COMPARISON OF MODELS FOR STIMULUS RECOGNITION DATA

Yoshio Takane, McGill University
and
Tadashi Shibayama, The University of Tokyo

Various models for stimulus recognition data were compared in terms of their goodness of fit (GOF). The models considered include the unrestricted similarity model, the euclidean distance model and the unique feature model. In all cases Luce's choice model was used to link the stimulus (dis)similarities to confusion probabilities. A maximum likelihood estimation procedure was developed to fit these models, and their GOF compared through the AIC statistic. Results were reported in detail for one data set (Keren & Baggen's data). Major findings were: (1) The unrestricted similarity choice model was found to be the best fitting model. (2) The euclidean distance model did not fit the data very well. (3) The unique feature model with the sets of features used in this study did not fit the data even as well as the poorly fitted euclidean distance model.

INTRODUCTION

In a recognition accuracy experiment a stimulus is randomly chosen from a set of n stimuli in each trial, and is presented to the subject under degraded stimulus presentation conditions. The subject's task is to identify the stimulus. For f_i replicated presentations of stimulus i we obtain a set of f_{ij} frequencies

*The work reported in this paper has been supported by Grant A6394 from the Natural Sciences and Engineering Research Council of Canada and by a leave grant from the Social Sciences and Humanities Research Council of Canada, both awarded to the first author. Major portions of this work were done while the first author was at the Institute of Statistical Mathematics in Tokyo, on leave from McGill University. The authors wish to express their gratitude to their colleagues at the Institute for their warm hospitality. The authors also wish to extend their gratitude to Dr. Gideon Keren at the Institute for Perception, TNO, The Netherlands, for providing his precious data. Thanks are also due to Drs. Jan de Leeuw, Tony Marley, Ivo Molenaar, Jim Ramsay, and Justine Sergent for their helpful comments on an earlier draft of this paper.

with which stimulus i is judged as stimulus j. A variety of models have been proposed for stimulus recognition data (e.g., Luce, 1963; Townsend, 1971; Nakatani, 1972; Keren & Baggen, 1981). These models attempt to predict a set of confusion probabilities, p_{ij}, that stimulus i is judged as stimulus j. Some attempts have been made to compare various aspects of these models (Townsend & Ashby, 1982; Townsend & Landon, 1982; Appelman & Mayzner, 1982), but no systematic comparisons have yet been made on the basis of a rigorous statistical criterion. In this paper we compare GOF of these models using the AIC statistic (Akaike, 1974).

All substantive models to be considered in this paper are based on stimulus similarities. The stimulus similarities are then assumed related to the observed form of data, confusion frequencies (probabilities), in a specific way. Thus, there are two major components in the models; a model of stimulus similarities and a model that relates stimulus similarities to confusion probabilities. The former is called the representation model of stimuli (or simply the similarity model) since it states a certain relationship among the stimuli. The latter is called the response model of data, since it links the represented relationship among stimuli to a specific form of data (Takane, 1981; Takane & Carroll, 1982). The following three similarity models will be considered in this paper: the unconstrained similarity model, the euclidean distance model and the unique feature model which is similar to Tversky's (1977) feature matching model. Combined with these similarity models Luce's choice model (Luce, 1959) is used to relate the stimulus similarities to observed confusion probabilities.

Nakatani's (1972) confusion choice model has also been tried as a response model. However, due to space limitation, results of this model will not be presented in this paper. (See Takane & Shibayama, in preparation.) Also, other models of stimulus recognition data, such as the all-or-none activation model and the overlap activation model (Townsend, 1971), have not been attempted, since these models have been consistently found inferior to the unrestricted similarity-choice model (Townsend & Ashby, 1982; Townsend & Landon, 1982). These models are interesting, however, and are worthy of a separate consideration, since they are both special cases of the unrestricted similarity-choice model and of the sophisticated guessing model, of which Nakatani's confusion choice model is also a special case (Pachella, Smith & Stanovitch, 1978; Smith, 1980; Townsend & Landon, 1982).

The three models and their variations were fitted to several sets of real data for goodness of fit comparisons. In this paper we report only one of them in some detail, namely Keren & Baggen's (1981) data on the recognition of segmented numerals. Their data have previously been analyzed by their own model (which is similar to our unique feature-choice model) as well as by the unrestricted similarity-choice model.

THE MODELS TO BE CONSIDERED

In this section we describe, in some detail, the models to be compared. Where appropriate we discuss interesting relationships among them and other existing models.

As stated earlier, models of stimulus recognition data predict confusion probabilities, p_{ij}. Different models are distinguished by different substructures they impose on p_{ij}. There is one model, however, which does not assume any specific substructures on p_{ij}. All the other models are considered special cases of this model.

The Null Model. It is well known that the maximum likelihood estimate of p_{ij} is given by f_{ij}/f_i when no further structural assumptions are made on p_{ij}. This is the least restrictive model for p_{ij} and thus serves as a benchmark model for the stimulus recognition data. In the context of the log-linear model (Bishop, Fienberg & Holland, 1975) this model is analogous to the saturated model. Since f_{ij} is minimal sufficient for p_{ij}, no other models of p_{ij} can possibly fit the data at hand better than this model. However, this model uses the largest number of parameters among the models to be considered. The effective number of parameters in this model is $n(n-1)$ where n is the number of stimuli used in an experiment. We have n^2 parameters, but for each i, we have to have $\sum_j p_{ij} = 1$, hence the effective number of parameters is reduced by n. This number of parameters taken into consideration, this model may not be the best model.

The Unrestricted Similarity-Choice Model. Based on the principle of "independence from irrelevant alternatives", Luce (1959) proposed a general choice model which has been widely used in diverse fields of scientific disciplines. The basic premise of the model is that each choice alternative is associated with a response strength which is assumed invariant over sets of choice alternatives, and that the probability of a particular choice alternative being chosen is proportional to its response strength (Constant Ratio Rule).

In the stimulus recognition context the Constant Ratio Rule leads to

$$p_{ij} = \frac{t_{ij}}{\sum_{k=1}^{n} t_{ik}} , \qquad (1)$$

where t_{ij}, which is not necessarily symmetric, is the response strength of stimulus j, when stimulus i is presented. Model (1) applied to a single confusion matrix, where the response alternatives are common across trials, is called the weak CRR model (Townsdne & Landon, 1982). The weak CRR model, however, is not any more restrictive than the null model. There is a scale indeterminacy in t_{ij}, for each i, and by setting $\sum_k t_{ik} = 1$ to remove this indeterminacy, we obtain $p_{ij} = t_{ij}$, which is equivalent to the null model.

Luce (1963) postulated that the response strength was proportional to both stimulus similarity and response bias, and arrived at the following model:

$$p_{ij} = \frac{w_j s_{ij}}{\sum_k w_k s_{ik}} , \qquad (2)$$

where s_{ij} is the similarity between stimuli i and j, and w_j is the response bias of stimulus j. It is usually assumed that $s_{ij} = s_{ji}$, and that $s_{ii} = 1 \geq s_{ij}$ for all i and $j \neq i$ and $\pi w_j = 1$. The latter restrictions are necessary in order to remove scale indeterminacies.

