JOURNAL OF MATHEMATICAL PSYCHOLOGY 21, 153–167 (1980)

Goodness of Fit Tests for Probabilistic Measurement Models

ALFRED HAMERLE AND GERHARD TUTZ

Abteilung Statistik, Universität Regensburg, Regensburg, West Germany

General goodness of fit tests for probabilistic response models are developed. The tests are applicable in psychophysics, in the theory of choice behavior and in mathematical learning theories. The necessary and sufficient constraints that a measurement model puts on the response probabilities are used for testing this model. In addition, representation theorems for some models are proved and the goodness of fit to experimental data is considered.

1. INTRODUCTION

Probabilistic response models describe a general class of mathematical models which are determined by a systematic connection of the probability of a response and a latent trait variable. Models of this type are found in various fields of psychology, for example in psychophysics, probabilistic choice behavior theory, the theory of dominance and preference and in mathematical learning theory. The general form of these models, as described in the following, is

$$P(R_{a_1 a_2 \cdots a_m}) = F\{G(h_1(a_1),\dots, h_m(a_m))\}$$

where $R_{a_1 a_2 \cdots a_m}$ is the response, $(a_1 ,\dots, a_m) \in A_1 \times \cdots \times A_m$ is a m-tuple of independent variables, $h_1 ,\dots, h_m$ are real valued functions, G is a real valued function and F is a distribution function. The a_i's might be some feature of the experimental situation, or they might describe the stimuli or the subject. Some examples of this form are the so-called Thurstone-Case-V models, where

$$P(R_{ab}) = \Phi(h(a) - h(b))$$

with Φ as the standard normal distribution, the BTL-model, as discussed in the following, or the probabilistic version of Additive Conjoint Measurement (Falmagne, 1978).

In these models the existence of scales $h_i(a_i)$ $(i = 1,\dots, m)$ has to be secured through a measurement theory approach in the form of representation theorems. Traditional measurement theory often uses deterministic models (see for example the extensive discussion in Krantz *et al.*, 1971). Taken in a strict sense, these models must be rejected when one axiom is not fulfilled. In probabilistic measurement these models must be treated as statistical models, and therefore validity must be shown by a goodness of fit

0022-2496/80/020153–15$02.00/0

test. An approach to this problem has been proposed by Falmagne (1978). He considers an equivalent model

$$P(R_{a_1a_2\cdots a_m}) = F(\theta_{a_1a_2\cdots a_m}),$$

where $\theta_{a_1a_2\cdots a_m}$ are solutions of the functional equation

$$\theta_{a_1\cdots a_m} = G(h_1(a_1),..., h_m(a_m)).$$

If constraints (axioms) on the θ's which secure the existence of the real valued functions $h_1(a_1),..., h_m(a_m)$ and G are found, he estimates, under the restriction that the axioms are fulfilled, $\theta_{a_1a_2\cdots a_m}$ instead of the scale values $h_i(a_i)$. These estimates are inserted into a theorem of Wilks (1962, p. 419) to test the goodness of fit of the proposed models. The application of this theorem, however, is restricted to the case where the constraints determine a system of linear equations in the θ's.

In the following a general approach to goodness of fit tests for these models is outlined, which is based on the constraints on the response probabilities. The approach does not use Wilks' results and linearity of the system of constraints is not necessary. In addition, an advantage of the proposed method as compared to Wilks' is that ML-estimation of parameters—which is often only possible with numerical methods—is not necessary. There are cases when the interest of the experimenter is not in the actual parameter values, but only in the goodness of fit. In the last sections the method is demonstrated with some probabilistic response models and with empirical data from attention measurement research.

2. Models with Two Response Categories—Same Sample Size for Each Experimental Condition

Many models deal with two response categories. This may be in the form of "yes–no" answer, the choice between two alternatives, a problem solved or not solved, or other variations. For each experimental condition which corresponds to a combination of the independent variables $A_1,..., A_m$, the response $R_{a_1a_2\cdots a_m}$ or the response $\bar{R}_{a_1a_2\cdots a_m}$ is registered. Test of significance derivation for this case is based on the results of Wald (1943), Neyman (1949) and Bhapkar (1966). Wald considered the following general problem:

Let $f(x_1,..., x_N, \theta_1,..., \theta_K)$ be the joint probability distribution of N independent and identically distributed random variables $X_1,..., X_N$ with K unknown parameters $\theta_1,..., \theta_K$. The parametric space Θ is some subset of the K-dimensional Euclidean space. It is assumed that partial derivations of f up to the second order with respect to the θ_i's exist and are continuous and the matrix

$$I(\theta) = -\frac{1}{N}\left(E_\theta \frac{\partial^2 \ln f}{\partial\theta_i\,\partial\theta_j}\right), \qquad i, j = 1,..., K,$$

is assumed to be positive-definite for all $\theta \in \Theta$.

