MULTIDIMENSIONAL MODELS OF PERCEPTION AND COGNITION

Edited by

F. Gregory Ashby

University of California at Santa Barbara

LAWRENCE ERLBAUM ASSOCIATES, PUBLISHERS

1992 Hillsdale, New Jersey Hove and London

SCIENTIFIC PSYCHOLOGY SERIES

Stephen W. Link & James T. Townsend, Editors

EDITED VOLUMES

F. Gregory Ashby Multidimensional Models of Perception and Cognition

Hans-Georg Geissler, Stephen W. Link, and James T. Townsend
Cognition, Information Processing, and Psychophysics: Basic Issues

MONOGRAPHS

William R. Uttal et al. The Swimmer: An Integrated Computational Model of a Perceptual Motor System

Stephen W. Link • The Wave Theory of Difference and Similarity

13 Structures in Stimulus Identification Data

Yoshio Takane
Tadashi Shibayama
McGill University

INTRODUCTION

Stimulus identification data have attracted considerable attention from many researchers (e.g., Ashby & Perrin, 1988; Keren & Baggen, 1981; Nosofsky, 1985b; J. E. K. Smith, 1980, 1982; Takane & Shibayama, 1986; Townsend & Ashby, 1982; Townsend & Landon, 1982). In a stimulus identification experiment one of n stimuli is randomly selected and presented on each trial, and the subject's task is to identify the stimulus. The basic data thus consists of a set $f_{j/i}$ (i = 1, . . . , n; j = 1, . . . , n) of frequencies of response j when stimulus i is presented, with $f_i = \Sigma_j f_{j/i}$ the total number of presentations of stimulus i.

Various models have been proposed for stimulus identification data (e.g., Ashby & Perrin, 1988; Keren & Baggen, 1981; Luce, 1963a; Nakatani, 1972; Shepard, 1957; Townsend, 1971). Typically, these models attempt to predict $p_{j/i}$, the probability of response j when stimulus i is presented. The models are distinguished by different submodels assumed for $p_{j/i}$. Two major classes of models have been proposed. (We exclude, from our account, the more recently proposed general recognition model by Ashby and Perrin, 1988, since it is treated in Chaps. 6–8, and 16. Also, see Ashby and Lee, 1991.) One class is similarity-choice models, and the other is sophisticated guessing models (J. E. K. Smith, 1980; Townsend & Landon, 1982).

In the similarity-choice models, a model of stimulus similarity is postulated, and the strength of a response when a stimulus is presented is defined as a function of the stimulus similarity and the bias for that response. A response is assumed chosen with probability proportional to its response strength relative to other alternative responses. This class of models includes the unrestricted

similarity-choice model (sometimes called the biased-choice model; Luce, 1963a). The Euclidean distance-choice model [sometimes called the MDS (Multidimensional Scaling) choice model; Shepard, 1957; Nosofsky, 1985b], and the unique feature-choice model (Keren& Baggen, 1981; Tversky, 1977). These models have been systematically compared by Takane and Shibayama (1986).

In the sophisticated guessing models, a presentation of a stimulus is assumed to generate an internal state, called a confusion set, characterized by a set of admissible responses. Its probability is defined as a function of stimulus similarities/dissimilarities. A response is chosen from the admissible responses in the confusion set with probability proportional to the response bias. The confusion probability is the sum, over all possible confusion sets, of probabilities leading to a certain response when a certain stimulus is presented. This class of models includes various versions of Nakatani's (1972) confusion-choice model, the all-or-none (AON) model (Broadbent, 1967; Townsend, 1971), the overlap (OVLP) activation model (Townsend, 1971), and the informed guessing model (Pachella, Smith, & Stanovich, 1978). The latter three models are also considered special cases of the similarity-choice models (see fourth section).

There has been considerable effort to establish relationships among various models of stimulus identification data. See Marley (chap. 12), Nosofsky (in press), Takane and Shibayama (1986), Townsend and Ashby (1982), Townsend and Landon (1982, 1983), and van Santen and Bamber (1981), each presenting a somewhat different viewpoint for relating the models.

This chapter compares goodness of fit (GOF) of the models in the two classes. A general strategy for model comparison is presented first (second section). It is illustrated by three examples for assessing the stability of confusion probabilities across contexts (across trials, across subjects, and across other stimulus conditions). In the third section, the similarity-choice models are briefly reviewed. Then, two related issues will be addressed, namely, fitting ADDTREE and EXTREE to stimulus identification data, and the problem of d (exponential) versus d^2 (Gaussian) in the exponent of the Euclidean distance-choice model. In the fourth section, the sophisticated guessing models are discussed along with their relationships. Empirical results are presented in the fifth section.

A STRATEGY FOR MODEL COMPARISON

Different models of stimulus identification data postulate different submodels for $p_{j/i}$. Whichever submodels are assumed, however, the likelihood of the total set of observations may be written as

$$L = \prod_i \prod_j (p_{j/i})^{f_{j/i}}, \tag{1}$$

where $p_{j/i}$ may be further constrained in various ways. Model parameters are determined in such a way as to maximize the log of L. Note that the definition of L does not include terms that are not related to model parameters. These terms do not affect the maximum likelihood estimates.

Once the maximum likelihood is obtained, Akaike's (1974) Information Criterion (AIC) can readily be calculated and used for model comparison. This statistic is defined as

$$AIC\ (q) = -2\ ln\ L^*(q) + 2n(q)\ , \qquad (2)$$

where $L^*(q)$ is the maximum likelihood and n(q) is the effective number of parameters in model q. The first term on the right side of Equation 2, $-2\ ln\ L^*(q)$, indicates a badness of fit of model q to the current data set. A reasonably good fit to the current data set is crucial; otherwise there is no way that the model can fit to future observations well. However, a goodness of fit to the current data set can be improved, as desired, by simply increasing the number of model parameters. An improved fit obtained this way, however, may not work favorably for predicting future observations, since additional parameters to be estimated tend to produce less reliable parameter estimates. To avoid overparametrization, the AIC penalizes additional use of parameters by adding 2n(q) to $-2\ ln\ L^*(q)$. The AIC is an estimate of $-2\ E\ [ln\ L(q)]$, minus twice the expected log-likelihood. However, on average, $-2\ ln\ L^*(q)$ underestimates $E[-2\ ln\ L(q)]$ as much as 2n(q). Adding 2n(q) to $-2\ Ln\ L^*(q)$ thus corrects this bias.

A smaller value of AIC indicates a better-fitting model. In an actual model comparison process, maximum likelihoods of competing models are obtained by fitting them to the data. The AICs are then calculated according to Equation 2, and the model associated with the smallest value of AIC is chosen as the best-fitting model. This procedure is called the minimum AIC procedure. Note that only relative values of AIC (which is larger or smaller) are relevant in the comparison process. Consequently, a constant may be added to AIC values without loss of generality. In Table 13.5, for example, AIC values are adjusted, so that the AIC of the saturated model is equal to zero. This is also why the maximum likelihood used to calculate AIC need not include the terms common to all models compared.

The minimum AIC optimizes predictability. Consequently, the best model identified by the minimum AIC procedure does not imply that the model is correct. It only means the model is best among competing models in the sense that it gives predictions closest to those produced by the correct model (i.e., future observations). The minimum AIC procedure eliminates certain restrictions associated with the asymptotic chi-square goodness-of-fit test. A significance level need not be chosen arbitrarily, and more than two models can be compared simultaneously. Also, the models compared need not be hierarchically ordered in their complexity.

The philosophy underlying the minimum AIC procedure is radically different from that underlying conventional statistical significance testing procedures. However, the following example, drawn from the situation in which the chi-square GOF test is also feasible, may illuminate what the minimum AIC procedure really does in much broader contexts. Let the model q = 1 be a special case of the model q = 2. The minimum AIC procedure selects a model according to whether $AIC\ (1) - AIC\ (2) \gtrless 0$. Since $AIC\ (1) = -2\ ln\ L^*(1) + 2n(1)$ and $AIC\ (2) = -2\ ln\ L^*(2) + 2n(2)$, $AIC\ (1) - AIC\ (2) \gtrless 0$ is equivalent to $-2[ln\ L^*(1) - ln\ L^*(2)] \gtrless 2[n(2) - n(1)]$. The left side of this inequality is equal to the asymptotic chi-square statistic. The minimum AIC procedure in this instance is thus equivalent to the chi-square GOF test with 2[n(2) − n(1)] used as the critical value. The significance level of this chi-square test can be found, if desired, by working backward from 2[n(2) − n(1)]. The minimum AIC procedure thus incorporates a built-in significance level. To ask if an observed difference in two AIC values is statistically significant is like asking if the observed difference in GOF is close to the built-in significance level or significantly far from it. No such significance tests are available in the statistical literature. See Sakamoto, Ishiguro, and Kitagawa (1986) for more detailed accounts of the AIC statistic.

