THE DUTCH IDENTITY: A NEW TOOL FOR THE STUDY OF ITEM RESPONSE MODELS

PAUL W. HOLLAND
EDUCATIONAL TESTING SERVICE

The Dutch Identity is a useful way to reexpress the basic equations of item response models that relate the manifest probabilities to the item response functions (IRFs) and the latent trait distribution. The identity may be exploited in several ways. For example: (a) to suggest how item response models behave for large numbers of items—they are approximate submodels of second-order loglinear models for $2^J$ tables; (b) to suggest new ways to assess the dimensionality of the latent trait—principle components analysis of matrices composed of second-order interactions from loglinear models; (c) to give insight into the structure of latent class models; and (d) to illuminate the problem of identifying the IRFs and the latent trait distribution from sample data.

Key words: IRT, manifest probabilities, loglinear models, second-order exponential models, moment problems, latent class models, Rasch model, dimensionality.

There are few mathematical tools that have proved useful in the study of the structure of item response models. This is especially true for the so-called "marginal maximum likelihood" approach in which the distribution of the latent variable is integrated out and the would-be analyser is left facing an intractable integral that must be evaluated numerically (Bock & Lieberman, 1970). While the EM algorithm (Bock & Aitkin, 1981) can be used to simplify computing this integral, this fact is mainly useful in computing maximum likelihood estimates for parametric models and does not lead to any insight into the structure of the models themselves.

The purpose of this paper is to introduce a simple tool that does, in some cases, make the integrals disappear and allows the structure of the model to appear in useful new ways. The remainder of this paper is organized as follows. Section 1 sets up the notation; section 2 states and proves the basic result—the Dutch Identity. Section 3 illustrates its use in several problems and section 4 contains additional discussion.

1. Notation

The notation follows that in Holland (1981), Cressie and Holland (1983), and Holland and Rosenbaum (1986). Let $C$ denote a population of examinees and $T$ a specific test. The $(0, 1)$-variable $x_j$ denotes a correct or incorrect response on item $j$ in $T$ and the response vector, $x$, is given by:

$$x = (x_1, \ldots, x_J).$$

Let the proportion of examinees in the population, $C$, who would produce response vector $x$ when tested with $T$ be denoted $p(x)$. Clearly then,
p(x) \geq 0 \quad \text{and} \quad \sum_x p(x) = 1.

The $2^J$ values, $p(x)$, are called the *manifest probabilities* because they can, in principal, be directly estimated from data. Let $X$ be the response vector on test $T$ of a randomly selected examinee from $C$. The probability function for $X$, $\text{Prob} \{X = x\}$, is just $p(x)$, that is,

$$\text{Prob} \{X = x\} = p(x).$$

Item response models restrict the form of the manifest probabilities, $p(x)$, in the following way. First of all, the value of a latent (unobservable) variable, $\theta$, is assumed to be associated with each examinee in $C$ such that given $\theta$, the coordinates of $X$ are independent, that is,

$$e(x = x|\theta) = \prod_j e(x_j = x_j|\theta).$$

The item response functions (IRFs), $P(X_j = 1|\theta) = P_j(\theta)$, are usually restricted in some way, for example, to be monotone increasing in $\theta$ or to have a specified functional form such as the one-, two-, or three-parameter logistic form (Birnbaum, 1968). Let the (cumulative) distribution function of $\theta$ over $C$ be denoted by $F(\theta)$.

Since $x_j$ is $(0, 1)$, we may write

$$P(X_j = x_j|\theta) = P_j(\theta)^{x_j}Q_j(\theta)^{1-x_j},$$

where $Q_j(\theta) = 1 - P_j(\theta)$. The conditional independence assumption may then be written as

$$P(X = x|\theta) = \prod_j P_j(\theta)^{x_j}Q_j(\theta)^{1-x_j}.$$

But, by the usual rules for manipulating conditional probabilities, we have

$$P(X = x) = \int P(X = x|\theta) \, dF(\theta),$$

and consequently all (locally independent) item response models may be viewed as restricting $p(x)$ to have the form

$$p(x) = \int \prod_j P_j(\theta)^{x_j}Q_j(\theta)^{1-x_j} \, dF(\theta),$$

for some choice of IRFs, $\{P_j(\theta)\}$, and some cdf, $F(\theta)$. In this representation, $\theta$ may be a scalar or a vector. I will suppose $\theta$ is a scalar for now but will relax this assumption in section 3.1.