The hypothesis to be tested is

$$H_0\colon \theta \in \Theta_0 \subset \Theta$$

where Θ_0 is defined by T independent conditions

$$C_t(\theta) = 0, \qquad t = 1,\ldots, T\ (T \leqslant K).$$

The partial derivatives of $C_t(\theta)$ up to the second order are assumed to be continuous functions. Let the $(T \times K)$-matrix $H(\theta)$ be defined by

$$H(\theta) = \left[\frac{\partial C_t(\theta)}{\partial \theta_k}\right], \qquad t = 1,\ldots, T, \quad k = 1,\ldots, K.$$

With

$$h'(\theta) := (C_1(\theta),\ldots, C_T(\theta))$$

for testing the hypothesis H_0, Wald used the statistic

$$Nh'(\hat{\theta})[H(\hat{\theta})\, I^{-1}(\hat{\theta})\, H'(\hat{\theta})]^{-1}\, h(\hat{\theta}), \tag{2.1}$$

where $\hat{\theta}$ is the unconstrained maximum likelihood estimate of θ. Wald showed that the statistic has a limiting χ^2 distribution with T degrees of freedom, if H_0 is true. To apply this general result as a test for probabilistic response models, probabilities given by

$$p_i = P(R_{a_1a_2\cdots a_m}) = F\{G[h_1(a_1),\ldots, h_m(a_m)]\}$$

are considered where every integer i corresponds to an experimental condition $a_1a_2 \cdots a_m$. I is the number of involved experimental conditions. Under each considered experimental condition a sample of N observations is drawn. Let the constraints which are necessary and sufficient for the numerical representation be given in the form

$$C_1(p) = 0,\ldots, C_T(p) = 0$$

where C_t are known functions and $p' = (p_1,\ldots, p_I)$ is the vector of response probabilities, $p_i \in (0, 1)$.

Let the random variables X_{in} be defined by

$$X_{in} = \begin{cases} 1 & \text{if in the } n\text{th observation in the } i\text{th sample} \\ & \text{the response } R_{a_1a_2\cdots a_m} \text{ occurs} \\ 0 & \text{otherwise,} \end{cases}$$

then the random variables $X_1,\ldots, X_N$ with $X_n' := (X_{1n},\ldots, X_{In})$ are independent and identically distributed with the unknown parameters $p_1,\ldots, p_I$. The likelihood function and the loglikelihood function are given by

$$f(x_1,\ldots, x_N\,; p) = \prod_{i=1}^{I} \prod_{n=1}^{N} p_i^{x_{in}}(1 - p_i)^{1-x_{in}} = \prod_{i=1}^{I} p_i^{y_i}(1 - p_i)^{N-y_i}$$

with $x_{in} = 0$ or 1 and $y_i = \sum_{n=1}^{N} x_{in}$ and

$$\ln f(x_1, \dots, x_N ; p) = \sum_{i=1}^{I} [y_i \ln p_i + (N - y_i) \ln(1 - p_i)].$$

The derivations take the form

$$\frac{\partial \ln f}{\partial p_k} = y_k \frac{1}{p_k} - (N - y_k) \frac{1}{1 - p_k}, \qquad k = 1, \dots, I,$$

$$\frac{\partial^2 \ln f}{\partial p_k^2} = -y_k \frac{1}{p_k^2} - (N - y_k) \frac{1}{(1 - p_k)^2},$$

$$\frac{\partial^2 \ln f}{\partial p_k \, \partial p_l} = 0, \qquad k \neq l,$$

and the first moments are

$$E \frac{\partial^2 \ln f}{\partial p_k^2} = -\frac{1}{p_k^2} \sum_{n=1}^{N} EX_{kn} - \frac{N}{(1 - p_k)^2} + \frac{1}{(1 - p_k)^2} \sum_{n=1}^{N} EX_{kn}$$

$$= -\frac{N}{p_k(1 - p_k)}, \qquad k = 1, \dots, I.$$

The maximum likelihood estimate of p_k is $q_k = y_k/N$ and Wald's statistic takes the form

$$h'(q)[H(q)\, S(q)\, H'(q)]^{-1}\, h(q) \tag{2.2}$$

with

$$S(q) = \begin{pmatrix} \frac{y_1(N - y_1)}{N^3} & & \\ & \ddots & \\ & & \frac{y_I(N - y_I)}{N^3} \end{pmatrix}.$$

3. Extension to R-Response Models

The generalization of this test to response models with more than two categories of response is possible to attain without difficulties. For each experimental condition $a_1 a_2 \cdots a_m$ there exist R categories of response ($R > 2$)

$$R^{(1)}_{a_1 a_2 \cdots a_m}, \dots, R^{(R)}_{a_1 a_2 \cdots a_m}.$$

The different combinations $a_1 a_2 \cdots a_m$ of the involved factor categories determine I independent experimental conditions with a sample size N for each combination. The response probabilities

$$p_{ir} := P(R^{(r)}_{a_1 a_2 \cdots a_m})$$

and

$$\sum_{r=1}^{R} p_{ir} = 1, \qquad i = 1, \ldots, I,$$

result in $I(R-1)$ unknown parameters.

