysics 29 (1981) 234–245.

D.G. Kendall, Seriation from abundance matrices, in: F.R. Hodson, D.G. Kendall and P. Tautu, eds., Mathematics in Archaeological and Historical Sciences (Edinburgh University Press, Edinburgh, 1971) pp. 215–247.

C.L. Krumhansl, Concerning the applicability of geometrical models to similarity data: the interrelationship between similarity and spatial density, Psychol. Rev. 85 (1978) 445–463.

R.D. Luce, Detection and recognition, in: R.D. Luce, R.R. Bush and E. Galanter, eds., Handbook of Mathematical Psychology, Vol. 1 (Wiley, New York, 1963) pp. 103–189.

R.D. Luce, Individual Choice Behavior (Wiley, New York, 1959).

R.M. Nosofsky, Stimulus bias, asymmetric similarity, and classification, Cog. Psychol. 23 (1991) 94–140.

R.G. Pachella, J.E.K. Smith and K.E. Stanovich, Qualitative error analysis and speeded classification, in: N.J. Castellan and F. Restle, eds., Cognitive Theory III (Lawrence Erlbaum Associates, Hillsdale, N.J., 1978) pp. 169–198.

D.M. Riefer and W.H. Batchelder, Multinomial modeling and the measurement of cognitive processes, Psychol. Rev. 95 (1988) 318–339.

J.E.K. Smith, On tests of quasi-independence in psychological research, Psychol. Bull. 80 (1973) 337–351.

J.E.K. Smith, Recognition models evaluated: A commentary on Keren and Baggen, Perception and Psychophysics 31 (1982) 183–189.

J.T. Townsend, Theoretical analysis of an alphabetic confusion matrix, Perception and Psychophysics 9 (1971) 40–50.

J.T. Townsend and D.E. Landon, An experimental and theoretical investigation of the constant-ratio rule and other models of visual letter confusion, J. Math. Psychol. 25 (1982) 119–162.

A. Tversky, Features of similarity, Psychol. Rev. 84 (1977) 327–352.