The following three examples demonstrate elementary uses of the minimum AIC procedure. More examples will be given in the following sections.

Example 1 (Constancy of $p_{j/i}$ Over Trials)

The product multinomial form of the likelihood function, Equation 1, presupposes independently replicated trials. This implies that $p_{j/i}$ is constant throughout the trials. Whether this is so can be verified by the following procedure. Trials are first grouped into several blocks. Let $p_{j/i(k)}$ and $f_{j/i(k)}$ denote, respectively, $p_{j/i}$ and $f_{j/i}$ for block k. The likelihood of $\{f_{j/i(k)}\}$, k = 1, . . . , K (where K is the number of blocks) is then stated as

$$L = \prod_k \prod_i \prod_j (p_{j/i(k)})^{f_{j/i(k)}} . \tag{3}$$

Under the hypothesis that $p_{j/i(k)} = p_{j/i}$ for all k, Equation 3 reduces to Equation 1. The problem thus becomes one of comparing the goodness-of-fit of Model 1 and Model 3. The maximum likelihood estimate (MLE) of $p_{j/i}$ in Equation 1 is given by $f_{j/i}/f_j$. This model uses n(n − 1) independent parameters. The MLE of $p_{j/i(k)}$ in Equation 3, on the other hand, is given by $f_{j/i(k)}/f_{j(k)}$, where $f_{j(k)} = \Sigma f_{j/i(k)}$. Model 3 uses Kn(n − 1) independent parameters.

Nosofsky (1987) collected "learning" identification data on 12 Munsell colors. The stimuli had constant hue (5R), but varied in brightness and saturation. The data were collected while subjects were still improving their performance. An experimental session was organized into three blocks of 108 trials each, and 34 subjects participated in the experiment. Model 1 and Model 3 were fitted to

Nosofsky's data. The AIC value was found to be 28,394.7 [(n(q) = 396] for Model 3 and 29,678.0 [(n(q) = 132] for Model 1, indicating that Model 3 was the better fitting model. As expected, confusion probabilities are not constant across the blocks. To justify Model 1, we must collect identification data after $p_{j/i}$'s have reached their "asymptotes" by a sufficient number of practice trials. Note that the result by no means implies that Model 3 is correct. It is possible that $p_{j/i(k)}$ may still vary within the blocks.

Example 2 (Constancy of $p_{j/i}$ Across Subjects)

Confusion data are often aggregated across subjects. This, however, assumes that $p_{j/i}$'s are constant across the subjects. Whether the $p_{j/i}$'s are constant can be tested similar to Example 1, provided that individual data are also available. "Blocks" in Example 1 are simply replaced by "subjects," and the problem reduces to a comparison between Model 3 and Model 1.

Townsend and Landon (1982) suspected significant individual differences in performing the stimulus identification task, and consequently analyzed each of four subjects' data separately. One of their original purposes was to test the Constant-Ratio Rule (see Example 3), and four different subsets of five letters (**A**, **E**, **F**, **H**, **X**) were employed. Set 1 included all letters, set 2 only (**A**, **E**, **F**, **H**), set 3 (**A**, **E**, **X**), and set 4 (**F**, **H**, **X**). Each stimulus was presented 240 times for each set and for each subject. For the purpose of testing the individual differences, only set 1 and set 2 were used here. Set 1 yielded the AIC value of 12,141.7 [n(q) = 80] for Model 3 and 12,230.4 [n(q) = 20] for Model 1. Set 2 yielded the AIC value of 8,953.9 [n(q) = 48] for Model 3 and 8,975.0 [n(q) = 12] for Model 1. For both sets, Model 3 was found to be the better-fitting model, indicating that the individual differences were indeed substantial. Townsend and Landon (1982) made the right decision in analyzing individual data.

Examples 1 and 2 concerned the constancy of $p_{j/i}$. When the constancy assumption is grossly violated, the use of Model 1 can be problematic. The problem is known as "overdispersion" (e.g., McCullagh & Nelder, 1983), in which the variance of $f_{j/i}$ becomes much larger than what is expected from the multinomial distributional assumption due to the variability in $p_{j/i}$ (Kraemer, 1988). The quasi-likelihood proposed by McCullagh and Nelder (1983) can incorporate an additional dispersion term to deal with the overdispersion problem often encounted in contingency table analyses.

Example 3 (Constant-Ratio Rule)

Let M and S denote sets of stimuli such that $S \subseteq M$ (i.e., S is a subset of M). Let $p_{j/i}(M)$ and $p_{j/i}(S)$ denote $p_{j/i}$ when the stimulus sets in identification tasks are restricted to M and S, respectively. The Constant-Ratio Rule (CRR; Clarke, 1957) stipulates that $p_{j/i}(S)$ is proportional to $p_{j/i}(M)$ for $i, j \in S$. That is,

$$p_{j/i}(S) = cp_{j/i}(M) , \tag{4}$$

for some $c \neq 0$. However, $\Sigma_{j\in S}p_{j/i}(S) = 1$, so that $c\,\Sigma_{j\in S}p_{j/i}(M) = 1$, or $c = 1/\Sigma_{j\in S}p_{j/i}(M)$. Using this expression for c, we can write Equation 4 more explicitly as

$$p_{j/i}(S) = p_{j/i}(M) \Big/ \sum_{k\in S} p_{k/i}(M) . \tag{5}$$

The right side of Equation 5 is equal to the conditional probability of $p_{j/i}(M)$ given S, which is denoted by $p_{j/i}(M/S)$. The CRR states that $p_{j/i}$ is essentially context-free, and the only effect of reducing the stimulus set M to S is that the probability of inadmissible responses under S is redistributed over admissible responses under S in proportion to $p_{j/i}(M)$ for $i, j \in S$. An alternative way of stating the same property is that $p_{k/i}/p_{j/i}$ does not depend on the stimulus set. That is,

$$\frac{p_{k/i}(M)}{p_{j/i}(M)} = \frac{p_{k/i}(S)}{p_{j/i}(S)} \qquad \text{for } i, j, k \in S . \tag{6}$$

Many researchers have investigated the CRR (Clarke, 1957; Hodge & Pollack, 1962; Morgan, 1974; Townsend & Landon, 1982). The latter two used the Likelihood Ratio Test. The CRR can also be tested by using the minimum AIC procedure. Let $f_{j/i}(M)$ and $f_{j/i}(S)$ be observed confusion frequencies, when the stimulus sets are M and S, respectively. Under the CRR, the ML estimate of $p_{j/i} = p_{j/i}(S) = p_{j/i}(M/S)$ is given by $\hat{p}_{j/i} = [f_{j/i}(S) + f_{j/i}(M)]/[f_j(S) + f_j(M/S)]$, where $f_i(S) = \Sigma_{j\in S}f_{j/i}(S)$ and $f_i(M/S) = \Sigma_{j\in S}f_{j/i}(M)$. This value of $\hat{p}_{j/i}$ is used in Model 1 to obtain the maximum likelihood under the CRR hypothesis. The CRR uses s(s − 1) parameters, where s is the number of stimuli in S. Under the non-CRR hypothesis the ML estimates of $p_{j/i(1)} \equiv p_{j/i}(S)$ and $p_{j/i(2)} \equiv p_{j/i}(M/S)$ are given by

$$\hat{p}_{j/i(1)} = \frac{f_{j/i}}{f_i(S)} \qquad \text{and} \qquad \hat{p}_{j/i(2)} = \frac{f_{j/i}(M)}{f_i(M/S)} ,$$

which are substituted for $p_{j/i(k)}$, k = 1,2, in Model 3 to obtain the maximum likelihood under this hypothesis. Model 3 in this case uses twice as many parameters as Model 1. The AICs are calculated in the same way as before, and the model comparison proceeds just as in the previous examples.

The foregoing procedure will be demonstrated with Townsend and Landon's (1982) data. This data set was briefly described in Example 2. There were four subjects (D.X., M.X., G.X., and A.X.), and each subject's data were analyzed separately. This is in accordance with the results of Example 2. Four stimulus sets were employed with set 1, with stimuli (**A, E, F, H, X**) serving as the master set for all the other three sets, set 2 with (**A, E, F, H**), set 3 with (**A, E, X**), and set 4 with (**F, H, X**). Results are reported in Table 13.1. For each pair of stimulus sets, the minimum AIC solution is indicated by an "a."