Equation (3) relates the manifest probabilities, $\{p(x)\}$, to the (latent) item response functions, $\{P_j(\theta)\}$, and the distribution of the latent variable, $F(\theta)$. In this paper, as in the discussion of the marginal maximum likelihood approach, (3) is taken to be the defining characteristic of any item response model. The integral in (3) is the "intractable integral" referred to earlier and is often an obstacle to the further understanding of these models.

The manifest probabilities, $\{p(x)\}$, are the governing quantities in the likelihood
function that arises when data are collected by randomly sampling \( N \) examinees from 
\( C \) and testing them with \( T \). In this situation let

\[
n(x) = \text{number of examinees in the sample producing response vector } x.
\]

Then, if \( N \) is small compared to the size of \( C \), \( \{n(x)\} \) follows a multinomial distribution with parameters \( N \) and \( \{p(x)\} \). The likelihood function is the multinomial probability function and (except for a multiplicative constant) is given by

\[
\prod_x p(x)^{n(x)}.
\]

Thus, the log likelihood function is

\[
L = \sum_x n(x) \log p(x).
\]

Hence, it is natural to study the structure of \( \log p(x) \), that is, the "log-manifest probabilities", and that is the approach taken in this paper.

In this context, a model for \( p(x) \) is simply a restriction on the form of \( p(x) \) in (4) and, in particular, the restrictions that define item response models are (3) and other possible restrictions on the form of \( \{P_j(\theta)\} \) and \( F(\theta) \).

Cressie and Holland (1983) studied the structure of the models defined by (3) and were successful in completely characterizing \( p(x) \) in the case of the Rasch model where the IRFs have the form specified by the Rasch model, that is,

\[
\log \frac{P_j(\theta)}{Q_j(\theta)} = a(\theta - b_j),
\]

and \( F \) is unrestricted. In (5), \( a \) is the (common) discrimination parameter, and \( b_j \) is the difficulty parameter for item \( j \). This paper generalizes a formula obtained by Cressie and Holland that reexpresses (3) in a useful way. This generalization is the Dutch Identity.

2. The Dutch Identity

Theorem 1 gives the basic result of this paper.

**Theorem 1: (The Dutch Identity).** If \( p(x) \) satisfies (3) then for any \((0, 1)\)-vector \( y \)

\[
\frac{p(x)}{p(y)} = E\left( \exp \left\{ \sum_j (x_j - y_j)\lambda_j(\theta) \right\} | X = y \right)
\]

where \( \lambda_j(\theta) \) is the item logit function,

\[
\lambda_j(\theta) = \log \left( \frac{P_j(\theta)}{Q_j(\theta)} \right).
\]

Before going through the easy proof of (6), I will make a few comments about it.

First of all, in (6), \( x \) varies over all possible response vectors while \( y \) is a fixed response vector. In a sense, \( y \) is an arbitrary choice of origin. In using Theorem 1 we may choose \( y \) to have desirable properties. The right-hand-side of (6) is the conditional expectation (given that \( X = y \)) of a certain function that involves the vector of item logit functions,
\[ \lambda(\theta) = (\lambda_1(\theta), \ldots, \lambda_J(\theta)). \] More specifically, it is the posterior moment-generating-function of \( \lambda(\theta) \) given that \( X = y \), evaluated at the point \( x - y \).

**Proof.** (This shortened argument was suggested by a referee.) From Bayes theorem it follows that for any \((0, 1)\)-vector, \( y \), we have

\[ dF(\theta) = \frac{p(y)}{P(X = y|\theta)} \frac{dF(\theta|X = y)}{P(X = y|\theta)}. \] (8)

Now substitute (8) into (2) yielding

\[ \frac{p(x)}{p(y)} = \int \frac{P(X = x|\theta)}{P(X = y|\theta)} dF(\theta|X = y). \] (9)

But from (1) and (7) it follows that

\[ P(X = x|\theta) = \exp \left\{ \sum_j x_j \lambda_j(\theta) \right\} \prod_j Q_j(\theta), \] (10)

and

\[ P(X = y|\theta) = \exp \left\{ \sum_j y_j \lambda_j(\theta) \right\} \prod_j Q_j(\theta). \] (11)

Finally, substitute (10) and (11) into (9) and (6) is proved. \( \square \)

This proof follows the type of argument used by Cressie (1982) to prove a similar identity useful in empirical Bayes applications. To my knowledge, the Dutch Identity has never been used in the analysis of item response models, although Cressie and Holland (1983) derived the special case of (6) in which \( y = 0 \). Finally, it should be mentioned that in (6) the fact that \( \theta \) is a scalar is not used and in fact \( \theta \) might be a vector, \( \theta \).