The above model is motivated by the fact that the more similar two stimuli are, the greater the chance is that they are confused. It is also the case that the greater the bias is for a particular resonse, the higher the probability is that the response is chosen. The response bias here refers to the "tendency of the subject to use some responses more frequently than others" (Smith, 1980). Whereas s_{ij} represents a certain perceptual quality (difference) between stimuli i and j, w_j is assumed to be independent of such a perceptual quality. Rather it is assumed more of a function of such circumstantial factors as frequency of occurrence. Among stimuli having a same degree of similarity those which are more familiar to the subject tend

to be chosen as a response more frequently than less familiar stimuli. Model (2) will be referred to as the unconstrained similarity-choice model in this paper, unconstrained in the sense that no further structural assumption is made on the similarity parameter.

It has been pointed out (Smith, 1982; Townsend & Landon, 1982) that the unconstrained similarity-choice model is equivalent to the log-linear quasi-symmetry model (Caussinus, 1965), which is stated as

$$p_{ij} = g\, a_i\, b_j\, c_{ij} \quad , \tag{3}$$

where $\prod_i a_i = \prod_j b_j = \prod_i c_{ij} = \prod_j c_{ij} = 1$ and $c_{ij} = c_{ji}$. Although parametrizations look different, one-to-one correspondence between (2) and (3) can be easily established by setting $w_j = b_j$, $s_{ij} = c_{ij}/(c_{ii}c_{jj})^{-1/2}$ and $(\sum_k w_k s_{ik})^{-1} = g\, a_i$. An important property of the quasi-symmetry model is the cycle condition, namely

$$p_{ij}\, p_{jk}\, p_{ki} = p_{ji}\, p_{ik}\, p_{kj} \tag{4}$$

(Bishop, Fienberg & Holland, 1975). Furthermore, from the theory of maximum likelihood estimation in the exponential family, of which the log-linear model is a special case, we immediately know that the maximum likelihood estimates in the unrestricted similarity-choice model satisfy: (1) Row and column marginals are always perfectly fit. (2) Discrepancies between observed and predicted confusion frequencies are skew-symmetric, which in turn implies that diagonals are always perfectly fit. These properties can be effectively used to check proper convergence of iterative procedures used to obtain the maximum likelihood estimates. (Note, however, that these properties do not generally hold for models in which s_{ij} and/or w_j are further constrained; e.g., the euclidean distance-choice model and the unique feature-choice model to be discussed in the following sections.)

The unconstrained similarity-choice model has been shown to account for stimulus recognition data very well in many situations (Townsend, 1971; Townsend & Landon, 1982). Keren & Baggen (1981), however, criticized the model on several accounts, of which we mention only those critically relevant for our purposes, and add our comments.

(1) The bias parameters are not perceptually independent of the stimulus similarities. For example, Keren & Baggen found strong

negative correlations between w_j and s_{ij}. Indeed there does not seem to be any good reason to believe that w_j is independent of perceptual qualities of the stimulus. On the contrary, it represents <u>whatever</u> is related to stimulus j, which makes stimulus j more or less plausible as a response. Since s_{ij} is also part of "stimulus characteristics" of stimulus j, it is no wonder that w_j is in some way related to s_{ij}. Recent evidence (Krumhansl, 1978; Podgorny & Garner, 1979) indicates that w_j is affected by such stimulus characteristics as stimulus discriminability (stimulus density around the stimulus) and stimulus complexity (difficulty of encoding the stimulus into a psychological representation) as well as other circumstantial factors. Indeed, psychological status of the bias parameters is at best ambiguous (Keren & Baggen, 1981). However, this does not imply that the bias parameters are unimportant. In fact, as pointed out by Smith (1982), Keren-Baggen's model (our unique feature-choice model), proposed as an alternative to the unrestricted similarity-choice model, also has a restricted form of bias components.

(2) The model uses too many parameters. Indeed it uses $(n-1)(n+2)/2$ parameters, which is the sum of the $n(n-1)/2$ similarity parameters and $n-1$, the number of bias parameters minus 1. Although this number is substantially smaller than $n(n-1)$, the effective number of parameters in the null model, it is still quite large. It seems advisable to attempt to reduce the number of parameters by hypothesizing some substructures on s_{ij}.

(3) There is no substantive theory behind s_{ij}; i.e., no theory regarding how similarity between two stimuli is perceived. Again this motivates some model for s_{ij} (i.e., some specific substructures on s_{ij}). In fact this as well as the argument in (2) has lead Keren & Baggen to develop their own, more parsimonious model for stimulus recognition data using Tversky's feature matching model to account for stimulus similarities.

<u>The Euclidean Distance-Choice Model</u>. The problem now is what substructures we may impose on s_{ij}. One outstanding possibility is the distance model that has been quite successfully used in multidimensional scaling. If we further assume that a function relating the distance to the similarity is an exponential decay function, we obtain

$$p_{ij} = \frac{w_j \exp(-d_{ij})}{\sum_k w_k \exp(-d_{ik})} . \qquad (5)$$

This specialized similarity-choice model will be called the euclidean distance-choice model, since the euclidean distance model is consistently used in this paper. The euclidean distance is formally defined, for a prescribed dimensionality A, as

$$d_{ij} = \left\{ \sum_{a=1}^{A} (X_{ia} - X_{ja})^2 \right\}^{1/2} , \qquad (6)$$

where x_{ia} and x_{ja} are coordinates of stimuli i and j, respectively, on dimension a. The effective number of parameters in the euclidean distance model depends on the dimensionality, and is given by nA-A(A+1)/2. This is usually much smaller than n(n-1)/2 for s_{ij}. The total number of parameters in the euclidean distance-choice model is thus (n-1)+nA-A(A+1)/2.

The euclidean distance-choice model is not at all new, although to the best of the authors' knowledge it has never been fitted directly (but see Getty, et al., 1979; Heiser, 1985.) As early as in 1957 (which even predates Luce's choice model) Shepard proposed a model identical in form to (5) in the stimulus generalization context. He also derived a formula for obtaining an estimate of d_{ij} from observed confusion probabilities, which may be used to obtain initial estimates of x_{ia}.