TABLE 13.1
Tests of Constant-Ratio Rule with Townsend and Landon's (1982) Data

Subject	Data Sets M	Data Sets S	Stimuli in M and Those in S (underlined)	CRR	Non-CRR	Difference in n(q)
D.X.	(1),	(2)	A E F H X	4,202.2[a]	4,207.4	12
	(1),	(3)	A E F H X	1,696.9[a]	1,698.5	6
	(1),	(4)	A E F H X	1,679.3	1,674.6[a]	6
M.X.	(1),	(2)	A E F H X	4,439.5[a]	4,454.8	12
	(1),	(3)	A E F H X	2,125.8[a]	2,131.7	6
	(1),	(4)	A E F H X	1,960.1	1,954.0[a]	6
G.X.	(1),	(2)	A E F H X	4,056.2[a]	4,070.4	12
	(1),	(3)	A E F H X	1,812.4[a]	1,817.4	6
	(1),	(4)	A E F H X	1,975.1[a]	1,975.5	6
A.X.	(1),	(2)	A E F H X	4,175.7[a]	4,192.0	12
	(1),	(3)	A E F H X	1,839.4	1,837.0[a]	6
	(1),	(4)	A E F H X	1,921.8	1,918.0[a]	6

[a]Minimum AIC.

Neither the CRR nor the non-CRR hypothesis is uniformly better than the other. However, there are twice as many cases supporting the CRR as cases against the CRR. This is probably because M and S are fairly similar in Townsend and Landon's data. Intuitively, the more similar M and S are, the higher is the chance that the CRR holds. The similarity between M and S may be measured by s/m, where s and m are the numbers of stimuli in S and M, respectively.

The pattern of cases in favor of the CRR across subjects and different pairs of stimulus sets agrees perfectly with the results obtained by Townsend and Landon (1982), using the Likelihood Ratio Chi-Square Test. For set 1 versus set 2, all the four subjects favored the CRR. This case had the highest similarity between M and S. For set 1 versus set 3, three favored the CRR, but for set 1 versus set 4 only one subject favored the CRR. These two cases shared three stimuli each, and consequently the similarity between M and S is considered approximately equal.

An error analysis was conducted to identify $p_{j/i}$'s for which the discrepancy between $\hat{p}_{j/i}$ under the CRR and $\hat{p}_{j/i}(S)$ and $\hat{p}_{j/i}(M/S)$ under the non-CRR was large. The discrepancy is measured by

$$z_{ij} = \pm \left\{ 2f_{j/i}(S)[\ln \hat{p}_{j/i}(S) - \ln \hat{p}_{j/i}] + 2f_{j/i}(M)\left[\ln \hat{p}_{j/i}\left(\frac{M}{S}\right) - \ln \hat{p}_{j/i} \right] \right\}^{1/2}$$

(Pierce and Schafer, 1986), which approximately follows the standard normal distribution. Note, however, that the z_{ij}'s are not independent across j. Stimulus-response pairs for which z_{ij} exceeds ± 2.58 [$\Pr(z \geq |2.58|) = 0.01$] are listed in Table 13.2. Stimuli **F** and **X** seem to be causing most of the problem.

TABLE 13.2
Stimulus Pairs That Violate CRR in Townsend and Landon's (1982) Data

Subject	Data Sets		Pair		p in the Master Set (M)	p in the Reduced Set (S)
	M	S	Stimulus	Response		
D.X.	(1),	(4)	F	X	.27	.13
M.X.	(1),	(4)	X	F	.06	.14
A.X.	(1),	(3)	X	A	.06	.13
A.X.	(1),	(4)	H	F	.07	.14

Hodge and Pollack's (1962) data (see fifth section) were also analyzed in a similar manner. Results were similar. The CRR does not seem to hold universally, and how likely it holds depends on the similarity between the stimulus sets.

SIMILARITY-CHOICE MODELS

All the models considered in the previous section impose some form of equality restrictions on $p_{j/i}$. This section deals with a group of models, called similarity-choice models, that specify explicit submodels under $p_{j/i}$. In this section, the similarity-choice models are briefly reviewed, and then two related issues are addressed, namely, fitting ADDTREE and EXTREE and comparing d and d^2 in the Euclidean distance-choice model.

Brief Review

In the similarity-choice models, $p_{j/i}$ is assumed proportional to the strength of response j when stimulus i is presented. Denote the response strength by t_{ij}. Then $p_{j/i} = v_i t_{ij}$ for some v_i. But since $\Sigma_j p_{j/i} = v_i \Sigma_j t_{ij} = 1$, $v_i = 1/\Sigma_j t_{ij}$, and thus

$$p_{j/i} = \frac{t_{ij}}{\sum_k t_{ik}} . \tag{7}$$

A variety of similarity-choice models are obtained by specializing t_{ij} in various ways.

In Luce's (1963a) unrestricted similarity-choice model, it is assumed that

$$t_{ij} = w_j s_{ij} , \tag{8}$$

where w_j (≥ 0, $\Sigma_j w_j = 1$) is the bias for response j, and s_{ij} is the similarity between stimuli i and j ($0 \leq s_{ij} = s_{ji} \leq s_{ii} = s_{jj} = 1$). By substituting Equation 8 in Equation 7, we can write the model more explicitly as

$$p_{j/i} = \frac{w_j s_{ij}}{\sum_k w_k s_{ik}} . \tag{9}$$

The model is called the unrestricted similarity-choice model, since there is no further restriction imposed on s_{ij}. The model is sometimes called the biased-choice model.

The unrestricted similarity-choice model is known to be a special case of the quasi-symmetry model for square contingency tables, which states

$$p_{ij} = ca_i b_j g_{ij} ,$$

where p_{ij} is the joint probability of row i and column j, c is some constant, a_i ($\Pi_i a_i = 1$) is the effect of row i, b_j ($\Pi_j b_j = 1$) is the effect of column j, and g_{ij} ($g_{ij} = g_{ji}$; $\Pi_i g_{ij} = \Pi_j g_{ij} = 1$) is the interaction effect between row i and column j. Since $p_{j/i} = p_{ij}/p_i$, where $p_i = \Sigma_j p_{ij}$, $p_{j/i}$ is given by

$$p_{j/i} = \frac{b_j g_{ij}}{\Sigma_k b_k g_{ik}} . \tag{10}$$

Model 9 is derived from Equation 10 by setting $s_{ij} = g_{ij}/(g_{ii}g_{jj})^{1/2}$ and $w_j = b_j g_{jj}{}^{1/2}/\Sigma_k b_k g_{kk}{}^{1/2}$. One important property of the quasi-symmetry model is the cycle condition (Caussinus, 1965)

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

which implies

$$p_{j/i}p_{k/j}p_{i/k} = p_{k/i}p_{j/k}p_{i/j} ,$$

for the unrestricted similarity-choice model (J. E. K. Smith, 1982). In the unrestricted similarity-choice model, $0 \leq s_{ij} = s_{ji} \leq s_{ii} = s_{jj} = 1$ also implies the column constraint

$$p_{i/i} \geq p_{i/j}$$

for all j. It has been shown (Townsend & Landon, 1982) that these two conditions are necessary and sufficient for the unrestricted similarity-choice model.

It follows from Equation 9 that

$$ln\left(\frac{p_{j/i}}{p_{i/i}}\right) = (ln\ w_j - ln\ w_i) + ln\ s_{ij} . \tag{11}$$

This implies that Model 9 decomposes $ln(p_{j/i}/p_{i/i})$ into two parts, a skew-symmetric part and a symmetric part, and represents the former by $ln\ w_j - ln\ w_i$ and the latter by $ln\ s_{ij}$.

In the Euclidean distance-choice model, stimuli are represented as points in multidimensional space. The distance between stimuli i and j, d_{ij}, is assumed related to s_{ij} by $s_{ij} = exp\ (-d_{ij})$ or $s_{ij} = exp(-d_{ij}^2)$. This leads to

$$p_{j/i} = \frac{w_j\ exp\ (-d_{ij})}{\Sigma_k w_k\ exp\ (-d_{ik})} , \tag{12}$$

or

$$p_{j/i} = \frac{w_j \exp(-d_{ij}^2)}{\sum_k w_k \exp(-d_{ij}^2)} . \quad (13)$$

Model 12 is sometimes called exponential MDS-choice model, Model 13 the Gaussian MDS-choice model (see also chap. 14; MDS stands for multidimensional scaling). There has been controversy over which of the two MDS-choice models, 12 or 13, better accounts for stimulus identification data (Nosofsky, 1985a, 1985b, 1986; Shepard, 1986, 1988). The issue will be addressed in the third section (see also chap. 11).

The MDS-choice model can also be derived from Krumhansl's (1978) distance-density model. Let $r_{ij} = d_{ij} + a_i^* + b_j^*$ or $r_{ij} = d_{ij}^2 + a_i^* + b_j^*$ be the distance-density model, where a_i^* and b_j^* are, respectively, the stimulus and response density parameters. Let $t_{ij} = \exp(-r_{ij})$ in Equation 7. Model 12 or 13 is obtained, depending on which r_{ij} is used to define t_{ij}. In either case, $w_j = \exp(-b_j^*)$.