Again the model is assumed to be determined by T independent constraints

$$C_t(p) = 0, \qquad t = 1, \ldots, T \; (T \leqslant I(R-1)).$$

Let

$$X_{irn} = \begin{cases} 1 & \text{if in the } n\text{th observation in the } i\text{th sample} \\ & \text{the response } R^{(r)}_{a_1 a_2 \cdots a_m} \text{ occurs} \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Then the joint probability distribution takes the form

$$f(x_{111}, \ldots, x_{IRN}; p) = \prod_{i=1}^{I} \prod_{r=1}^{R} \prod_{n=1}^{N} p_{ir}^{x_{irn}}$$

$$= \prod_{i=1}^{I} \prod_{r=1}^{R-1} p_{ir}^{y_{ir}} \left[1 - \sum_{s=1}^{R-1} p_{is}\right]^{N - \sum_{s=1}^{R-1} y_{is}}$$

where

$$y_{ir} = \sum_{n=1}^{N} x_{irn}, \qquad r = 1, \ldots, R-1, \quad i = 1, \ldots, I.$$

and the derivations and first moments are

$$\frac{\partial^2 \ln f}{\partial p_{ik}\, \partial p_{il}} = -\frac{N - \sum y_{ir}}{(1 - \sum p_{ir})^2}, \qquad k \neq l,$$

$$\frac{\partial^2 \ln f}{\partial p_{ik}^2} = -\frac{y_{ik}}{p_{ik}^2} - \frac{N - y_{ir}}{(1 - \sum p_{ir})^2}, \qquad k = 1, \ldots, R-1,$$

$$E\, \frac{\partial^2 \ln f}{\partial p_{ik}^2} = -N \left(\frac{1}{p_{ik}} + \frac{1}{1 - \sum p_{ir}} \right),$$

$$E\, \frac{\partial^2 \ln f}{\partial p_{ik}\, \partial p_{il}} = -N \frac{1}{1 - \sum p_{ir}}.$$

Then the matrix $I(p)$ is given by

$$I(p) = \begin{pmatrix} A_1 & & - \\ & \ddots & \\ - & & A_I \end{pmatrix}$$

with the symmetric $(R-1) \times (R-1)$-matrices

$$A_i = \begin{pmatrix} \frac{1}{p_{i1}} + \frac{1}{p_{iR}}, & \frac{1}{p_{iR}}, \dots, & \frac{1}{p_{iR}} \\ & \ddots & \\ \cdots & & \frac{1}{p_{i,R-1}} + \frac{1}{p_{iR}} \end{pmatrix}, \qquad i = 1,\dots, I.$$

With the inverse of A_i

$$A_i^{-1} = \begin{pmatrix} p_{i1} - p_{i1}^2, & -p_{i1}p_{i2}, \dots, & -p_{i1}p_{i,R-1} \\ & \ddots & \vdots \\ & \cdots & p_{i,R-1} - p_{i,R-1}^2 \end{pmatrix} \tag{3.2}$$

$I^{-1}(p)$ takes the form

$$I^{-1}(p) = \begin{pmatrix} A_1^{-1} & & - \\ & \ddots & \\ - & & A_I^{-1} \end{pmatrix}$$

with A_i^{-1} from (3.2).

With $q_{ir} = y_{ir}/N$ $(i = 1,\dots, I; r = 1,\dots, R-1)$ the test statistic of Wald is

$$Nh'(q)[H(q)\, I^{-1}(q)\, H'(q)]^{-1}\, h(q), \tag{3.3}$$

where

$$h'(q) = (C_1(q),\dots, C_T(q)),$$

and the $(T \times I(R-1))$-matrix

$$H(q) = \left(\frac{\partial C_t(p)}{\partial p_{ir}}\right)_{p=q}, \qquad \begin{array}{l} t = 1,\dots, T, \\ i = 1,\dots, I, \\ r = 1,\dots, R-1. \end{array}$$

If H_0 holds, then the statistic in (3.3) has a limiting chi square distribution with T degrees of freedom.