It is interesting to note that model (5) can also be derived from a seemingly unrelated model. Krumhansl (1978; see also Bentler & Weeks, 1978; Takane & Sergent, 1983; Carroll, 1983; Winsberg & Carroll, 1984) proposed the distance-density model to account for the effect of stimulus density around a particular stimulus. The stimulus density affects stimulus discriminability, which in turn causes violations of minimality and symmetry (Tversky, 1977) required of the distance. In order to account for the effect she proposed a model which is a combination of regular distance function (usually euclidean) and density components. Specifically, her model is:

$$\tilde{d}_{ij} = d_{ij} + ac_i + bc_j , \qquad (7)$$

where $\tilde{d}_{ij}$ is the "distance-density" between stimuli i and j, d_{ij} is the usual euclidean distance between stimuli i and j, c_i and c_j (≥ 0) are density parameters, and a and b are the weights to account for asymmetry. The density parameters are supposed to take a larger value when the stimulus density around a particular point is higher, and consequently stimulus discriminability is lower for the stimulus. When there are a lot of stimuli similar to the stimulus, stimulus similarities between the stimulus and all other stimuli tend to be "diluted."

Podgorny and Garner (1979) gave a different interpretation to Krumhansl's model in the choice reaction time context. Since it takes more time to judge less similar stimuli the same, and also more time to identify more complicated stimuli, the density parameter in Krumhansl's model should be, in some way, related to the stimulus complexity. Under this new interpretation c_i does not even have to be nonnegative. Furthermore, we may completely distinguish between stimulus complexity and response complexity, and replace ac_i and bc_j by less restrictive parameters, a_i and b_j. Then the model becomes

$$\tilde{d}_{ij} = d_{ij} + a_i + b_j \quad . \tag{8}$$

Takane and Sergent (1983) call a_i and b_j stimulus specificities following the terminology used in common factor analysis. Carroll (1983) called the above model the "common dimension model."

If we assume $t_{ij} = \exp(-\tilde{d}_{ij})$ in (1), it can be written that

$$p_{ij} = \frac{\exp(-\tilde{d}_{ij})}{\sum_k \exp(-\tilde{d}_{ik})} = \frac{w_j \exp(-d_{ij})}{\sum_k w_k \exp(-d_{ik})} \tag{9}$$

where $w_j = \exp(-b_j)$ which is identical to model (5). Thus, one interpretation of the bias parameter is stimulus simplicity. Another interpretation is stimulus discriminability (which is inversely related to the stimulus density). This explains why w_j tends to be negatively correlated to a_{ij}. If stimulus j is similar to many stimuli (many large a_{ij}'s), the stimulus density around the stimulus is high, its discriminability low, and the response bias tends to be small. It seems that the bias parameter is a many-faceted entity.

A comparison between the euclidean distance-choice model (5) and the unique feature-choice model (Keren-Baggen's model) to be described in the next section is interesting, particularly in the light of the relationship between Shepard's model (the euclidean distance-choice model) and Krumhansl's distance-density model, as illustrated above. Krumhansl (1982) did not seem to be aware of this relationship in discussing the relationship between her model and Keren-Baggen's.

The Unique Feature-Choice Model. The standard distance model is not the only way to constrain the similarity parameters or the response strength. For example, if stimuli are characterized by a set of identifiable features, stimulus dissimilarity between two stimuli may be defined as a linear combination of features not commonly possessed by the two. Let

$$y_{ia} = \begin{cases} 1, & \text{if stimulus } i \text{ possesses feature } a, \\ 0, & \text{otherwise,} \end{cases}$$

and define $X_{ija}^{(D_1)} = y_{ia}(1-y_{ja})$ and $X_{ija}^{(D_2)} = y_{ja}(1-y_{ia})$. Then $X_{ija}^{(D_1)}$ takes the value of one, if stimulus i, but not stimulus j, possesses feature a, and zero, otherwise. Similarly, $X_{ija}^{(D_2)}$ takes the value of one, if stimulus j, but not stimulus i, possesses feature a, and zero, otherwise. The linear combination of the unique features may then be written as

$$\tilde{g}_{ij} = \sum_a (X_{ija}^{(D_1)} b_a + X_{ija}^{(D_2)} c_a), \tag{10}$$

where $b_a (\geq 0)$ and $c_a (\geq 0)$ are the weights representing importance of unique feature a in overall dissimilarity. Note that in general $\tilde{g}_{ij} \neq \tilde{g}_{ji}$. If we replace $\tilde{d}_{ij}$ in (9) by $\tilde{g}_{ij}$, or equivalently assume $t_{ij} = \exp(-\tilde{g}_{ij})$ in (1), we obtain

$$p_{ij} = \frac{\exp(-\tilde{g}_{ij})}{\sum_k \exp(-\tilde{g}_{ik})} . \tag{11}$$

Model (11) is called the unique feature-choice model in this paper.

The above model is essentially equivalent to the general version of Keren-Baggen's (1981) model, which was initially derived from Tversky's (1977) feature matching model. In Tversky's model features commonly possessed by two stimuli are also supposed to take part in overall (dis)similarity between the two stimuli. The common features may be indicated by $X_{ija}^{(C)} = v_{ia} v_{ja}$. However, model (11) is invariant over the transformation of the form, $\tilde{g}_{ij} \longrightarrow \tilde{g}_{ij} + \tilde{g}_i$ (where $\tilde{g}_i$ is specific to i), and as pointed out by Smith (1982), $X_{ija}^{(C)}$ and $X_{ija}^{(D_1)}$ are completely redundant in the context of stimulus recognition experiments. This can be easily seen by pointing out $X_{ija}^{(C)} = v_{ia} - X_{ija}^{(D_1)}$ and y_{ia} is constant for specific i. Thus, either one of $X_{ija}^{(C)}$ and $X_{ija}^{(D_1)}$

can be dropped without loss of generality. We arbitarily dropped $X_{ija}^{(C)}$ in (10). The rank deficiency and consequent parameter indeterminacies in the original Keren-Baggen's model were overlooked by Keren & Baggen (1982; also Krumhansl, 1982).

Smith (1982) has pointed out that $\exp(-\tilde{R}_{ij})$ in (11) can be rewritten in the form, $w_j s_{ij}$, so that the unique feature-choice model is indeed a special case of the unrestricted similarity-choice model with particular constraints on w_j and s_{ij}. That is,

$$w_j = \prod_a (u_a)^{y_{ja}}$$

and

$$s_{ij} = \prod_a (v_a)^{|y_{ia} - y_{ja}|^r}, \quad r \geq 1,$$

where $\ln u_a = (b_a - c_a)/2$ and $\ln v_a = -(b_a + c_a)/2 (\leq 0)$. Furthermore,

$$-\ln s_{ij} = \sum v_a^* |y_{ia} - y_{ja}|^r, \tag{12}$$

where $v_a^* = -\ln v_a (\geq 0)$, can be interpreted as the r-th power of the Minkowski power distance with the power equal to r. When r=1, the unique feature-choice model is equivalent to the city-block distance-choice model with stimulus features serving as (prescribed) dimensions. Each dimension(=feature) has only two levels, $y_{ia}=1$ or 0, according to presence or absence of the feature.