In the unique feature-choice model (Keren & Baggen, 1981; Tversky, 1977) t_{ij} is specialized in yet another way. Let

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

and define $x_{ija} = y_{ia}(1 - y_{ja})$, and $x_{jia} = y_{ja}(1 - y_{ia})$. The x_{ija} takes the value 1 if stimulus i, but not stimulus j, has feature a, and is zero otherwise. The x_{jia}, on the other hand, takes the value 1 if feature a is unique to stimulus j. Let h_{ij} be the (asymmetric) dissimilarity between stimuli i and j, defined by

$$h_{ij} = \sum_a (x_{ija} b_a + x_{jia} c_a) , \quad (14)$$

where b_a and c_a are the dissimilarity contributions of feature a, when the feature is unique to stimulus i ($x_{ija} = 1$) and stimulus j ($x_{jia} = 1$), respectively. Let $t_{ij} = \exp(-h_{ij})$ in Equation 7. Then

$$p_{j/i} = \frac{w_j^* \exp(-e_{ij})}{\sum_k w_k^* \exp(-e_{ik})} , \quad (15)$$

where

$$w_j^* = \exp\left(\sum_a u_a^* y_{ja}\right) ,$$

and

$$e_{ij} = \Sigma v_a^* |y_{ia} - y_{ja}|^q , \quad (16)$$

with $q \geq 1$, $u_a^* = (b_a - c_a)/2$ and $v_a^* = (b_a + c_a)/2$. This indicates that the unique feature model is a special case of Model 9, in which both w_j and s_{ij} are constrained in special ways (J. E. K. Smith, 1982); that is, $w_j = w_j^*$ and $s_{ij} = \exp$

$(-e_{ij})$. The e_{ij} in Equation 16 can be considered the q-th power of the Minkowski power metric. In the special case $q = 2$, e_{ij} is equal to the square of the Euclidean distance. Thus, the unique feature-choice model can also be viewed as a special case of Equation 13, where $w_j = w_j^*$ and $d_{ij}^2 = e_{ij}$ defined by a set of prescribed features.

It may be assumed that c_a is proportional to b_a; that is, $c_a = cb_a$ for some c but for all a. Then Equation 14 reduces to

$$h_{ij} = \sum_a (x_{ija} + x_{jia}c)b_a \,. \tag{17}$$

This model is called the restricted unique feature model, as opposed to Equation 14, which is called the general unique feature model.

When we further assume that $b_a = c_a$, Equation 14 reduces to the symmetric-difference model

$$q_{ij} = \sum_a (x_{ija} + x_{jia})b_a \,, \tag{18}$$

which may be substituted for d_{ij}^2 in Equation 13 to obtain

$$p_{j/i} = \frac{w_j \, exp\,(-q_{ij})}{\sum_k w_k \, exp\,(-q_{ik})} \,. \tag{19}$$

This model plays an important role in fitting ADDTREE and EXTREE to stimulus identification data.

Fitting ADDTREE and EXTREE

In the additive similarity tree (ADDTREE; Sattath & Tversky, 1977; see also chap. 3), dissimilarity between two stimuli is represented by the length of a path connecting them. The extended similarity tree (EXTREE; Corter & Tversky, 1986) is similar, except that some segments of paths have markers. Whenever a path connecting two stimuli includes segments with common markers, those segments are excluded from the total path length. That is, they are not counted toward the overall dissimilarity between the stimuli. There is one-to-one correspondence between ADDTREE and EXTREE representations and the symmetric difference model, Equation 18, defined on a set of features determined by the tree structure. Given the tree structure, then, ADDTREE and EXTREE can be fitted to the stimulus identification data by using Equation 18.

Let us illustrate, using Keren and Baggen's (1981) data. This data set pertains to 10 rectangular digits used in digital clocks and calculators. There were eight subjects, but stimulus exposure duration was adjusted for each subject to minimize individual differences in performance, and the data were pooled across the subjects. Keren and Baggen's data have been analyzed and reanalyzed previously by several authors (Keren & Baggen, 1981; J. E. K. Smith, 1982; Takane &

Shibayama, 1986). Takane and Shibayama, in particular, used their data to systematically compare various similarity-choice models.

Figure 13.1 displays an EXTREE structure derived from Keren and Baggen's data by Corter and Tversky (1986). A feature set can be extracted from the tree that defines the symmetric-difference metric. In the tree a feature corresponds with a branch. In the figure it corresponds with a segment of a path between a terminal node (representing a stimulus) and the root that connects all the stimuli in the stimulus set. There are 19 such segments numbered from 1 to 19. Four of them are marked by special symbols (**C**, **D**, **E**, & **H**). Segments having direct contact with terminal nodes represent features unique to the stimuli corresponding to the nodes. Digit 2, for example, has only one feature (feature 1) unique to the stimulus. Digit 1 has features 2, 4, 5, and 19. Feature sets that characterize other stimuli can be obtained in a similar manner. Table 13.3 displays the feature indicator matrix for the 10 digits corresponding to the EXTREE structure presented in Figure 1.

Dissimilarity between digits 1 and 2 is obtained by summing the contributions

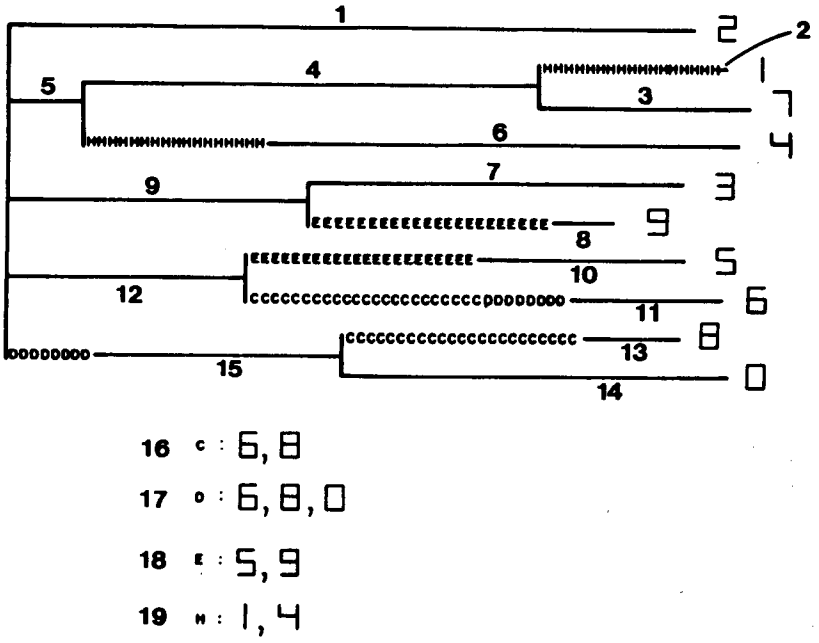

FIG. 13.1. Optimal EXTREE structure for Keren & Baggen's (1981) data. *Note.* From "Extended Similarity Trees" by J. E. Corter and A. Tversky, 1986, *Psychometrika, 51,* p. 443. Copyright 1986 by the Psychometric Society. Reprinted by permission.

TABLE 13.3
Feature Matrix Corresponding to the EXTREE Structure Given in Figure 13.1 for Keren & Baggen's (1981) Data

	Features																	
Stimuli	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*17*	*18*	*19*
1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
5	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0
6	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1	0	0
7	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	1	0	0
9	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0

of features 1, 2, 4, 5, and 19. Marked feature 19 is included because it is unique to digit 1. Dissimilarity between digits 1 and 4 is defined by features 2, 4, and 6. Feature 19 is not included because it is commonly shared by the two digits. Our analysis obtains optimal weights for the features, which, when added, give overall dissimilarities between the stimuli.

The ADDTREE structure (not displayed) obtained by Corter and Tversky (1986) is similar to the EXTREE structure. It is identical to the EXTREE structure without marked features except that digit groups (8, 0), (5, 6), (3, 9), (1, 7, 4), and (2) are not joined simultaneously. In ADDTREE, (8, 0) and (5, 6) are joined first, then (8, 0, 5, 6) with (3, 9), and then (8, 0, 5, 6, 3, 9), (1, 7, 4) and (2) simultaneously. The ADDTREE structure contains 17 features. The feature indicator matrix can be constructed in the same way as in Table 13.3.

Both the ADDTREE and the EXTREE structures were fitted to Keren and Baggen's data. ADDTREE yielded the AIC value of 164.1 with 26 parameters. (There are nine independent bias parameters.) EXTREE provided the AIC value of 56.1 with 28 parameters. With just two more parameters this improvement in fit is rather impressive. EXTREE also does considerably better than the Euclidean distance-choice models with comparable numbers of parameters. (See Table 13.8.) However, EXTREE does not fit as well as the best fitting model found by Takane and Shibayama (1986), that is, the unrestricted similarity-choice model.