3. Some Applications of the Dutch Identity

3.1 An Item Response Model That is a Second-order Loglinear Model

An item response model for \( p(x) \) involves an integral, but loglinear models for \( p(x) \) are much simpler and merely state that \( \log p(x) \) is linear in some parameters, that is,

\[ \log p(x) = \alpha + b(x)\beta \] (12)

where \( \beta \) is a (column) vector of free parameters of length \( K \), \( b(x) \) is a (row) vector of \( K \) known constants, and \( \alpha \) is the normalizing constant that insures that the \( p(x) \) sum to 1. Loglinear models for \( p(x) \) are loglinear models for \( 2^J \)-contingency tables. These are widely used (e.g., Bishop, Fienberg, & Holland, 1975). Some examples are given below. Throughout the rest of this paper, superscript \( t \) denotes vector or matrix transpose, for example, \( \beta^t \).

1. **Independence.** The coordinates of \( X = (X_1, \ldots, X_J) \) are independent if and only if

\[ \log p(x) = \alpha + \sum_j \beta_j x_j. \] (13)
In this case \( b(x) = (x_1, \ldots, x_J) \) and \( \beta' = (\beta_1, \ldots, \beta_J) \).

2. **Extended Rasch model.** In Cressie and Holland (1983) the following model is discussed in detail

\[
\log p(x) = \alpha + \sum_j \beta_j x_j + \sum_k \gamma_k \delta(k, x_+). \tag{14}
\]

where

\[ x_+ = \sum_j x_j, \]

and

\[ \delta(k, x_+) = \begin{cases} 
1 & \text{if } x_+ = k \\
0 & \text{otherwise}. 
\end{cases} \]

If

\[ b(x) = (x_1, \ldots, x_J, \delta(0, x_+), \ldots, \delta(J, x_+)) \]

and

\[ \beta' = (\beta_1, \ldots, \beta_J, \gamma_0, \ldots, \gamma_J), \]

then (14) defines the class of extended Rasch models as defined in Cressie and Holland (1983). If the \( \{\gamma_k\} \) are restricted by the inequalities indicated by Cressie and Holland, then (14) defines the class of Rasch models.

3. **Second-order exponential models.** Tsao (1967) defines a second-order exponential (SOE) model by

\[
\log p(x) = \alpha + \sum_j \beta_j x_j + \sum_{r<s} \gamma_{rs} x_r x_s. \tag{15}
\]

In this case

\[ b(x) = (x_1, \ldots, x_J, x_1 x_2, x_1 x_3, \ldots, x_{J-1} x_J), \]

and

\[ \beta' = (\beta_1, \ldots, \beta_J, \gamma_{12}, \gamma_{13}, \ldots, \gamma_{J-1,J}). \]

Tsao’s SOE models are examples of the general loglinear model.

An interesting question is whether or not an item response model satisfying (3) can ever be equivalent to a SOE model. This section shows that from the Dutch Identity one may construct an item response model that is a submodel of the class of SOE models. The next section suggests that this construction is far more general than it might first appear. I will state the results as a corollary to Theorem 1 in which \( \theta \) is a column vector.

**Corollary 1.** If, for some choice of \( y \), the posterior distribution of \( \theta|X = y \) is a \( D \)-dimensional normal, that is,

\[ F(\theta|X = y) \text{ is } N_D(\mu_y, \Sigma_y), \]
and if the item logit functions \( \lambda_j(\theta) \) are linear, that is,

\[
\lambda_j(\theta) = \lambda_j(\mu_y) + a_j(\theta - \mu_y)
\]

where \( a_j = (a_{ij}, \ldots, a_{Dj}) \) then

\[
\log p(x) = \alpha + (x - y)'\lambda(\mu_y) + 1/2(x - y)'\Lambda\Sigma_y\Lambda'(x - y),
\]

where

\[
A' = (a'_1, \ldots, a'_j), \quad \text{and} \quad \alpha = \log p(y).
\]

I first prove this result using Theorem 1 and then comment on it.