4. R-RESPONSE MODELS AND VARIABLE SAMPLE SIZE

Sometimes it is impossible to get the same sample size in each experimental condition, possibly because of the experimental design itself or because of the occurrence of "missing values." In these cases Wald's results are not directly applicable. So in this section the theory is based on Neyman (1949). Again let the hypothesis H_0 (i.e., the model) be determined by T independent functions of p

$$C_t(p) = 0, \qquad t = 1,\ldots, T\ (T \leqslant I(R-1)). \tag{4.1}$$

The sample size of the ith experimental condition is N_i, $i = 1,\ldots, I, \sum N_i = N$. Neyman has shown that H_0, which is the response model, can be tested by using the $\chi_1{}^2$-statistic with

$$\chi_1{}^2 = \sum_{i=1}^{I} \sum_{r=1}^{R} \frac{(y_{ir} - N_i\hat{p}_{ir})^2}{y_{ir}}, \tag{4.2}$$

where $\hat{p}$'s are minimum-$\chi_1{}^2$-estimates subject to the constraints in (4.1). If H_0 holds, the test statistic is distributed in the limit ($N_i \to \infty$, $N_i/N \to c_i < \infty$, $i = 1,\ldots, I$) as χ^2 with T degrees of freedom. Moreover, Neyman has shown that H_0 can also be tested by Pearson's χ^2-statistic

$$\chi^2 = \sum_{i=1}^{I} \sum_{r=1}^{R} \frac{(y_{ir} - N_i\hat{p}_{ir})^2}{N_i\hat{p}_{ir}} \tag{4.3}$$

or the likelihood ratio statistic

$$-2 \ln \lambda = 2 \sum_{i=1}^{I} \sum_{r=1}^{R} y_{ir} \ln \frac{y_{ir}}{N_i\hat{p}_{ir}} \tag{4.4}$$

where $\hat{p}$'s are any BAN (best asymptotical normally distributed) estimates. The three statistics have the same limiting distribution. For details see Neyman (1949).

Here the minimum-$X_1{}^2$ estimates are considered, for Bhapkar (1966) studied the connection with the test statistic used in the former sections. He pointed out that the $X_1{}^2$-statistic of Neyman, if defined, i.e. if all y_{ir} are positive, is algebraically identical to Wald's statistic. This is true in the case where the restrictions

$$C_t(p) = 0, \qquad t = 1,\ldots, T,$$

are linear, and also in the nonlinear case, often occurring in probabilistic response models. In this case Neyman used "linearized" restrictions

$$C_t(q) + \sum_{i=1}^{I} \sum_{r=1}^{R-1} \left[\frac{\partial C_t(p)}{\partial p_{ir}}\right]_{p=q} (p_{ir} - q_{ir}) = 0 \tag{4.5}$$

where

$$q_{ir} = \frac{y_{ir}}{N_i}; \qquad q' = (q_{11}, \ldots, q_{I,R-1}),$$

and showed that the minimum-χ_1^2 estimates subject to constraints (4.5) are also BAN estimates.

Because the statistics of Wald and Neyman are algebraically equivalent, the results of the former sections are applicable. If R equals 2, the minimum of (4.2) subject to the restrictions (4.5) is given by the test statistic (2.2) in Section 2, where the diagonal elements of $S(q)$ have to be substituted by

$$\frac{y_i(N_i - y_i)}{N_i^3}, \qquad i = 1, \ldots, I.$$

If $R > 2$, the minimum is given by the test statistic (3.3) in Section 3, where for the determination of $I^{-1}(p)$ the matrices A_i $(i = 1, \ldots, I)$ have to be substituted by

$$\tilde{A}_i = NN_iA_i.$$

Therefore, if the sample size is not equal, these test statistics can be used to evaluate probabilistic response models. The test statistics have a limiting χ^2 distribution with T degrees of freedom ($N_i \to \infty$, $N_i/N \to c_i < \infty$, $i = 1, \ldots, I$). The response model has to be rejected if the value of the test statistic is greater then $\chi^2_{1-\alpha,T}$.

5. Some Models

In this section some probabilistic response models are discussed which will serve to demonstrate the application of the derived goodness of fit tests. The first one, related to the "strict utility model" (Block & Marschak, 1960) and the Rasch model (Rasch, 1960, 1966), is called "strict latent trait model" and may be generalized without difficulties to multifactorial designs. The second one is a Birnbaum model (see Lord & Novick, 1968, Chaps. 17–20) which involves a special interaction between the factors and is determined by strict unlinear constraints on the response probabilities. The third one is the well-known BTL-model (see for example Luce & Suppes, 1965).