It follows from (10) that

$$\tilde{R}_{ij} = \sum_a \left\{ (X_{ija}^{(D_1)} + X_{ija}^{(D_2)}) \left(\frac{b_a+c_a}{2}\right) + (X_{ija}^{(D_1)} - X_{ija}^{(D_2)}) \left(\frac{b_a-c_a}{2}\right) \right\}.$$

Note that $X_{ija}^{(D_1)} + X_{ija}^{(D_2)} = |v_{ia}-v_{ja}|^r$ and $X_{ija}^{(D_1)} - X_{ija}^{(D_2)} = v_{ia} - v_{ja}$, so that, with the above definations of w_j, w_i and s_{ij},

$$\exp(-\tilde{R}_{ij}) = \frac{w_j}{w_i} s_{ij},$$

which in turn is equal to p_{ij}/p_{ii} according to (2). Since $1/w_i$ cancels out in the numerator and the denominator, (11) reduces in form to (2). It is interesting to note that $\ln s_{ij}$ pertains to symmetric and $\ln w_j - \ln w_i$ to skew-symmetric part of $-\tilde{R}_{ij}$.

When it is assumed $b_a = bc_a$ (the weights applied to the same features in $X_{ija}^{(D_1)}$ and $X_{ija}^{(D_2)}$ are proportional) in (10), we obtain

$$\tilde{g}_{ij} = \sum_a (X_{ija}^{(D_1)} b + X_{ija}^{(D_2)}) c_a, \tag{13}$$

which is equivalent to the restricted version of Keren-Baggen's model. Whereas the general unique feature-choice model has 2m independent parameters (assuming $X_{ija}^{(D_1)}$ and $X_{ija}^{(D_2)}$ are all linearly independent), model (13) has only m+1 parameters, where m is the number of features. If further $b_a = c_a$ is assumed in (10) (or b=1 is assumed in (13)), we obtain

$$\tilde{g}_{ij} = \sum_a (X_{ija}^{(D_1)} + X_{ija}^{(D_2)}) c_a, \tag{14}$$

$w_j = 1$ for all j and $-\ln s_{ij} = \sum_a (c_a^*) |y_{ia} - y_{ia}|^r$, where $c_a^* = -\ln c_a$, which is essentially how similarity between two stimuli is defined in Medin & Schaffer's (1978) cue context model for classification learning.

The fact that the unique feature-choice model is a special case of the unrestricted similarity-choice model, and that it is also a special case of the distance-choice model, makes comparisons among these models even more interesting.

Goodness of Fit Evaluation. As pointed out earlier the different models are basically distinguished by different substructures they impose on p_{ij}. However, once p_{ij} is specified, the likelihood of the total set of observations is stated in the same way (in the form of product multinomials) for all the models; i.e.,

$$L = \prod_i \prod_j (p_{ij})^{f_{ij}} .$$

One advantage of the maximum likelihood method is that we can calculate the AIC statistic (Akaike, 1974) for relatively straight forward goodness of fit comparisons. This statistic is defined as

$$AIC = -2 \ln L^* + 2q ,$$

where L^* is the maximized value of L and q is the effective number of parameters in the fitted model. The smaller the value of AIC, the better the fit of the model.

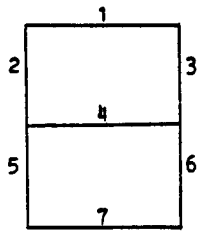
The maximized log likelihood indicates a goodness of fit of a model to the data at hand. In general we can improve this kind of goodness of fit by including more parameters in the model. However, it is not necessarily the case that this inflated model performs better for future observations, because a larger number of parameters tend to produce less reliable parameter estimates. This leads to the idea that the maximum likelihood should be penalized for additional use of parameters. Specifically, the log likelihood is penalized by -q in order to obtain an unbiased estimate of the expected log likelihood, which is the basic construction of the AIC statistic. The use of the AIC statistic for model evaluations has successfully been demonstrated in a variety of psychometric models (Takane, 1981; Takane & Carroll, 1981; Takane & Sergent, 1983).

EMPIRICAL RESULTS

The models described in the previous sections were applied to several data sets. In this paper we report only one of them in some detail, namely Keren & Baggen's (1981) data. (For other results, see Takane & Shibayama, in preparation.) Their data were previously analyzed by Smith (1982), using both the unrestricted similarity-choice model and the unique feature-choice model. Our results are somewhat different from his, since the former are based on Keren & Baggen's original data, while the latter on the data recovered from observed proportions reported in Keren & Baggen (1981) by multiplying the total number of replications per stimulus. In this recovery process, existence of missing data was completely ignored.

Keren & Baggen (1981) conducted a recognition experiment with ten segmented numerals (digits 0 through 9). Seven line segments (three horizontal and four vertical) were arranged in the shape of 8 (see Figure 1), and each of ten single-digit numerals was defined by a subset of the seven line segments. For example, 2 is composed of line segments 1, 3, 4, 5 and 7, and 4 of line segments 2, 3, 4 and 6, etc.

Figure 1. The seven segmental features



Each numeral was presented 36 times to each of eight subjects for a total of 288 responses to each stimulus. Stimulus exposure time was adjusted for each subject to make the average rate of correct responses about .7 for each subject. The data were aggregated over the subjects, assuming there were no substantial individual differences. "No response" was allowed to avoid guessing. There were about 4% of the trials in which the "no response" was elicited. Those trials were omitted from subsequent analyses. Thus, strictly speaking the total number of "valid" responses, f_i, for each i should be treated as a random variable. In the present analysis, however, f_i was treated as fixed. This can be justified, since so far as we assume a same model for f_i, that part of the likelihood function (pertaining to this model) cancels out in the model comparison process, and the likelihood (14) may be treated as if it were a conditional likelihood, conditional upon f_i.

Table 1 summarizes the results of fitting the models to the data. All the three models described in the previous sections were applied. Both the general and the restricted versions, (10) & (13), of the unique feature model were fitted. Features used in the unique feature model are the seven segmental features used to construct the stimuli (7-feature case). Two additional features, deemed psychologically important, were originally conceived by Keren & Baggen and analyses were repeated including these two features (9-feature case). Those additional features were "open to the left" and "open to the right." Digits 2, 3, 5, 7 and 9 are open to the left, while digits 2, 5, and 6 are open to the right. (While this definition conforms with that of Smith's (1982), there is no guarantee that it agrees with Keren & Baggen's (1981) who did not provide the information.)

The values of the AIC statistic indicate that the unrestricted similarity-choice model is the best fitting model. The four-dimensional euclidean solution is the best solution

Table 1. Summary of GOF statistics for Keren-Bagpen's data

0.	Null Model	17.4 (90)
Similarity Model		
1.	Unrestricted Similarity Model	5.6* (54)
2.	Euclidean Distance Model	
	dim=2	220.3 (26)
	dim=3	79.2 (33)
	dim=4	35.9 (39)
	dim=5	40.9 (44)
3.	General Unique Feature Model	
	7 features	199.3 (14)
	9 features	95.3 (16)
4.	Restricted Unique Feature Model	
	7 features	273.7 (8)
	9 features	122.5 (10)

AIC-6170 (top)
effective number of model parameters in parentheses (bottom)
*minimum AIC solution

among the euclidean distance-choice models obtained by systematically varying the dimensionality from two to five. But even the best euclidean solution is nowhere near the performance of the unrestricted similarity-choice model. Increasing the dimensionality beyond four dimensions does not improve the GOF of the euclidean distance model. This suggests that the euclidean distance model is not appropriate for the data.