**Proof.** From (6) it follows

\[
\log p(x) = \alpha + \log E(\exp \{(x - y)\lambda(\theta)\}|X = y),
\]

where \( \alpha = \log p(y) \). But by assumption \( \theta|X = y \sim N(\mu_y, \Sigma_y) \), and since \( \lambda(\theta) \) is a linear function of \( \theta \), the posterior distribution of \( \lambda \) is also multivariate normal. Hence,

\[
E(\lambda|X = y) = \lambda(\mu_y),
\]

and

\[
Cov (\lambda|X = y) = \Lambda\Sigma_y\Lambda'.
\]

Now remember that the expected value in (17) is the moment generating function (mgf) of \( \lambda \) evaluated at \( (x - y) \). However, the mgf of a normal variable \( Z \) with mean vector \( \mu \) and covariance matrix \( \Sigma \) evaluated at the vector \( s \) is

\[
E[\exp \{stZ\}] = \exp \{st\mu + 1/2st\Sigma s\}. \quad (19)
\]

Applying (19) to (18) and (17) with \( s = x - y \) yields (16).

To see that (16) is the same as (15), expand the terms in (16) and collect them to form

\[
\log p(x) = (\alpha + 1/2y'By - y'\lambda(\mu_y)) + x'(\lambda(\mu_y) - By) + 1/2x'Bx, \quad (20)
\]

where \( B = \Lambda\Sigma_y\Lambda' \). Now suppose \( B = \Gamma + D_b \) where \( \Gamma \) has a zero diagonal, \( b \) is the diagonal of \( B \) and \( D_b \) is the diagonal matrix based on \( b \). Then (16) is equivalent to

\[
\log p(x) = (\alpha + 1/2y'By - y'\lambda(\mu_y)) + x'(\lambda(\mu_y) - By + b) + 1/2x'Tx, \quad (21)
\]

since \( x_i^2 = x_i \).

If we now make the substitutions

\[
\alpha' = \alpha + 1/2y'By - y'\lambda(\mu_y),
\]

and

\[
\beta = \lambda(\mu_y) - By + b,
\]

we see that (16) is equivalent to

\[
\log p(x) = \alpha' + x'\beta + 1/2x'Tx, \quad (22)
\]

which is just a matrix way of expressing (15) because \( \Gamma \) is symmetric. **QED.**
The fact that an item response model exists that is a nontrivial example of a SOE model (i.e., is not independent) is quite interesting in its own right. Lord (1962) showed that second-order linear (as opposed to loglinear) models can produce score distributions that correspond to tests with negative reliabilities. This is not true of the model specified in (16) because \( A \Sigma y A^t \) is nonnegative definite.

3.2. Item Response Models With Large Numbers of Items

It might be thought that the example given in Corollary 1 is very special but the purpose of this section is to conjecture that it is a limiting form for all "smooth" unidimensional item response models.

When the number of items, \( J \), is large, \( \theta \) is a scalar, \( F \) has a density over the whole real line, and \( y \) is a "typical" response vector, then the posterior distribution of \( \theta \) given \( X = y \) can be expected to be approximately normal, that is,

\[
dF(\theta|X = y) = \frac{1}{\sigma_y} \phi \left( \frac{\theta - \mu_y}{\sigma_y} \right) d\theta,
\]

where \( \phi(z) \) is the unit normal density function. Approximate normality of posterior distributions is called the "Bernstein-Von Mises Theorem" (Basawa & Rao, 1980) and holds in many Bayesian models, Dawid (1970). For graphs that illustrate this phenomena see Thissen, Steinberg and Wainer (1988, pp. 156-157). At present I know of no thorough discussion of the asymptotic posterior normality of latent variable distributions and this would appear to be a interesting area for further research.

If the item logit functions, are differentiable they have the expansion

\[
\lambda_j(\theta) = \lambda_j(\mu_y) + \sum_{i=1}^J \left( \frac{\partial \lambda_j}{\partial \theta} \right) (\theta - \mu_y) + o(|\theta - \mu_y|). \tag{24}
\]