(5.1) Definition. Let A and B be sets, and let $p_{ij} := P(R_{a_ib_j})$ and $P(\bar{R}_{a_ib_j}) = 1 - P(R_{a_ib_j})$ be the response probabilities for the two responses, $R_{a_ib_j}$ and $\bar{R}_{a_ib_j}$, in the combination of factors a_ib_j with $a_i \in A$, $b_j \in B$.

(1) The triple (A, B, P) is a *strict latent trait model*, if there exist functions $h_1 : A \to \mathbb{R}$ and $h_2 : B \to \mathbb{R}$, such that for all $a_i \in A$, $b_j \in B$

$$P(R_{a_ib_j}) = \psi[h_1(a_i) - h_2(b_j)]$$

where $\psi(x) = 1/(1 + e^{-x})$

(2) The triple (A, B, P) is a *Birnbaum model*, if there exist functions $h_1\colon A \to \mathbb{R}$, $h_2\colon B \to \mathbb{R}$, $h_3\colon B \to \mathbb{R}$, such that for all $a_i \in A$, $b_j \in B$

$$P(R_{a_i b_j}) = \psi[h_3(b_j)(h_1(a_i) - h_2(b_j))],$$

where $\psi(x) = 1/(1 + e^{-x})$.

For the following models the existence of a numerical representation which depends upon constraints on the response probabilities is given by representation theorems. The representation theorems can be formulated in various forms. Some kind of representation theorems showing the equivalence of these models to a system of constraints on the response probabilities have been given by various authors, for example by Falmagne (unpublished manuscript, see also Micko (1970)) for the models in (5.1) and for example by Block and Marschak (1960), Luce (1959) or Luce and Suppes (1965) for the BTL-model. However, the representation theorems given by these authors are not sufficient for the incorporation into goodness of fit tests. Therefore, in this section modified representation theorems are derived showing the equivalence of the models to a minimal set of constraints and, moreover, the independence of constraints which is essential for the construction of the goodness of fit tests.

(5.2) Representation Theorem (strict latent trait model). *Let $A, B, P(R_{a_i b_j})$ be as in Definition* (5.1). *Then the following conditions are equivalent*

(1) *The triple (A, B, P) is a strict latent trait model.*

(2) *For $i = 2,..., I, j = 2,..., J$, the following conditions hold,*

$$C_{ij}^{S}(p) := l_{11} - l_{i1} + l_{ij} - l_{1j} = 0, \tag{5.3}$$

where

$$l_{kl} := \ln(p_{kl}/(1 - p_{kl})).$$

Moreover, the $(I - 1)(J - 1)$ constraints in (5.3) *are independent.*

Proof. 1. If (A, B, P) is a strict latent trait model, the constraints in (5.3) follow immediately.

2. Assume that (2) is true. Let $h_1(a_i) := 1 - l_{11} + l_{i1}$, $i = 1,..., I$, and $h_2(b_j) := 1 - l_{1j}$, $j = 1,..., J$, then from (5.3) for all $n \in \{1,..., I\}$, $m \in \{1,..., J\}$,

$$h_1(a_n) - h_2(a_m) = (1 - l_{11} + l_{n1}) - (1 - l_{1m}) = l_{n1} + l_{1m} - l_{11} = l_{nm}.$$

By substitution, it follows that

$$p_{nm} = \frac{\exp(h_1(a_n) - h_2(b_m))}{1 + \exp(h_1(a_n) - h_2(b_m))},$$

i.e., the strict latent trait model.

3. To show the independence of the $(I-1)(J-1)$ constraints let λ_{ij} ($i = 2,\ldots, I$, $j = 2,\ldots, J$) be real numbers such that for all $p \in (0, 1)^{IJ}$

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \lambda_{ij} C_{ij}^{S}(p) = 0$$

holds.

Consider now fixed $n \in \{1,\ldots, I\}$, $m \in \{1,\ldots, J\}$ and let $p_{ij} = \frac{1}{2}$ if $(i, j) \neq (n, m)$; then

$$\lambda_{nm} \ln \frac{p_{nm}}{1 - p_{nm}} = 0$$

and from this $\lambda_{nm} = 0$ follows immediately.