Apart from the fact that it did not provide the best-fitting model, there is an obvious limitation to this analysis. It presupposes that the optimal tree structure is known beforehand. The "optimal" tree structures used were obtained by Corter and Tversky (1986), but there is no guarantee that they are optimal with respect to the model and the fitting criterion used here. Indeed, the weight for feature 8 in EXTREE was estimated to be negative, and in order to avoid the improper

solution the feature had to be eliminated. This indicates that the true optimal structure could be different from the one fitted.

Table 13.4 shows the stimulus-response pairs for which the discrepancy between observed and predicted probabilities is large. For the pairs of stimuli for which the predicted probability is underestimated, new markers could be added to part of the features unique to the stimuli, increasing the similarity between them. Overestimation of confusion probabilities is more difficult to deal with. There is one such pair in the table, stimulus 2 and response 9. One way to handle this case is obviously to increase the weight for feature 2 while attaching a common marker to this feature as well as features unique to all other digits except digit 9. This is admittedly ad hoc, however; the meaning of the marked feature is not entirely clear.

Comparison Between d and d^2

As noted, two versions of the Euclidean distance-choice model have been proposed, d and d^2 in the exponent. There has been some controversy as to which of the two works better. This section examines existing empirical evidence and reports results on a numerical experiment designed to shed some light on the issue.

Kornbrot (1978) compared the GOF of d and a logistic version of Thurstone's successive categories model (SCM) for unidimensional stimuli (pure tones varying in intensity) and found that the latter fitted the data considerably better. Nosofsky (1985a) reanalyzed Kornbrot's data by d^2 and found that it provided comparable fits to SCM. Nosofsky (1985b, 1986) also presented data sets involving multidimensional stimuli (semicircles of varying size with a spoke in each semicircle oriented in a different angle from a horizontal baseline) to which d^2 fitted appreciably better than d. Ashby and Perrin (1988) report analyses of Townsend, Hu, and Ashby's (1981) data, for which they found d^2 worked consistently better.

Table 13.5 summarizes these results. In the table the AIC values have been

TABLE 13.4
Stimulus Pairs with Large Discrepancies Between Observed and Predicted Probabilities from EXTREE

Stimulus	*Response*	*Observed p*	*Predicted p*
2	6	.047	.013
1	6	.015	.002
0	7	.025	.007
2	9	.007	.046
3	7	.040	.019
9	8	.082	.046

TABLE 13.5
Comparison Between d and d^2 for Various Data Sets

Source		*Data*	d	d^2
Kornbrot (1978)	1	D.P. (natural cond.)	33.4	−44.5[a]
	2	D.J. (natural cond.)	60.3	−19.1[a]
dim = 2	3	D.P. (biased cond.)	350.9	−23.4[a]
	4	D.J. (biased cond.)	296.3	−47.0[a]
Nosofsky (1985b)	1	Subject 1	216.7	−233.4[a]
dim = 2	2	Subject 2	212.1	−162.4[a]
(1986)	1	Subject 1	1,262.2	−71.2[a]
dim = 2	2	Subject 2	1,015.0	7.9[a]
(1987)	1	Block 1	21.7[a]	792.1
dim = 2	2	Block 2	−54.0[a]	472.2
	3	Block 3	−48.1[a]	379.2
Townsend & Landon (1982)	1	D.X. (5 stimuli)	−14.9[a]	48.0
		(4 stimuli)	5.4[a]	21.8
dim = 2	2	M.X. (5 stimuli)	−11.3[a]	17.8
		(4 stimuli)	−4.1[a]	11.8
	3	G.X. (5 stimuli)	15.5[a]	141.7
		(4 stimuli)	−1.6[a]	40.4
	4	A.X. (5 stimuli)	−9.4[a]	44.8
		(4 stimuli)	1.6[a]	29.6
Townsend, Hu, & Ashby, (1981)[b]	1	Observer 1 (gap)	21.5	5.9[a]
	2	Observer 2 (gap)	19.6	4.3[a]
	3	Observer 3 (gap)	10.5	−2.0[a]
dim = 2	4	Observer 4 (gap)	−0.1	−3.2[a]
	5	Observer 1 (no gap)	11.5	−0.9[a]
	6	Observer 2 (no gap)	1.5	−5.0[a]
	7	Observer 3 (no gap)	4.3	−5.2[a]
	8	Observer 4 (no gap)	13.1	5.9[a]

[a]A better solution.
[b]Results obtained by Ashby & Perrin (1988).

adjusted so that the AIC of the saturated model takes the value of zero in each case. Table 13.5 is supplemented by the results from two more studies that, unlike those mentioned, favor d. Nosofsky (1987) used colors of constant hue (5R in Munsell notation), but varying in brightness and saturation, and Townsend and Landon (1982) used five letters of the alphabet and their subsets (see Examples 2 and 3). The results reported in Table 13.5 are very orderly in the sense that one model is consistently better than the other within each study. Data sets collected within a study use the same stimuli and the same experimental procedure.

Shepard (1986, 1988) attributed Nosofsky's (1985b, 1986) results to peculiarities in his data collection procedures and argued that d must be favored on empirical and theoretical grounds. Indeed, almost all the data sets analyzed prior to Nosofsky favored d. Data sets to be discussed in the fifth section present four

such examples. They are Keren and Baggen's (1981) data on rectangular digits (used previously), Wickelgren's (1965) data on different forms of verbs, Hodge and Pollack's (1962) data on tones varying in intensity, frequency, and duration (briefly mentioned in Example 3), and Clark and Stafford's (1969) data on consonant-vowel combinations. All of these data favor d in the exponent of the Euclidean distance-choice model. Tables 13.8 and 13.9, in particular, present systematic comparisons of various models fitted to Keren and Baggen's data and Hodge and Pollack's data, respectively.

Table 13.6 presents a summary of existing empirical results concerning the choice between d and d^2. The table indicates for each study mentioned whether d or d^2 is favored, whether the data are individual or aggregated, whether the stimuli have separable or integral dimensions, and whether practice trials were extensive, moderate, or minimal.

The last three variables are thought to be important in distinguishing the two cases. Individual data tend to favor d^2, although there is one exception. Townsend and Landon's (1982) data are individual data, yet favor d. Stimuli with separable dimensions tend to favor d^2, although in some cases deciding whether relevant stimulus dimensions are separable or integral is not so straightforward. For some stimulus sets, it is difficult even to see any obvious dimensional structures. The letters used by Townsend and Landon, the rectangular digits used by Keren and Baggen, the different verb forms used by Wickelgren, and the consonant-vowel combinations used by Clark and Stafford are such examples. They are tentatively classified as having "integral" dimensions, meaning that there are no obvious separable dimensions or that the task involved may require something more cognitive than perceptual (e.g., integration of unidentifiable dimensions).

A larger number of practice trials seem to favor d^2. Nosofsky's (1987) data, unlike his two previous studies, favor d. The data were obtained with minimal practice trials. The reason, however, could be that the data are aggregated or that the stimuli have integral dimensions. Because of the generally high correlations among the variables considered, it is impossible to tear apart their confounding effects and draw any sensible conclusion as to which variable is causing a particular effect. To isolate the effects of the variables, one must conduct a factorial experiment in which all possible combinations of levels of the variables are equally represented.

One thing seems clear, however. Whether s_{ij} is a strictly convex (downward) function of d or whether there is an inflection point near $d = 0$ is not likely to be an important factor in deciding which of the two models works better in particular situations. What counts most is what happens where d is large. Whereas $exp\ (-d^2)$ approaches 0 very quickly as d increases, $exp\ (-d_{ij})$ does not. Although part of the difference is mitigated by adjusting the overall size of the stimulus configuration, it still remains that $exp\ (-d_{ij})$ has a heavier tail. The unsquared distance d is thus favored in situations where the heavy tail is required. Aggre-

TABLE 13.6
A Summary of Empirical Results Concerning the Choice Between d and d^2

Data Set	d/d^2	*Data (Individual/ Aggregated)*	*Stimuli (Separable/ Integral)*	*Practice (Extensive/ Moderate/ Minimal)*	*Description of Stimuli*
Kornbrot (1978)	d^2	Ind	Sep (unidimensional)	Ext	1-s bursts of 500-Hz pure tones varying in intensity
Nosofsky (1985b)	d^2	Ind	Sep	Ext	Circles of varying size with a spoke in each circle oriented in a different angle
(1986)	d^2	Ind	Sep	Ext	
(1987)	d	Agg	Int	Min	Colors of constant hue, but varying in brightness and saturation
Townsend & Landon (1982)	d	Ind	Int(?)	Mod	Five letters of alphabet (A, E, F, H, X) their subset
Townsend, Hu, & Ashby (1981)	d^2	Ind	Sep	Mod	A vertical and a horizontal line segment presented together, alone or not presented
Keren & Baggen (1981)	d	Agg	Int(?)	Mod	Rectangular digits
Wickelgren (1965)	d	Agg	Int	?	Forms of verbs
Hodge & Pollack (1962)	d	Agg	Int(?)	Ext	Tones varying in intensity, frequency, & duration
Clark & Stafford (1965)	d	Agg	Int	?	Vowel-consonant and consonant-vowel combinations

gated data tend to require heavier tails because of the individual differences in identification performance. Minimal practice trials also tend to require heavier tails, because at initial stages subjects can confuse rather distinct stimuli (i.e., they make the sort of errors that they should not make, if only well-practiced).