This is quite a contrast to other stimulus recognition situations in which the euclidean distance model performed reasonably well. In these situations, however, sets of stimuli employed are characterized by sets of attributes on which the stimuli are distinguished in the amounts of the atributes they possess (e.g., tones defined by intensity, frequency and duration in Hodge & Pollack's (1962) data). The attributes always exist. Only their amounts differ. In contrast the digits used in Keren & Baggen's experiment are made up of features which are either present or absent. The binary features are not easily amenable to a simple dimensional organization. (See, however, the argument leading to (12).)

The above explanation for the poor fit of the euclidean distance model to Keren & Baggen's data, however, is not as straightforward as it seems. For example, Sergent & Takane (in preparation) found a remarkably good fit of two-dimensional euclidean representations of the same set of stimuli using two-choice reaction time data. This suggests that the "nature" of stimuli alone does not completely determine the best representation. It is more likely that the "nature" of stimuli interacts with the kind of similarity measures and stimulus presentation conditions employed, to determine the best representation of the stimuli. The stimulus recognition data used in the present study are usually taken under much more adverse stimulus presentation conditions (in order to create confusions purposely) than the reaction time data. The latter are usually taken under clear viewing, longer stimulus exposure conditions in order to minimize the number of errors (Takane & Sergent, 1983). As such, the two measures may reflect different aspects of underlying processes (Santee & Egeth, 1982; Sergent & Takane, 1985).

There are other conceivable reasons for the poor fit of the euclidean distance model. The conditions under which Keren & Baggen's data were obtained are still much more favorable than those typical of stimulus recognition data. As a consequence the data are extremely diagonally dominant. The average diagonal entry (=correct identification rate) is approximately .7 with ten

stimuli, which makes off-diagonal entries (proper confusion rates) rather small. This in turn makes the stimulus more discriminable, less similar to each other. However, large distances are precisely where the euclidean distance model tends to break down, and where bad fits count most. A larger dimensionality may be necessary to accommodate larger distances. This is indicated by the limiting situation in which the data are perfectly diagonal. In such a case n stimuli are represented as an infinitely large n-1 dimensional simplex. Under the assumption that $s_{ij} = \exp(-d_{ij})$, the effect of the viewing conditions should come in the form of a power function on s_{ij} (i.e., $(s_{ij})^b$), in order to be completely absorbed by the uniform expansion or contraction of the stimulus configuration. However, no systematic investigation has yet been conducted to assess the exact nature of the various experimental conditions on the recognition probabilities. The best representation may still differ depending on the experimental conditions, even if the same measure of stimulus similarity is used, since the different experimental conditions may require different processing strategies.

The goodness of fit of the unique feature model is rather disappointing. The best fit is obtained in the 9-feature general version of the model. (Note that the effective number of parameters for this solution is 16 instead of 18. This is because of linear dependencies among X_{ija}'s.) Still, it is much worse than that of the unconstrained similarity-choice model. It is even worse than that of the four-dimensional euclidean distance model. This is consistent with Smith (1982), but contrary to Keren & Baggen's conclusion that their model fitted the data reasonably well; i.e., their model could successfully reduce the number of parameters in the unconstrained similarity-choice model without significantly impairing the goodness of fit of the model. Indeed, the number of parameters was greatly reduced. However, the price they had to pay for this seems to be too high. Although what is a satisfactory model is ultimately a matter of subjective judgment, the use of a relatively insensitive criterion for goodness of fit is often misleading. It tends to overlook relatively small, yet empirically meaningful regularities in the data.

However, all these by no means imply that the unique feature model is hopeless. For example, with many fewer parameters the two versions of the 9-feature unique feature model performed much better than the two-dimensional euclidean distance model. This suggests that if only we can find an appropriate set of features that define stimulus dissimilarities, the unique feature model may turn out to be the better model. In particular the list of features used in this study may be quite incomplete. Furthermore, except those two appended features (which may be viewed as interactions among the seven segmental features), features were assumed to be independent. It is quite possible that some of these segments do interact. For example, a combination of segments 3 and 6 may define a new feature. It is rather arbitrary to take these segments always as separate features.

One disadvantage of the unique feature model is that a set of features defining the stimulus dissimilarity have to be known in advance in order to fit the model. There are 2^n possible features for n stimuli, which are all possible subsets of the n stimuli. A systematic way to obtain a minimal and sufficient set of features from this list is urgently needed. The kind of error analysis done by Smith (1982) may be quite useful in detecting features effective to account for stimulus similarities underlying the stimulus recognition data. For examle, stimulus pairs 1-6 (also 6-1), 5-1 (also 1-5), 4-2, and 1-2 have large standardized residuals using the nine features. This gives some insight into what is possibly wrong with the nine features (e.g., Digit "1" should not be defined solely in terms of features 3 and 6. Features 2 and 5 are equally good to define "1".) and what additional features may have to be taken into account. It would also be helpful to devise a computer program to do multiple comparisons interactively in the context of the log-linear quasi-symmetry model.

CONCLUDING REMARKS

In this paper we presented a fairly elaborate analysis of a set of stimulus recognition data. Although our conclusions differ, we support Keren-Baggen's venture to find a more parsimoneous representation for the digits. Although our attempt is not yet successful, we contend our general methodology, the maximum likelihood estimation of model parameters and the GOF

comparisons through AIC, is quite effective in detecting what components and further studies are necessary to develop a yet better model for the stimulus recognition data.

One potential obstacle in our attempt is the possible individual differences. Since the data were provided in an aggregated form, there was no way to detect the possible systematic individual differences in Keren & Baggen's data. However, quite often the subjects differ in their perceptual ability, response style, response strategy, etc. which tend to influence the results of stimulus recognition experiments. For example, Sergent & Takane (in preparation), using the reaction time data, found remarkable individual differences in the way the digit stimuli are processed by two subjects. Although in Keren & Baggen's experiment stimulus display conditions were adjusted for each subject so that the average error rate was approximately equal across the subjects, it by no means guarantees complete elimination of the individual differences. If such is the case, an error model (in the form of product multinomials) is not completely justified. Ideally individual (unaggregated) data have to be analyzed. Unfortunately most of the published data are in an aggregated form.

REFERENCES

Akaike, H. A new look at the statistical model identification. IEEE Transactions on automatic control. 1974, 19, 716-723.