(5.4) Representation Theorem (Birnbaum model). *Let $(A, B, P(R_{a_i b_j}))$ be as in Definition* (5.1). *Then the following conditions are equivalent*:

(1) *The triple (A, B, P) represents a Birnbaum model.*

(2) *For $i = 3,\ldots, I$, $j = 2,\ldots, J$, the following conditions hold,*

$$C_{ij}^{B}(p) := (l_{11} - l_{21})(l_{1j} - l_{ij}) - (l_{1j} - l_{2j})(l_{11} - l_{i1}) = 0, \tag{5.5}$$

where $l_{nm} = \ln(p_{nm}/(1 - p_{nm}))$. Moreover, the $(I-2)(J-1)$ constraints in (5.5) *are independent.*

Proof. Equations (5.5) follow immediately from the Birnbaum model. Assume the conditions in (5.5) are fulfilled. Let

$$\begin{aligned} h_1(a_i) &:= l_{i1} - l_{11} + 1 \qquad \text{where } i = 1,\ldots, I, \\ h_2(b_j) &:= 1 - l_{1j}(l_{11} - l_{21})/(l_{1j} - l_{2j}), \\ h_3(b_j) &:= (l_{1j} - l_{2j})/(l_{11} - l_{21}) \qquad \text{where } j = 1,\ldots, J. \end{aligned}$$

Consider now

$$\begin{aligned} h_3(b_j)(h_1(a_i) - h_2(b_j)) &= \frac{l_{1j} - l_{2j}}{l_{11} - l_{21}} \cdot (l_{i1} - l_{11}) + l_{1j} \\ &= \frac{l_{1j} - l_{2j}}{l_{11} - l_{21}} \frac{l_{1j} - l_{ij}}{l_{1j} - l_{ij}} (l_{i1} - l_{11}) + l_{1j} \\ &= \frac{l_{1j} - l_{2j}}{l_{11} - l_{21}} \frac{l_{11} - l_{i1}}{l_{1j} - l_{ij}} (l_{ij} - l_{1j}) + l_{1j}\,. \end{aligned}$$

Then, from the constraint $C_{ij}^{B}(p) = 0$, which is equivalent to

$$\frac{l_{1j} - l_{2j}}{l_{11} - l_{21}} \frac{l_{11} - l_{i1}}{l_{1j} - l_{ij}} = 1,$$

$h_3(b_j)(h_1(a_i) - h_2(b_j)) = l_{ij}$ follows immediately. Use of the exponential function and solving for p_{ij} yields the Birnbaum model.

The proof of the independence of the constraints is analogous to that in Theorem (5.2).

(5.6) DEFINITION (BTL model). Let A be a set of I objects, and let p_{ij} be the probability that object a_i is preferred to object a_j ($i, j = 1,..., I; i < j$). (A, P) is a BTL model, if there exists a real function $h: A \to \mathbb{R}$ such that for all $a_i, a_j \in A$,

$$p_{ij} = \psi[h(a_i) - h(a_j)]$$

where $\psi(x) = 1/(1 + e^{-x})$, $i, j = 1,..., I; i < j$.

Note that Definition (5.6) describes a special type of BTL models, where only $\binom{I}{2}$ comparisons are considered. In some experimental situations, however, another definition of the BTL model, in which also p_{ii}'s occur, may be expedient. Of course, a representation theorem for the BTL model is well known showing the equivalence of the model to a system of constraints on the response probabilities, given by the "product rule"

$$\frac{p_{ij}}{1 - p_{ij}} \frac{p_{jk}}{1 - p_{jk}} = \frac{p_{ik}}{1 - p_{ik}}.$$

In the following a modified version is derived, using for the numerical representation a minimal set of independent constraints.

(5.7) REPRESENTATION THEOREM (BTL model). *Let* (A, P) *be as in Definition* (5.6). *Then the following conditions are equivalent.*

(1) (A, P) *is a* BTL *model.*

(2) *The* $\frac{1}{2}I(I + 1) - I + 1$ *independent conditions hold*

$$\begin{aligned} l_{1j} + l_{iI} - l_{1I} - l_{ij} &= 0, \qquad i = 2,..., I - 2; \quad j = i + 1,..., I - 1, \\ l_{1j} + l_{jI} - l_{1I} &= 0 \qquad j = 2,..., I - 1, \end{aligned} \tag{5.8}$$

where

$$l_{ij} := \ln \frac{p_{ij}}{1 - p_{ij}}, \qquad i, j = 1,..., I.$$

Proof. 1. Equations (5.8) follow immediately from the BTL model in (5.6).

2. Assume that conditions (5.8) hold. Define

$$\begin{aligned} h(a_1) &:= 0, \\ h(a_i) &:= -l_{1i}, \qquad i = 2,..., I. \end{aligned}$$

For $i = 2,..., I - 2; j = i + 1,..., I - 1$, the first equation in (5.8) yields

$$l_{ij} = l_{1j} + l_{iI} - l_{1I}$$

and from the second equation in (5.8) we derive

$$l_{1I} = l_{1i} + l_{iI}\,.$$

Hence

$$l_{ij} = l_{1j} + l_{iI} - l_{1i} - l_{iI} = l_{1j} - l_{1i} = h(a_i) - h(a_j).$$