The following numerical experiment on Nosofsky's (1987) data clarifies our argument. The data were collected while subjects were still in learning phases. Significant improvements in identification performance can be observed over blocks of trials, as verified in Example 2. The $\min(f_{ij}, c)$ (where c is some integer) was subtracted from each of the off-diagonal elements of the confusion frequency tables and added back to the corresponding diagonal elements. This is supposed to simulate what happens when subjects make fewer and fewer careless mistakes as they become more proficient in the task. Both d and d^2 were fitted to the "corrected" tables with the value of c incremented systematically.

The results are shown in Table 13.7. In all three blocks, d fitted better for smaller values of c, but the difference between d and d^2 diminished as the value of c increased until d^2 took over.

TABLE 13.7
Numerical Experiments on Nosofsky's (1987) Data

	Error Frequency Corrected $-c$	d	d^2
Block 1	0	13,781.7[a]	14,452.1
	−12	6,072.5[a]	6,236.3
	−15	4,925.7[a]	4,940.6
	−18	3,990.6	3,939.1[a]
	−19	3,722.9	3,672.9[a]
	−20	3,466.3	3,426.2[a]
	−21	3,208.2	3,176.1[a]
Block 2	0	8,409.4[a]	8,935.6
	−3	6,544.4[a]	6,582.0
	−4	6,141.2[a]	6,141.8
	−5	5,789.6	5,774.9[a]
	−6	5,619.5	5,429.6[a]
	−7	5,317.4	5,139.1[a]
	−8	5,049.8	4,831.3[a]
	−9	4,756.9	4,525.4[a]
Block 3	0	6,123.2[a]	6,550.5
	−3	4,752.8[a]	5,094.8
	−4	4,379.0[a]	4,652.3
	−5	4,020.4[a]	4,216.6
	−6	3,555.6	3,517.8[a]
	−7	3,276.3	3,234.7[a]
	−8	3,098.1	2,984.8[a]
	−9	2,899.3	2,804.1[a]

[a]Minimum AIC solution.

A larger value of c was needed to get this reversal in block 1, which included the most initial trials. The results are as expected. If, however, the variabilities in p over subjects and over trials are indeed what make the data more in line with d, then there is not much substance in this model, because those are the situations that should be avoided in collecting stimulus identification data according to the results of the second Section.

Our finding is consistent with Ennis (1988a; see also Ennis, Palen, & Mullen, 1988; chap. 11), who attempted to reconcile the two models by postulating multivariate distributions for the stimulus representations. For confusable stimuli for which "perceived similarity may vary from moment to mement because of variation in the mental representations of the stimulus objects" (Ennis, 1988a, p. 408), d^2 provides the better model. However, for discriminable stimuli that "cannot be confused because of this variation" (Ennis, 1988a, p. 408), d could be the better model. Shepard (1988), on the other hand, argues that for discriminable stimuli that "are nevertheless close enough in psychological space to be judged . . . likely to have the same important consequence" (Shepard, 1988, p. 416), d should be favored. Clearly, what Shepard has in mind is the stimulus generalization context. However, how relevant is his argument to the stimulus identification context? Stimulus identification data are usually collected under a pressure to minimize the error probability and consequently, rather distinct, from stimulus generalization data. None of the data sets for which we found d fitted better are, strictly speaking, stimulus generalization data.

SOPHISTICATED GUESSING MODELS

In this section, sophisticated guessing models (SGM) are discussed in some detail. The general form of the SGM is presented first, followed by its specialization, Nakatani's (1972) confusion-choice model. The symmetric SGM, another specialization of the general SGM, is then discussed along with its relation to the similarity-choice models, the AON and OVLP models. Some introductory remarks about the models were given in the introduction.

In the SGM, a presentation of a stimulus is assumed to generate a confusion set c, which is a set of admissible responses. A response is chosen among the admissible responses according to the bias of the response. Let C_j denote the set of all confusion sets that include j as an admissible response plus the null set, which includes no admissible responses. Let $p_{c/i}$ be the probability of confusion set c when stimulus i is presented. Then the general SGM can be written as

$$p_{j/i} = \sum_{c \in C_j} p_{c/i} \left(\frac{b_j}{\sum_{k \in c} b_k} \right), \tag{20}$$

where b_j is the bias parameter for response j (J. E. K. Smith, 1980). This b_j is

analogous to w_j in the similarity-choice models. One exception should be allowed in Equation 20. When c is null, $\Sigma_{k \in c} b_k$ should be interpreted as the sum of b_k over all possible (not necessarily admissible) responses. Most SGM do not allow the null confusion set, but some, such as Nakatani's confusion-choice model, do.

Nakatani's Confusion-Choice Model

Various specializations of Equation 20 are possible. In Nakatani's (1972) confusion-choice model, it is assumed that

$$p_{c/i} = \prod_k (p_{ik})^{r_{k/c}} (1 - p_{ik})^{1 - r_{k/c}} , \tag{21}$$

where p_{ik} is the marginal probability of response k being admissible for stimulus i, and $r_{k/c}$ is defined as

$$r_{k/c} = \begin{cases} 1 & \text{if } k \in c, \\ 0 & \text{otherwise} . \end{cases}$$

In Nakatani's original model, stimuli are represented in multidimensional Euclidean space. Each p_{ik} is assumed to be a decreasing function of the distance d_{ik} between stimulus i and response k (which is supposed to coincide with stimulus k). More specifically, d_{ik} is assumed to follow the standard normal distribution, and p_{ik} is set equal to the probability that d_{ik} falls within a threshold denoted by t_k. For ease of computation, however, this was replaced by

$$p_{ik} = \{1 + exp\ (-(t_k - d_{ik}))\}^{-1} \tag{22}$$

here.

The preceding model is analogous to the Euclidean distance-choice model since it assumes a representation of stimuli in multidimensional Euclidean space, and confusion probabilities are related to distances between the stimulus points. The d_{ik} in Equation 22 may be replaced by the unrestricted dissimilarity parameter δ_{ik} ($\delta_{ii} = 0$), analogous to s_{ik} in the unrestricted similarity-choice model. This, however, turns out to be equivalent to what Townsend and Landon (1982) called the modified Nakatani model, which does not restrict p_{ik} in any way. This follows from

$$logit\ (p_{ik}) = ln \left(\frac{p_{ik}}{1 - p_{ik}} \right) = t_k - \delta_{ik} ,$$

and

$$logit\ (p_{kk}) = t_k ,$$

which establish the one-to-one correspondence between $\{p_{ik}\}$ and $\{\delta_{ik}, t_k\}$. Alternatively, d_{ik} in Equation 22 may be replaced by h_{ij} in Equations 14 or 17 or by q_{ij}

in Equation 18. This leads to the unique feature versions of Nakatani's confusion-choice model.

There is one important departure in our implementation of Nakatani's confusion-choice model. The bias parameters are defined as

$$b_j = \frac{f_j}{f} \tag{23}$$

(Hojo, 1982), where $f_j = \Sigma_i f_{j/i}$ and $f = \Sigma_j f_j$. In Nakatani's original procedure, b_j was estimated according to the Least-Squares Criterion, whereas t_k was calculated in an ad hoc manner. Ideally, both b_j and t_k should be estimated according to a well-defined statistical criterion. This was attempted. Too often, however, it led to numerical difficulties. It was decided that t_k was estimated by the maximum likelihood method, and b_j by Equation 23. This decision was dictated because there was no ready-made formula available for t_j, whereas Equation 23 for b_j was obvious (Hojo, 1982).

Symmetric SGM, AON, and OVLP Models

In the symmetric SGM, it is assumed that $p_{c/i} = p_{c/j}$ whenever i and j are in c, and that $p_{c/i} = 0$ whenever i is not in c (Noreen, 1978). The first assumption states that the probability of a confusion set evoked by a stimulus in the set is equal across all stimuli in the confusion set (i.e., $P_{c/i} \equiv p_c$ for all $i \in c$). The second assumption states that the confusion sets evoked by stimulus i always include response i as an admissible response. (Nakatani's confusion-choice model does not satisfy these conditions.) Under these assumptions, Equation 20 becomes

$$p_{j/i} = \sum_{c \in C_{ij}} p_c \left(\frac{b_j}{\Sigma_{k \in c} b_k} \right), \tag{24}$$

where C_{ij} is the set of all the confusion sets that include i and j.