Appelman, I.E., & Mayzner, M.S. Application of geometric models of letter recognition: Distance and density. Journal of Experimental Psychology: General, 1982, 111, 60-100.

Bentler, P.M., & Weeks, D.G. Restricted multidimensional scaling models. Journal of Mathematical Psychology, 1978, 17, 138-151.

Bishop, Y.M.M., Fienberg, S.E., and Holland, P.W. Discrete multivariate analysis: theory and practice. Cambridge, Mass: The MIT Press, 1975.

Carroll, J.D. Common dimension analysis. Paper presented at the multidimensional data analysis workshop. Paris, 1983.

Caussinus, H. Contribution a l'analyse statistique tableau de correlation. Ann. Fac. Sci. University Toulouse, 1965, 29, 77-182.

Getty, D.J., Swets, J.A., Swets, J.B., and Green, D.M. On the prediction of confusion matrices from similarity judgments. Perception & Psychophysics, 1979, 26, 1-19.

Heiser, W.J. On the selection of a stimulus set with prescribed structure from empirical confusion frequencies. RR-85-08, Department of Data Theory, University of Leiden, 1985.

Hodge, M.H., & Pollack, I. Confusion matrix analysis of single and multidimensional auditory displays. Journal of Experimental Psychology, 1962, 63, 129-142.

Keren, G., & Baggen, S. Recognition models of alphanumeric characters. Perception & Psychophysics, 1981, 29, 234-246.

Krumhansl, C.L. Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. Psychological Review, 1978, 84, 445-463.

Krumhansl, C.L. Density versus feature weights as predictors of visualk identifications. Comments on Appelman and Mayzner Journal of Experimental Psychology: General, 1982, 11, 101-108.

Luce, R.D. Individual choice behavior: A theoretical analysis. New York: Wiley, 1959.

Luce, R.D. Detection and recognition. In R.D. Luce et al. (Eds.), Handbook of mathematical psychology (Vol. 1). New York: Wiley, 1963.

Medin, D.L., & Schaffer, M.M. Context theory of classification learning. Psychological Review, 1978, 85, 207-238.

Nakatani, L.H. Confusion-choice model for multidimensional psychophysics. Journal of Mathematical Psychology, 1972, 9, 104-129.

Pachella, R.G., Smith, J.E.K., & Stanovich, K.E. Qualitative error analysis and speeded classification. In Castellan, N.J., and Restle, F. (Eds.) Cognitive theory (Vol.3). Hillsdale, N.J.: Earlbaum, 1978.

Podgorny, P., & Garner, W.R. Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. Perception & Psychophysics, 1979, 26, 37-52.

Santee, J.L., & Egeth, H.E. Do reaction time and accuracy measure the same aspects of letter recognition? Journal of Experimental Psychology: Human Perception and Performance, 1982, 8, 489-501.

Sergent, J., & Takane, Y. Structures in two-choice reaction time data. Submitted for publication, 1985.

Sergent, J., & Takane, Y. Individual differences in structures of two-choice reaction time data. A manuscript in preparation.

Shepard, R.N. Stimulus response generalization: A stochastic model relating generalizaiton to distance in psychological space. Psychometrika, 1957, 22, 325-345.

Smith, J.E.K. Models of identification. In Nickerson, R.S. (Ed.) Attention and performance VIII. Hillsdale, N.J.: Earlbaum, 1980.

Smith, J.E.K. Recognition models evaluated: A commentary on Keren & Baggen. Perception & Psychophysics, 1982, 31, 183-189.

Takane, Y. Multidimensional successive categories scaling: A maximum likelihood method. Psychometrika, 1981, 46, 9-28.

Takane, Y., & Carroll, J.D. Nonmetric maximum likelihood multidimensional scaling from directional rankings of similarities. Psychometrika, 1981, 46, 389-405.

Takane, Y., & Sergent, J. Multidimensional scaling models for reaction times and same-different judgments. Psychometrika, 1983, 48, 393-423.

Takane, Y., & Shibayama, T. Structures in stimulus recognition data. (In preparation).

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

Townsend, J.T., & Ashby, F.G. Experimental test of contemporary mathematical models of visual letter recognition. Journal of Experimental Psychology: Human Perception and Performance, 1982, 8, 834-864.

Townsend, J.T., & Landon, D.E. An experimental and theoretical investigation of the constant-ratio rule and other models of visual letter confusion. Journal of Mathematical Psychology, 1982, 25, 119-162.

Tversky, A. Features of similarity. Psychological Review, 1977, 84, 327-352.

Winsberg, S., & Carroll, J.D. A nonmetric method for a multidimensional scaling model postulating common and specific dimension. Paper presented at the Psychometric Society Meeting, 1984.

DECONFUSING CONFUSION MATRICES

DISCUSSION OF THE PAPER BY TAKANE AND SHIBAYAMA

Ivo W. Molenaar
University of Groningen

The paper 'Comparison of models for stimulus recognition data' by Takane and Shibayama is highly welcome: researchers have analyzed data of this type for several decades now, but systematic comparisons of different models have been rare, and not all contributions in the past have been based on clearly specified statistical models and efficient estimation procedures. Hopefully the present comment will be helpful in obtaining still more progress and clarity. After some general questions about stimulus recognition data, this comment discusses one by one the models introduced by T & S, with a few remarks on the specific data that they reanalyzed and a sketch of a very crude nonparametric model.

ESTABLISHING CONFUSION

In a typical paper on stimulus recognition data, an experiment of the following type is reported. Some four to eight students are used as experimental subjects. The stimulus set contains n elements, where n varies between 8 and 26 depending on the experiment. Stimuli can be e.g. tones, or shades of red, that are manipulated to vary in a few relevant dimensions. Other experiments use the digits 0,1,...9 or the 26 letters of the alphabet as stimuli. After a short training session, each subject is repeatedly presented each of the stimuli in an unpredictable order. Presentation conditions are degraded to the extent that most subjects, when presented stimulus i, say, correctly identify it as i in some 30-90

percent of the replications and give a different answer in the remaining cases. The responses are summarized in a row-conditional n x n 'confusion matrix' P of which the i,j element p(j|i) denotes the fraction of presentations of stimulus i for which the subject thought that stimulus j was presented (the T & S notation p_{ij} for p(j|i) hides its asymmetry). The aim is a parsimonious description of this matrix leading to more insight in the occurrence of confusions during the human perceptual process.

As a statistician without much involvement in the stimulus recognition literature, I feel inclined to ask some preliminary questions, about the area as a whole rather than about the T & S paper.

Why would one and the same class of mathematical models for the P-matrix be suitable for all such experiments ? Perception is sometimes auditory, sometimes visual. The stimulus set is well known (digits, letters) or specific to the experiment (tones, shades of red). In the latter case the experimenter knows that stimuli differ on a few dimensions only (it is not clear from the publications whether the subject knows this too).