For $i = 1; j = 2,..., I$, we obtain from the definition of $h(a_i)$

$$l_{1j} = h(a_1) - h(a_j).$$

For $i = 2,..., I - 1; j = I$, the second equation of (5.8) yields

$$l_{iI} = l_{1I} - l_{1i} = h(a_i) - h(a_I).$$

3. Independence. Let

$$\sum_{\substack{i=2,\ldots,I-2\\ j=i+1,\ldots,I-1}} \lambda_{ij}(l_{1j} + l_{iI} - l_{1I} - l_{ij}) + \sum_{j=2,\ldots,I-1} \gamma_j(l_{1j} + l_{jI} - l_{1I}) = 0, \qquad \text{for all } p \in (0, 1).$$

For any $k \in \{2,..., I - 2\}$ and $l \in \{k + 1,..., I - 1\}$ let $p_{ij} = \frac{1}{2}$ for $(i, j) \neq (k, l)$. This implies

$$\lambda_{kl} l_{kl} = 0 \qquad \text{for all } p_{kl} \in (0, 1)$$

and therefore

$$\lambda_{kl} = 0.$$

Let $m \in \{2,..., I - 1\}$ and $p_{ij} = \frac{1}{2}$ for $(i, j) \neq (1, m)$. Then we have

$$\gamma_m l_{1m} = 0 \qquad \text{for all } p_{1m} \in (0, 1)$$

and therefore

$$\gamma_m = 0.$$

The problem of scale transformations and uniqueness theorems is not relevant for the problem of goodness of fit tests considered here. For this reason uniqueness theorems are not formulated. However, it should be remarked, that the BTL model and the strict latent trait model yield difference scales whereas the Birnbaum model yields a specific interval scale: admissible transformations of h_1, h_2, h_3 are of the type

$$h_1'(a) = \alpha h_1(a) + \beta, \qquad h_2'(b) = \alpha h_2(b) + \beta, \qquad h_3'(b) = (1/\alpha)\, h_3(b).$$

6. Applications to Empirical Data

The goodness of fit of the first two models to empirical data, which are from the field of attention measurement, is evalued below. Schmalhofer (1978) varied combinations of the factors "attention" and "sensory stress" in a reaction time experiment, and registered the frequencies of correct responses under the various conditions.

In a fixed interval of time (variation of the degree of attention), the subject has to determine which of two lamps has been turned off first. The various levels of the factor "sensory stress" and, therefore, the difficulty of the detection task, was controlled by various intervals of time between turning off. The time intervals for the two factors "attention (A)" and "sensory stress (S)" are given in milliseconds. From a sample size of 50 in each combination the following frequencies occurred:

	S		
A	10	20	30
300	0.56	0.58	0.64
350	0.58	0.74	0.92
400	0.60	0.76	0.96
450	0.82	0.96	0.98
500	0.74	0.90	0.98

The constraints of the response models, which occur in the test statistic

$$h'(q)[H(q)\, S(q)\, H'(q)]^{-1}\, h(q),$$

are implicit in the matrix H and the vector h, and are given directly by the independent conditions in the representation theorems (5.2) and (5.4). For the strict latent trait model with the $(I-1)(J-1) = 8$ conditions, h takes the form

$$h'(q) = (C_{22}^S(q),\ldots, C_{44}^S(q))$$

and the (8×15)-matrix

$$H(q) = \left[\frac{\partial C_{ij}(p)}{\partial p_{kl}}\right]_{p=q}.$$

From $C_{ij}^S(p) = l_{11} - l_{i1} + l_{ij} - l_{1j}$ the derivations are

$$\begin{aligned}
\frac{\partial C_{ij}^S(p)}{\partial p_{kl}} &= \frac{1}{p_{kl}(1-p_{kl})}, && k=1,\ l=1;\quad k=i,\ l=j,\\
&= -\frac{1}{p_{kl}(1-p_{kl})}, && k=1,\ l=j;\quad k=i,\ l=1,\\
&= 0 && \text{otherwise.}
\end{aligned}$$

If the model is true, the statistic is asymptotically distributed as χ^2 with 8 degrees of freedom.