Hence, in a neighborhood of \( \mu_y \) they are essentially linear. But, if \( \sigma_y \) is small, as it will be for large enough \( J \), the \( o(\cdot) \) term in (24) can be ignored because \( \theta \) will lie, with high posterior probability, inside the neighborhood where \( \lambda_j(\theta) \) is linear. Consequently, \( \lambda \) will be approximately multivariate normal with mean vector, \( \lambda(\mu_y) \), and with a covariance matrix given by \( (\partial \lambda/\partial \theta) \sigma_y^{-2} (\partial \lambda/\partial \theta)^t \) which is a \( J \times J \) matrix of rank 1. Hence, because of Corollary 1, it is a reasonable conjecture that the following equation will hold approximately for large \( J \) for any unidimensional item response model for which the IRFs are smooth, and \( F \) is continuous:

\[
\log p(x) = \alpha + (x - y)^t \lambda(\mu_y) + 1/2\sigma_y^{-2} (x - y)^t \left( \frac{\partial \lambda}{\partial \theta} \right) \left( \frac{\partial \lambda}{\partial \theta} \right)^t (x - y). \tag{25}
\]

Equation (25) defines a submodel of the class of SOE models in (15) in which the second-order parameters are restricted to a multiplicative form. In terms of the free parameters that can be independently estimated, (25) is of the following log-quadratic form:

\[
\log p(x) = \alpha + (x^t \beta) + (x^t y)^2. \tag{26}
\]

Equation (26) does not define a loglinear model but rather a submodel of the class of SOE models that has only \( 2J \)-parameters rather than the full set of \( J + (J^2) \) parameters of the general SOE model.

The conjecture that (26) describes many item response models when \( J \) is large depends on these conditions: (a) \( \theta \) is one-dimensional, (b) \( F \) has a density, (c) \( y \) is
chosen so that \(F(\theta|X = y)\) is approximately normal with a small variance, \(\sigma^2\), and (d) \(\lambda_j(\theta)\) is differentiable. Most models in use explicitly or implicitly assume that (a), (b), and (d) hold. In addition, the existence of \(y\) satisfying (c) is just an example of the approximate normality of posterior distributions which is also just a Bayesian version of the central limit theorem and can be expected to hold under very general conditions. I think that the only issue is how large \(J\) needs to be in order for this approximate normality to hold. This is a worthy topic for research, but is not pursued here.

One implication of (26) is that there can be at most two parameters per item consistently estimated for long tests. This is in accord with the general observation that it is often difficult to estimate three or more item parameters in an unrestricted fashion for data sets that involve many examinees and many items, even though IRFs are often parameterized with more than two parameters.

It is easy to weaken the near linearity assumption in (24) to a quadratic model for \(\lambda_j(\theta)\). Under the approximate posterior normality assumption, the moment generating function in (17) can still be evaluated exactly and results in a new “log-nonlinear” model that is worthy of further investigation.

### 3.3. The Study of Test Dimensionality

If \(\theta\) is a vector parameter and has an approximate normal posterior distribution \(F(\theta|X = y)\) for some \(y\), with mean \(\mu_y\) and covariance matrix \(\Sigma_y\), then the generalization of (25) is

\[
\log p(x) = \alpha + (x - y)'\lambda(\mu_y) + 1/2(x - y)' \left( \frac{\partial\lambda}{\partial\theta} \right) \Sigma_y \left( \frac{\partial\lambda}{\partial\theta} \right)' (x - y). \tag{27}
\]

Letting

\[
\beta_j = \lambda_j(\mu_y) \quad \text{and} \quad R = \left( \frac{\partial\lambda}{\partial\theta} \right) \Sigma_y \left( \frac{\partial\lambda}{\partial\theta} \right)',
\]

we have

\[
\log p(x) = \alpha + (x - y)'\beta + 1/2(x - y)'R(x - y). \tag{28}
\]

Equation (28) says that \(p(x)\) satisfies an SOE model in which the matrix of second-order interactions is \(R\). However, the rank of \(R\) is the rank of \(\Sigma_y\) which is the same as the dimensionality of the latent variable \(\theta\). Hence, (28) suggests that a way to “factor-analyze” dichotomous items is to fit a SOE model to the \(2^J\) table, \(\{n(x)\}\), and then to perform a principal components analysis of the matrix of second-order interactions, \(R\). This method will be especially appropriate when there are large numbers of items. It does not make any assumption other than those made in section 3.2.