The model can be rewritten as

$$p_{j/i} = b_j z_{ij}, \tag{25}$$

where

$$z_{ij} = \sum_{c \in C_{ij}} \left(\frac{p_c}{\Sigma_{k \in c} b_k} \right)$$

and $\Sigma_j b_j = 1$. Note that $z_{ij} = z_{ji}$ (z_{ij} is symmetric). This is the reason for the name symmetric SGM. Also, $z_{ii} \geq z_{ij}$ for all j. This inequality holds, since $C_{ii} = C_i \supseteq C_{ij}$. Model 25 can be further rewritten in the form of Equation 9 by setting

$$s_{ij} = \frac{z_{ij}}{(z_{ii} z_{jj})^{1/2}} \quad \text{and} \quad w_j = \frac{b_j (z_{jj})^{1/2}}{\Sigma_k b_k (z_{kk})^{1/2}}$$

(J. E. K. Smith, 1980). This implies that the symmetric SGM is a special case of the similarity-choice model. (Note that Equation 25 satisfies the cycle condition and the column constraint, which are necessary and sufficient for the unrestricted similarity-choice model.) However, the reverse is not necessarily true. Although Equation 9 can be put in the form of Equation 25 by letting

$$z_{ij} = \frac{s_{ij}}{v_i v_j} \sum_k v_k w_k \qquad \text{and} \qquad b_j = \frac{v_j w_j}{\sum_k v_k w_k} ,$$

where $v_i = \Sigma_j w_j s_{ij}$, it could lead to inadmissible values of p_c in the symmetric SGM.

Equation 25 along with the conditions on z_{ij} ($z_{ii} \geq z_{ij} = z_{ji} \leq z_{jj}$) is important, since any models that can be expressed in the form of Equation 25 are special cases of the symmetric SGM and, consequently, special cases of the similarity-choice model. In the all-or-none model (Townsend, 1971), it is assumed that stimulus i is identified perfectly with probability p_i, but with the remaining probability, $1 - p_j$, a confusion state is evoked that elicits response j with probability b_j^*. The model can be formally written as

$$p_{j/i} = \begin{cases} (1 - p_i) b_j^* & \text{for } j \neq i , \\ p_i + (1 - p_i) b_i^* & \text{for } j = i , \end{cases} \tag{26}$$

where $\Sigma_j b_j^* = 1$. This model can be rewritten in the form of Equation 25 by setting

$$z_{ij} = a_i a_j \left(\frac{\Sigma_k b_k^*}{a_k} \right) \qquad \text{for } j \neq i$$

and

$$b_j = \frac{b_j^*}{a_j} \Big/ \left(\sum_k \frac{b_k^*}{a_k} \right) ,$$

where $a_i = 1 - p_i$ and

$$z_{ii} = \left[a_i^2 + \frac{a_i(1 - a_i)}{b_i^*} \right] \sum_k a_k b_k^* .$$

It can be easily verified that

$$z_{ii} \geq z_{ij} = z_{ji} \leq z_{jj} .$$

By implication, the AON model is also a special case of the similarity-choice models.

In the overlap model (Townsend, 1971), it is assumed that with probability p_{ii} stimulus i is identified perfectly. With probability p_{ij}, a confusion state is generated in which the only admissible responses are i and j. The response j is chosen with probability $b_j/(b_i + b_j)$. The model is written as

$$p_{j/i} = \begin{cases} p_{ij}\left(\dfrac{b_j}{b_i + b_j}\right) & \text{for } j \neq i\,, \\ p_{ii} + \sum\limits_{k \neq i} p_{ik}\left(\dfrac{b_i}{b_i + b_k}\right) & \text{for } j = i\,, \end{cases} \tag{27}$$

where $\Sigma\, b_j = 1$ and $\Sigma\, p_{ij} = 1$ for all i. This model is also a special case of the symmetric SGM. Let

$$z_{ij} = \frac{p_{ij}}{b_i + b_j} \qquad \text{for } j \neq i\,,$$

and

$$z_{ii} = \sum_{k \neq i} z_{ik} + \frac{p_{ii}}{b_i}\,.$$

Then Equation 27 can be rewritten in the form of Equation 25, indicating that the OVLP model is a special case of the symmetric SGM, which in turn is a special case of the similarity-choice models. Again, it can be easily verified that $z_{ii} \geq z_{ij} = z_{ji} \leq z_{jj}$. There is one-to-one correspondence between parameters in the OVLP model and those in the unrestricted similarity-choice model. However, a proper solution in the latter may correspond with an improper (inadmissible) solution in the former. A proper solution in the former, on the other hand, always leads to a proper solution in the latter.

The informed guessing model (Pachella et al., 1978) is similar to the OVLP model, except that it has one additional parameter. This additional parameter represents the probability of an uninformative confusion state assumed possible in the informed guessing model. In this confusion state, any response is admissible, and a response is chosen with probability equal to its response bias. When the probability of this confusion state is assumed to be zero, the informed guessing model reduces to the OVLP model. It can be easily verified, however, that even with the additional parameter the informed guessing model is a special case of the symmetric SGM.

A hierarchy of the SGM is presented in Figure 13.2. It would be interesting to compare the GOF of the models discussed in this section as well as those discussed in the third section on an empirical basis.

SOME EMPIRICAL RESULTS

The models discussed in the previous section were applied to four data sets, and the results are reported in this section. The four data sets are from Keren and Baggen (1981), Wickelgren (1965), Hodge and Pollack (1962), and Clark and Stafford (1969). All are empirically interesting. However, they are all aggregated data, and the results may be confounded with individual differences.

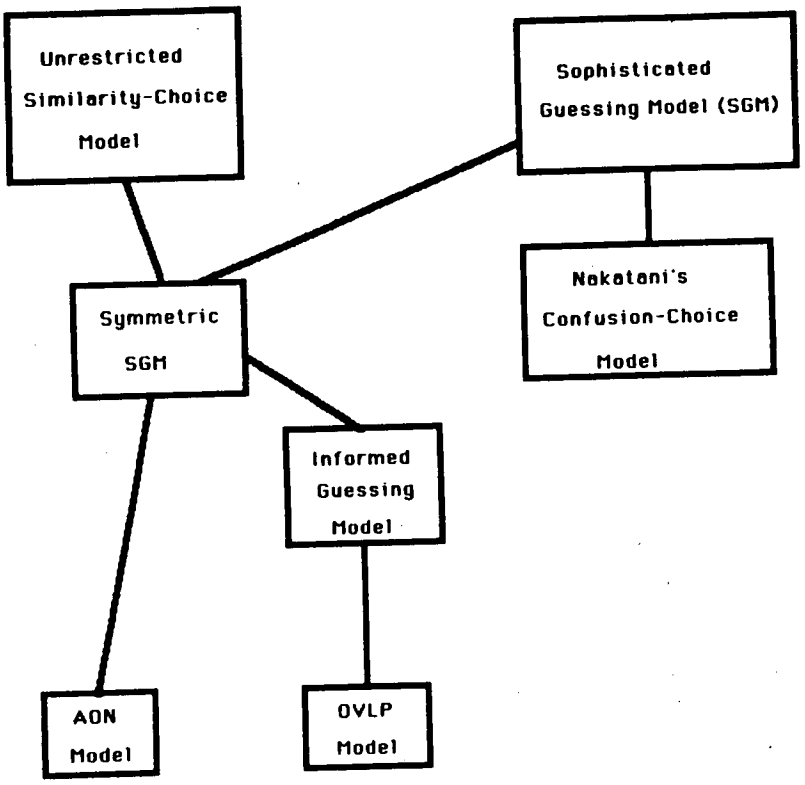

FIG. 13.2. A hierarchy of sophisticated guessing models.

Keren and Baggen's Data

The Keren and Baggen data set was used by Takane and Shibayama (1986) to compare similarity-choice models discussed in the third section. It was found that the unrestricted similarity-choice model fit the data best. ADDTREE and EXTREE were fit as special cases of the unique feature-choice model. However, neither ADDTREE nor EXTREE fit as well as the unrestricted similarity-choice model.