Why would experiments of this type be externally valid for those real life situations in which confusions occur? In real life a subject will usually have some external information on the base rate at which the stimuli occur and on the transition probabilities between subsequent stimuli. Reading segmented numerals from a digital clock I know that some digits are impossible in some positions, and I may use my vague knowledge of the hour of the day. Most experiments present the stimuli in equal frequency in random order, removing such external clues. This may hold for identifying a telephone number mentioned across a noisy channel, but I could not find many similar situations.

Does the experimenter want to describe what people do do, or what people should do ? Subjects may be instructed to give instantaneous and spontaneous answers, or to strive for a rational strategy. In the latter case it is vital to report how much the subjects knew about the stimuli themselves and on the frequency and order of presentation. It is disap-

pointing that most publications say almost nothing about the instructions to the subject, and even omit the total number of times that each stimulus was presented.

Is there any cogent reason why all experimental subjects would produce the same P-matrix apart from random fluctuations ? T & S have added a paragraph to their last section on this issue, after my oral discussion in Cambridge. Most researchers in this area aggregate across subjects, at best stating verbally the absence of large differences between subjects. I am not inclined to believe a prima vista that 36 subjects seeing each numeral eight times lead to the same aggregated result that Keren & Baggen obtained for eight subjects 36 times each. The simple likelihoods used do not distinguish generalization to future trials from generalization to future subjects.

MODELING CONFUSION

T & S present various models that will be individually commented. It is good that their presentation stresses the relations between the models, and underlines both their statistical and their psychological aspects, fully in the spirit of Takane's admirable paper in Psychometrika 1981. The distinction between the similarity model and the probability transform is very useful, although only the combination of both can be tested. It is certainly illuminating to study the differences in fit for the Keren & Baggen data, but I should like to stress that I see no reason why one model would be uniformly superior to the others across many data sets.

THE SATURATED MODEL. I agree that this is a useful baseline or null model. T & S remark in their goodness of fit section that 'it is not necessarily the case that this inflated model performs better for future observations' : I would go a step further and predict that this will be rarely the case as soon as there is any system in the data. Keren & Baggen (1981, p.234) explicitly mention the desirability of a more parsimonious and psychologically more meaningful description.

UNRESTRICTED SIMILARITIES. There are many reasons to multiply the similarities s(i,j) by response bias parameters w(j) when modeling the choice probability p(j|i). In some contexts this has a natural Bayesian interpretation. If w(j) is the estimated base rate or prior probability that j is displayed, and s(i,j) is the probability that the degraded i presented is in fact j, one could say that the subject uses the posterior probabilities in the choice task. This remark is a digression, because in this interpretation s(i,j) would probably not be symmetric, for the data sets discussed the subjects had no reason to assume unequal prior probabilities, and choosing according to the posterior probabilities is suboptimal for most of the plausible loss functions. Nevertheless, this interpretation may be psychologically meaningful in some real life tasks. If I see a degraded symbol in a text, which I consider to be a 'Q' with probability .7, say, and an 'O' with probability .3, it is wise to take into account that 'O' has a much higher base rate than 'Q'. This leads to different strategies for Latin, English and Dutch texts, on rational grounds.

EUCLIDEAN DISTANCES. I agree that the distance model has been successfully used in multidimensional scaling as a parsimonious representation of a distance matrix. This was often achieved by the nonmetric variant, however. T & S use the metric variant plus the exponential decay. This is rather restrictive, which may account for the poor performance in Table 1. The extension to 'distance densities' is interesting.

UNIQUE FEATURES. It is surprising to see how the number of common features turns out to be irrelevant. Below (12) T & S state that the unique feature model 'is equivalent' to the city block distance model when r=1. This is mathematically correct, but the interpretation is quite different. In the distance case all coordinates are free parameters; the unique feature case assumes that the coordinates y_{ia} are known, leaving only two free parameters per dimension : one could say $b_a + c_a$ as the relevance of the a-th dimension for the similarities and $b_a - c_a$ as its relevance for the asymmetry. For me the major weakness of models of this kind is that the investigator must specify in advance which set of m independently acting dichotomous features adequately describes the

perceived similarities. It must also be known for each stimulus and feature whether the former possesses the latter. For the Keren & Baggen data, the seven line segments are plausible candidates for the first seven features. They add 'open to the right' (like 6) and 'open to the left' (like 9) as 'psychologically meaningful' other features. I would be surprised if this set of nine features were superior to many other sets one could pick. It would be interesting to investigate the model fit for other sets of features, perhaps not even including the seven segments.

WHAT IS A GOOD MODEL ?

There is a vast literature on this topic. T & S opt for the AIC index, which is defensible but leaves undiscussed others like the Schwarz information criterion and the James & Mulaik & Brett parsimonious fit index, as well as model modification indices proposed by Jöreskog, Bentler and Bonett and others. If we leave the domain of the likelihood we have still more choices, among which the criteria used by Keren & Baggen, who emphasize interpretability and psychologically meaningful processes rather than pure fit.

In this respect it is interesting that the Nakatani model, to be discussed in a separate paper by T & S, supposes that some subset of responses is dismissed as 'too dissimilar'. This harmonizes with 'bounded rationality' discussed in the analysis of human decision makers. For the complementary subset of admissible responses, however, Nakatani's probabilities are proportional to the response attractiveness parameters b(j) without further influence of the similarities s(i,j). I should like to propose to use the product b(j)s(i,j) rather than b(j) in this selection from the subset of admissible responses. It might also be instructive to interrogate subjects whether certain responses are indeed ruled out by them before deciding between the remaining ones.

For the segmented numerals example, and also for the alphabet as a stimulus set, I wonder whether discrepancies between the symbols as

presented and other more usual fonts have influence on the subject's decisions. The largest off-diagonal probability (not displayed by T & S, by lack of space) is that 27 percent of the presentations of '7' lead to the response '1'. Here a font influence becomes plausible: the segmented seven has a weird appearance.

Again by lack of space, there is no discussion of the fitted values with their residuals, or of the parameter estimates. This might have reinforced the useful remarks made by T & S on the misfits. It is interesting to note that the percentage correct varies from 55% for 9 to 88% for 1, and that there are marked asymmetries both before and after correcting for different $p(i|i)$.

Analyzing confusion data for the capitals of the alfabet, Heiser (1985) obtains some empirical support for the exponential transformation from distance to probability. To my knowledge there are no studies in which externally established similarity data are compared to the fitted $s(i,j)$ values. Calling the latter 'similarities' means that similarity is defined as 'what explains confusion within a specific model'; it would be interesting to see if this claim can be substantiated.

NONPARAMETRIC MODELS

Keren and Baggen (1981, p.241) give a graph showing that the probability of confusion decreases with the number of distinguishing segments (i.e. segments present in exactly one of the two numerals). Smith (1982, p.185) discusses the same point, and adds that under tachistoscopic conditions a vertical line will be taken for a '1' whether at the left or at the right of the field (this leads to less distinguishing segments for the pair 1,6).