The Birnbaum model is determined here by the $(I-2)(J-1)=6$ constraints. With

$$C_{ij}^{B}(p)=(l_{11}-l_{21})(l_{1j}-l_{ij})-(l_{1j}-l_{2j})(l_{11}-l_{i1})$$

h takes the form

$$h(q)'=(C_{32}^{B}(q),\ldots,C_{53}^{B}(q))$$

and the (6×15)-matrix

$$H(q)=\left[\frac{C_{ij}^{B}(p)}{p_{kl}}\right]_{p=q}$$

contains the derivations

$$\begin{aligned}
\frac{\partial C_{ij}^{B}(p)}{\partial p_{kl}} &= \frac{1}{p_{11}(1-p_{11})}(l_{2j}-l_{ij}), && k=1, \quad l=1,\\
&= \frac{1}{p_{1j}(1-p_{1j})}(l_{i1}-l_{21}), && k=1, \quad l=j,\\
&= \frac{1}{p_{21}(1-p_{21})}(l_{ij}-l_{1j}), && k=2, \quad l=1,\\
&= \frac{1}{p_{2j}(1-p_{2j})}(l_{11}-l_{i1}), && k=2, \quad l=j,\\
&= \frac{1}{p_{i1}(1-p_{i1})}(l_{1j}-l_{2j}), && k=i, \quad l=1,\\
&= \frac{1}{p_{ij}(1-p_{ij})}(l_{21}-l_{11}), && k=i, \quad l=j,\\
&= 0 && \text{otherwise.}
\end{aligned}$$

If the Birnbaum model is true, the statistic is asymptotically distributed as χ^2 with 6 degrees of freedom. With a χ^2 of 15.76 the hypothesis of the adequacy of a strict latent trait model had to be rejected. However, postulating the Birnbaum model resulted in a χ^2 of 5.33 and, therefore, could not be rejected.

Therefore, one can assume that the data correspond with the Birnbaum model, where the levels of "sensory stress" are represented by two scales. The "sensory stress" represents the complexity of the task. Thus an interesting analogy results in the interpretation of the scales of the Birnbaum model as item difficulty and item discriminating power in the psychological theory of mental tests (see Lord & Novick, 1968). But here, the interpretation of the scales shall not be investigated further.

The calculation of the values of the test statistic was done with a FORTRAN program at the Rechenzentrum der Universität Regensburg.

7. Concluding Remarks

(1) Although most models use two categories, the proposed goodness of fit tests are also applicable to multicategory models. For example, it is possible that the experimenter at each trial selects a subset of a finite set of stimuli and presents it to the subject. The task of the subject is to choose one stimulus (multichoice paradigm).

(2) The estimation of the unknown parameters of the model makes sense only if the model is correct. If the model has to be rejected for experimental data, the tedious estimation procedure is to no purpose. An advantage of this approach is that the explicit estimation of the unknown parameters is not necessary for testing the goodness of fit.

(3) A possible development of this work concerns the case of unequal response categories. If for different factor combinations the number of response categories is different, applicability of the proposed test statistics is not secured. Though it is assumed that the theory can be extended to this case, it still has to be investigated.

References

Bhapkar, V. P. A note on the equivalence of two test criteria for hypothesis in categorical data. *Journal of the American Statistical Association*, 1966, **61**, 228–235.

Block, H. D., & Marschak, J. Random orderings and stochastic theories of responses. In I. Olkin *et al.* (Eds.), *Contributions to probability and statistics.* Stanford, Calif., 1960. Pp. 97–132.

Falmagne, J. C. Probabilistic choice behavior theory: Axioms as constraints in optimization. In J. N. Castellan, Jr., & F. Restle (Eds.), *Cognitive theory.* New York, 1978. Vol. 3.

Krantz, D. H., Luce, R. D., Suppes, P., & Tversky, A. *Foundations of measurement I.* New York, 1971.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores.* Reading, Mass., 1968.

Luce, R. D. *Individual choice behavior.* New York, 1959.

Luce, R. D., & Suppes, P. Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology, III.* New York, 1965. Pp. 249–410,

Micko, H. C. Eine Verallgemeinerung des Messmodells von Rasch mit einer Anwendung auf die Psychophysik der Reaktionen. *Psychologische Beiträge*, 1970, **12**, 4–22.

Neyman, J. Contributions to the theory of the χ^2-test. In *Proceedings of the Berkeley Symposium of Mathematical Statistics and Probability.* Berkeley: University of California Press, 1949. Pp. 239–273.

Rasch, G. *Probabilistic models for some intelligence and attainment tests.* Kopenhagen, 1960.

Rasch, G. An item analysis which takes individual differences into account. *British Journal of Mathematical and Statistical Psychology*, 1966, **19**, 49–57.

Schmalhofer, F. *Experimente zur gleichzeitigen Messung von sensorischer Beanspruchung und Aufmerksamkeit.* Diplomarbeit, Regensburg, 1978.

Suppes, P., & Zinnes, J. Basic measurement theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology I.* New York, 1963. Pp. 1–63.

Thurstone, L. L. A law of comparative judgment. *Psychological Review*, 1927, **34**, 273–286.

Wald, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 1943, **54**, 426–482.

Wilks, S. S. *Mathematical Statistics.* New York, 1962.

Received: June 27, 1979