Sophisticated guessing models, including the AON and the OVLP models and Nakatani's confusion-choice model with various (dis)similarity models, were fit in this study, and results are reported in Table 13.8. The main entries in the table are the AIC values and the effective numbers of parameters in the fitted models given in parentheses. It is immediately clear that the AON model does not provide a good fit. This model seems too naive, not only for this data set but also for all other data sets discussed in this section. The OVLP model, on the other

TABLE 13.8
Summary of GOF Statistics for Keren and Baggen's (1981) Data

	Response Model		
Similarity Model	*Similarity-Choice Model*		*Confusion-Choice Model*
0. Saturated Model		17.4 (90)	
1. Unrestricted similarity model	5.6 (54)		8.6 (54)
2. Euclidean distance model	d	d^2	
dim = 2	220.3 (26)	1,148.8 (26)	42.6 (27)
dim = 3	79.2 (33)	438.2 (33)	11.5 (34)
dim = 4	35.6 (39)	90.9 (39)	9.5 (40)
dim = 5	40.9 (44)	20.5 (44)	7.9 (45)
3. Unique feature model			
General asymmetric			
7 features	199.3 (14)		11.6 (24)
9 features	95.3 (16)		−3.8[a](26)
Restricted asymmetric			
7 features	273.7 (8)		18.0 (18)
9 features	122.5 (10)		5.3 (20)
Symmetric-difference			
7 features	197.5 (16)		16.1 (17)
9 features	91.5 (18)		9.4 (19)
4. ADDTREE	164.1 (26)		131.4 (25)
EXTREE	56.1 (28)		26.5 (28)
5. All-or-none model		680.3 (19)	
6. Overlap model			
Proper		5.6 (54)	

AIC-6170.
Effective number of model parameters in parentheses.
[a]Minimum AIC solution.

hand, yielded a proper solution with the GOF equivalent to that of the unrestricted similarity-choice model.

Nakatani's confusion-choice model with various (dis)similarity submodels compares favorably with its similarity-choice model counterparts. In particular, one version of the unique feature model combined with the confusion-choice model was found to work remarkably well. The stimuli used in Keren and Baggen's experiment were 10 rectangular digits defined by subsets of seven segmented features (the seven-feature case). Two additional features, open left and open right, were included in the nine-feature case. The general asymmetric unique feature model (Model 14), with the nine features, combined with the confusion-choice model, proved to be the best fitting model among all the models fitted. This model uses far fewer parameters [n(q) = 26] than the unrestricted similarity-choice model [n(q) = 54]. This result was somewhat surprising, after having obtained disappointing results with the unique feature-choice models (Takane & Shibayama, 1986) using the same set of features. It seems,

however, that whenever the OVLP model yields a proper solution without explicit constraints on its parameters, Nakatani's confusion-choice model generally works very well, and a unique feature model may be found that provides an excellent fit to the data.

Wickelgren's Data

The results for Wickelgren's (1965) data are remarkably similar to those of Keren and Baggen's data. Again, the OVLP model yielded a proper solution without explicit constraints on model parameters, Nakatani's confusion-choice model generally worked very well, and a version of the unique feature confusion-choice model turned out to be the best-fitting model. The stimuli used in Wickelgren's experiment were eight consonant-vowel combinations (fa, af, fo, of, na, an, no, on). Subsets of the stimuli were presented to the subjects, who were asked to recall them shortly after. Three features of the stimuli, the order of the two types of elements (CV or VC), the type of vowel (a or o), and the type of consonant (f or n), were taken as an initial feature set. All possible interactions among the three features (1 and 2, 1 and 3, 2 and 3, and 1, 2, and 3) were then added to form a seven-feature set. The symmetric-difference unique feature model (Model 18) with the seven features combined with the confusion-choice model was found to be the best-fitting model. From the result of the four-dimensional Euclidean distance confusion-choice model, it was conjectured that a subset of the seven features (original three features, 1, 2, and 3, plus the interaction between 1 and 3) might work even better. This was tried, but was found to be not as good as the full seven-feature set.

Hodge and Pollack's Data

The results are somewhat different for Hodge and Pollack's (1962) data, which are summarized in Table 13.9. Hodge and Pollack's data were briefly mentioned in Example 3. The data pertain to eight tones constructed by factorial combinations of two levels each of three physical attributes: frequency (1,000 Hz, 1,006 Hz); intensity (80 dB, 81 dB), and duration (320 ms, 367 ms).

The OVLP model yielded an improper solution; some probabilities were estimated to be negative, and nonnegativity constraints had to be imposed to obtain a proper solution. The GOF of the proper solution (AIC = 57.4) was, however, appreciably worse than that of the improper solution (AIC = 13.8). Still, Nakatani's confusion-choice model fared reasonably well in comparison with its similarity-choice model counterparts. However, the saturated model, in which no special submodels were assumed under $p_{j/i}$, turned out to be the best-fitting model (AIC = 6.5).

The best nonsaturated model was the four-dimensional unsquared Euclidean distance-choice model (AIC = 9.0). The first three of the four dimensions in this model roughly corresponded with the three physical attributes (frequency, inten-

TABLE 13.9
Summary of GOF Statistics for Hodge and Pollack's (1962) Data

	Response Model		
Similarity Model	*Similarity-Choice Model*		*Confusion-Choice Model*
0. Saturated Model		6.5[a](56)	
1. Unrestricted similarity model	13.8 (35)		13.5 (35)
2. Euclidean distance model	d	d^2	
dim = 2	171.7 (20)	958.9 (20)	57.9 (21)
dim = 3	14.4 (25)	284.9 (25)	26.7 (26)
dim = 4	9.0[a](29)	80.1 (29)	10.1 (30)
dim = 5	14.9 (32)	45.0 (32)	11.5 (33)
3. Unique feature model			
General asymmetric			
4 features	245.3 (8)		25.8 (16)
7 features	28.1 (14)		20.7 (22)
Symmetric-difference			
4 features	116.5 (11)		22.4 (12)
7 features	28.1 (14)		14.7 (14)
4. All-or-none model		539.5 (15)	
5. Overlap model			
Improper		13.8 (35)	
Proper		57.4 (35)	

AIC-14450.
Effective number of model parameters in parentheses.
[a]Minimum AIC solution.

sity, and duration) of the stimuli. The fourth dimension represented the three-way interaction among the three attributes. Since all the attributes had only two levels, they were coded into binary features. The unique feature models were fit using these features. However, no unique feature models were found to fit as well as the saturated model. Subsequently, the feature set was incremented by including all two-way interactions, which improved the fit, but not as much as desired. Note that with the seven features (three main effects plus interactions among them) the general asymmetric unique feature-choice model and the symmetric-difference unique feature-choice model provide an identical GOF, which seems to be the case in general.

Clark and Stafford's Data

The fourth data set was reported by Clark and Stafford (1969). The results were similar to those of Hodge and Pollack's data. The stimuli were eight different forms of verbs embedded in sentences. Subjects were shown the sentences and asked to remember the verb. Verbs differed in tense (present or past), in perfective form, and in progressive form. An example would be: (a) watch, (b)

watched, (c) is watching, (d) was watching, (e) has watched, (f) had watched, (8) has been watching, and (i) had been watching.

As in Hodge and Pollack's data, an improper solution was obtained from the OVLP model. This was due to large proportions of errors in the two data sets. The OVLP model requires $p_{i/i} \geq \Sigma_j p_{i/j}$ (J. E. K. Smith, 1980). The difference between the improper solution and the constrained proper solution is much larger, however, in Clark and Stafford's data than in Hodge and Pollack's data. The informed guessing model may be a better choice under this circumstance. However, the informed guessing model suffers from a different kind of problem; model parameters in the informed guessing model are not uniquely determined.

Nakatani's confusion-choice model worked reasonably well. However, the minimum AIC solution was found to be the two-dimensional unsquared Euclidean distance-choice model (AIC = 5.8). The two dimensions in this model roughly corresponded with two of the three defining features of the verb forms: perfective or not perfective and progressive or not progressive. Attempts were made to fit the unique feature models using the defining features of the stimuli and the interactions among them. However, no unique feature models were found to fit better than the best Euclidean distance-choice model.

CONCLUDING REMARKS

This chapter compared a number of existing models of stimulus identification data. One important model was omitted, the general recognition model by Ashby and Perrin (1988). This model is very general and can explain a variety of phenomena that could not be explained by other models. It was not considered here, despite its promise, primarily because it is still under development and because it is too general. In most cases, only specialized models can be fit, and it is not clear what specializations are necessary in particular situations. This situation can improve rapidly (see chaps. 6–8 and 16), however, and the full comparison between this model and the kinds of models discussed in this chapter would undoubtedly be interesting. Such attempts are already underway (Ashby & Lee, 1991).

ACKNOWLEDGMENTS

This work has been supported by Grant A6394 from the Natural Sciences and Engineering Research Council of Canada to the first author. Thanks are due to Greg Ashby, Tony Marley, and an anonymous reviewer for their helpful comments on an earlier version of this paper, to Milton Hodge for providing Hodge and Pollack's data, to Jim Corter for providing the ADDTREE and EXTREE structures, to Marion McGlynn for running the analyses in the third section, and to Marina Takane for preparing Figure 13.2.