A very simple nonparametric model just predicts that confusion probabilities within each line of the matrix are partially ordered according to the number of distinguishing features. For responses equally far from the stimulus in this counting measure no order prediction is made. Such predictions are confirmed to a large extent in both the Keren & Baggen

and the Hodge & Pollack data. Let C denote the number of correct predictions and D the number of cases where the reverse order was observed, disregarding ties in the data or in the predictions like in Kendall's tau-b. The almost trivial prediction that $p(i|i)$ exceeds any other $p(j|i)$ was omitted. Then in the Keren & Baggen 7 features case one tries to make 36 predictions about the order of two probabilities for each of the ten numerals. For all 360 one obtains C=229, D=48 and 83 ties. Per numeral the C/D ratios for i=0,1,...,9 are 16/12, 24/3, 22/6, 21/1, 20/8, 22/8, 26/5, 25/1, 28/0, 25/4. When 0 is presented no less than 8 percent of the answers is '1', although 0 has four additional features; this contributes to the disappointing 16/12 ratio. Comparable results are found for the Hodge Pollack data, with less ties.

This primitive model can be viewed as a nonmetric variant of the T & S unique feature (10) and (12) with r and all b- and c-parameters equal to one. It is obvious that it has far less predictive power than a parametric one. It does not predict why some $p(i|i)$ are larger than others, but the same holds for the quasi-symmetry model where the maximum likelihood estimates exactly reproduce the diagonal elements. The rough nonparametric model could be useful, however, in a preliminary screening which of the many possible feature sets is promising for a parametric analysis. The nonparametric unidimensional unfolding models presented by Van Schuur and Van Blokland at the Fourth European Meeting of the Psychometric Society in Cambridge can be viewed as born from the same desire to check some relevant order relations by simple means before passing to a specific parametric model. Frequent violations from the predicted order are a general alarm signal. An occasional gross violation may indicate a specific event, like the 7,1 confusion mentioned above.

More generally it may be wise to strive for insight via the study of some large residuals in a simple model with substantive plausibility rather than to add more and more parameters until a model fully fits the data. Although they do not analyse residuals in detail, T & S will probably agree with this plea. The tradeoff between parsimony and fit should not be too much shifted by our ability to fit complex models with our powerful computers : small is beautiful.

A REPLY TO IVO MOLENAAR

Yoshio Takane and Tadashi Shibayama

We thank very much Ivo Molenaar for his penetrating comments on our paper. Although there are numerous points raised, we will have to concentrate on only a few of them here.

One of the major problems relates to possible individual differences in the confusion process. That is, p_{ij} (confusion probability), may differ from one subject to another. The problem seems that in most cases these individual differences are not substantively interesting, and yet their magnitude is often too large to be safely ignored in modeling stochastic components of the data. Over the past ten years or so, however, more researchers have come to realize the potential danger of aggregating data (e.g., Morgan, 1974; Smith, 1980; Townsend & Ashby, 1982; Townsend & Landon, 1982).

We may use the same model comparison idea (as illustrated in the paper) to test whether there are indeed substantial individual differences, provided that unaggregated data are available. Let $p_{ij}^{(k)}$ represent the probability that stimulus i is judged as stimulus j by subject k, and let $f_{ij}^{(k)}$ be the observed frequency corresponding to $p_{ij}^{(k)}$. Then under the individual differences hypothesis we have

$$L_{ID} = \prod_{k} \prod_{i,j} (p_{ij}^{(k)})^{f_{ij}^{(k)}}$$

The maximum of L_{ID} is achieved at $\hat{p}_{ij}^{(k)} = f_{ij}^{(k)} / f_{i.}^{(k)}$, where $f_{i.}^{(k)} = \sum_j f_{ij}^{(k)}$. Under the no individual differences hypothesis, on the other hand, we have $p_{ij}^{(k)} = p_{ij}$ for all k, and consequently

$$L_{NID} = \prod_{i,j} (p_{ij})^{f_{ij}},$$

where $f_{ij} = \sum_k f_{ij}^{(k)}$. The maximum of L_{NID} is attained at $\hat{p}_{ij} = f_{ij} / f_{i.}$, where $f_{i.} = \sum_j f_{ij}$. The AIC statistic can be readily evaluated for the two hypotheses, and may be used for the goodness of fit comparison. We have applied the above procedure to Townsend & Landon's (1982) data, and have found significant individual differences. Thus, Townsend & Landon's decision to analyze individual data was indeed a sensible decision. Unfortunately, nothing could be done for Keren & Baggen's data, since they are provided in an aggregated form.

The second point relates to empirical status of s_{ij}. Luce (1961) called it a similiarity parameter as a quantity which satisfies the following properties:

1) $1 = s_{ii} \geq s_{ij}$ for all i and j

2) $s_{ij} = s_{ji}$ (by definition)

3) $s_{ij} \geq s_{ik}\, s_{kj}$.

Note that these properties are analogous to the three metric axioms. Just as the distance is defined as a quantity which satisfies the metric axioms, the similarity (as a formal concept) is defined as a quantity that satisfies the above properties. Once this is accepted, it only remains to be seen that s_{ij} estimated from data indeed satisfies these properties.

A question remains, however, as to whether the above properties constitute a necessary and sufficient set of properties characteristic of the similarity. For example, s_{ij} is by definition symmetric, but is the similarity necessarily symmetric? Recent evidence indicates the contrary (Tversky, 1977; Krumhansl, 1978), although it may still be useful to retain the symmetry in the formal use of the word similarity. To quote Smith (1980; p 132):

> "Although the linguistic habits of subjects clearly lead them to use the term asymmetrically, I expect the analyst in most cases will be aided by separating such usage into symmetric and antisymmetric parts and retaining the scientific term similarity for the symmetric part."

This is exactly what is done in Luce's similarity choice model. As has been shown in the paper, the similarity choice model (and all its special cases) decomposes response strength into symmetric and antisymmetric parts, and associate the former with stimulus similarity and the latter with response bias.

The similarity should also exhibit certain invariance properties. For example, it should not be affected by such factors as stimulus presentation frequencies and pay-offs. However, it is known that not only the bias parameters but also s_{ij} change rather drastically as a function of such manipulations (Townsend & Ashby, 1982). It is also affected by the response set (the set of possible responses) and by other stimulus presentation conditions (Townsend & Landon, 1982). As the bias parameter has many faces, being affected by various factors, s_{ij} is no better than the bias parameter in this regard, and further investigations are necessary to isolate these factors. In fact, both the euclidean distance choice model and the unique feature choice model represent attempts to remedy this situation. Unfortunately, these attempts are not yet quite successful.

Additional References:

Luce, R.D. A choice theory analysis of similarity judgments. Psychometrika, 1961, 26, 151-163.

Morgan, B.J. On Luce's choice axiom. Journal of Mathematical Psychology, 1974, 11, 107-123.