More generally, I am tempted to propose that a test measures \(D\) dimensions in population \(C\) if representation (3) holds for its manifest probabilities in population \(C\) with \(\theta = (\theta_1, \ldots, \theta_D)\) and if there is a response vector \(y\) such that \(F(\theta|X = y)\) is more concentrated about its center in every direction than \(F(\theta)\) is. If \(F(\theta)\) and \(F(\theta|X = y)\) both possess covariance matrices, \(\Sigma\) and \(\Sigma_y\), then this condition could be expressed as

\[
\Sigma > \Sigma_y \tag{29}
\]

in the sense of positive definiteness. This proposal is based on the idea that if the test really measures all of the coordinates of \(\theta\) then, for at least one response vector, \(y\), our knowledge of \(\theta\) ought to be more precise in every \(\theta\)-direction if the response \(y\) is observed than if the test is not given. A referee also suggested a weaker alternative
definition in which each coordinate of \( \Theta \) has a \( y \) such that conditioning on \( X = y \) decreases the dispersion of that coordinate of \( \Theta \).

3.4. Latent Class Models

The simplest latent class model has just two latent classes, which we can label by two real numbers \( \theta_1 \) and \( \theta_2 \). Then (3) reduces to

\[
p(x) = \sum_{i=1}^{2} \left\{ \prod_{j} P_{j}(\theta_{i})^{x_{j}} Q_{j}(\theta_{i})^{1-x_{j}} \right\} p_{i}, \tag{30}
\]

where \( p_1 + p_2 = 1 \) are the proportions of examinees in \( C \) with \( \theta_1 \) and \( \theta_2 \), respectively.

This latent class model violates the assumption that \( F \) is continuous in the strongest possible way, that is, \( F \) is a two-point distribution. However, the Dutch Identity, (6), is still valid for this case. The posterior distribution, \( F(\theta|X = y) \), is also a two-point distribution concentrated on \( \theta_1 \) and \( \theta_2 \) with

\[
p_1(y) = P(\theta = \theta_1|X = y)
\]

and

\[
p_2(y) = P(\theta = \theta_2|X = y).
\]

Hence the moment generating function in (6) is given by

\[
E \left( \exp \left\{ \sum_{j} (x_{j} - y_{j}) \lambda_{j} \right\} |X = y \right) = p_1(y) \exp \left\{ \sum_{j} (x_{j} - y_{j}) \lambda_{j_1} \right\} + p_2(y) \exp \left\{ \sum_{j} (x_{j} - y_{j}) \lambda_{j_2} \right\}
\]

where \( \lambda_{j_1} = \lambda_{j}(\theta_1), \lambda_{j_2} = \lambda_{j}(\theta_2) \). Applying the Dutch Identity yields

\[
\frac{p(x)}{p(y)} = p_1(y) \exp \left\{ \sum_{j} (x_{j} - y_{j}) \lambda_{j_1} \right\} \left[ 1 + \frac{p_2(y)}{p_1(y)} \exp \left\{ \sum_{j} (x_{j} - y_{j}) (\lambda_{j_2} - \lambda_{j_1}) \right\} \right].
\]

Let \( \delta_{j} = \lambda_{j_2} - \lambda_{j_1} \) and \( \beta_{j} = \lambda_{j_1} \), then taking logs we have

\[
\log p(x) = \alpha + \sum_{j} (x_{j} - y_{j}) \beta_{j} + \log \left( 1 + \exp \left\{ \gamma + \sum_{j} (x_{j} - y_{j}) \delta_{j} \right\} \right), \tag{31}
\]

where \( \alpha = \log \left( p(y)p_1(y) \right) \) and \( \gamma = \log \left( p_2(y)/p_1(y) \right) \).

Let \( \text{LP}(t) \) be the "logistic potential" function, that is,

\[
\text{LP}(t) = \log (1 + \exp \{t\}).
\]

(Note that the derivative of \( \text{LP}(t) \) is the logistic function, hence the name "logistic potential".)

We may express (31) as

\[
\log p(x) = \alpha + (x - y)^{\beta} + \text{LP}(\gamma + (x - y)^{\delta}). \tag{32}
\]

Thus, the Dutch Identity reveals that the two-class latent class model for dichotomous data is a log-nonlinear model of a very special form, (32). It is like the log-
quadratic model in (26) with the quadratic term replaced by LP(γ + (x − y)rβ). Different choices of y can affect α and γ in (26) but not β and β. This representation of the two-point latent class model may be helpful in suggesting alternative ways of fitting these models.

3.5. What Does an Observed Response Vector Tell us About the Value of a Latent Variable?

The estimation of θ in item response models is problematic. The LOGIST program, Wingersky (1983), produces "maximum likelihood" estimates of θ, ̂ θ, while the approach used in BILOG, Mislevy and Bock (1982), produces posterior expectations of θ given each possible response vector, y, that is, E(θ|X = y). However, in my opinion, it has always been a mystery as to exactly what these quantities really mean since

1. the scale of θ depends on the assumption made about Pj(θ),
2. for some choices of F, E(θ|X = y) may not exist, (more of a technical point than an often occurring phenomenon),
3. the "likelihood function" used in LOGIST to compute ̂ θ's is not the real likelihood function (i.e., the probability of the observed data) for many applications—for example, when examinees are sampled from a well-defined population the likelihood function in (4) is the correct one. In (4) θ is not a parameter that can be maximized over.

The Dutch Identity provides a key to understanding this mystery. The equation

\[ p(x) / p(y) = E(e^{(x-y)'λ}|X = y) \tag{33} \]

may be re-expressed in the following way. Let \( r = x - y \) and let \( S_y = \{ r: y + r = x, \text{ where } x \text{ is a } 0/1 \text{ vector} \} \).

Thus \( S_y \) is the set of all \((0, 1, -1)\)-vectors \( r \) such that when added to \( y \) we get a \((0, 1)\)-vector, \( x \), back. Clearly, \( S_y \) depends on \( y \). Now (6) can be written as

\[ E(e^{r'λ}|X = y) = \frac{p(y + r)}{p(y)} \quad \text{all } r \in S_y. \tag{34} \]

Hence (34) says that for each fixed value of \( y \), the moment generating function for the conditional distribution of \( λ \) given that \( X = y \) evaluated at each \( r \in S_y \) equals the ratio \( p(y + r)/p(y) \). Because the manifest probabilities are, in principle (i.e., with enormous samples of examinees), the most that the data can ever determine, (34) implies that for each \( y \), the values of \( E(e^{r'λ}|X = y) \) for \( r \in S_y \) are the most that we can know about \( λ \). Suppose we let

\[ g_r(λ) = e^{r'λ}, \tag{35} \]

then (34) says that

\[ E(g_r(λ)|X = y) = \frac{p(y + r)}{p(y)} \tag{36} \]

for all \( r \in S_y \). Thus, (36) is an example of the generalized moment problem (Kemperman 1968). Equation (36) says that all we can know, even with infinite numbers of examinees, are the values of the expectation of \( g_r(λ) \) for all \( r \in S_y \) for this conditional
distribution. Kemperman shows how knowing some generalized moments can be used to infer facts about the underlying distribution, which in this case is the conditional distribution of $\lambda$ given $X = y$. These facts consist of bounds on the distribution, that is

$$L_y(S) \leq P(\lambda \in S | X = y) \leq U_y(S).$$

In (37) $S$ represents any (measurable) set in $J$-dimensional space, $\lambda$ is the vector of item logits, $L_y(S)$ is a lower bound and $U_y(S)$ is an upper bound. Kemperman illustrates some cases where similar bounds can be explicitly computed. In the present case $L_y(S)$ and $U_y(S)$ are functions of the manifest probabilities, $\{p(x)\}$. The probability, $P(\lambda \in S | X = y)$, as $S$ varies over the subsets of $J$-space, describes what is knowable about $\lambda$ given the response vector, $y$, of some examinee. Suppose that we know exactly the conditional distribution of $\lambda$ given $X$. What would that tell us about $\theta$? If we knew the IRF's, $\{P_j(\theta)\}$, then an event of the form $\lambda = \lambda(\theta) \in S$ would be equivalent to one of the form $\theta \in T$ so that the conditional probability distribution of $\theta$ given $X$ could, in principle, be computed from that of $\lambda$ given $X$. The issue of model fit also arises in the following way. Are the $A_j$ values obtained via a known item logit function $\lambda_j(\theta)$ consistent with the conditional distribution of $\lambda$ given $X$? When the conditional distribution is not known but only bounded, as in (37), the results are even weaker conclusions about model fit and the posterior distribution of $\theta$ given $X$. I believe that further examination of (36) and (37) is an important area of future research into the foundations of item response models because it can shed more light on what can be recovered about latent variable models from the manifest data.

The central role of $k(\theta)$ in (36) suggests that item response model building ought to be in terms of $\lambda_j(\theta)$ rather than $P_j(\theta)$. In fact, the 1- and 2-parameter logistic models are simply examples of linear models for $\lambda_j(\theta)$.

### 3.6. The Rasch Model

The Rasch model has a one-parameter logistic item response function, that is, the logit function, $\lambda_j(\theta)$, has the following linear form:

$$\lambda_j(\theta) = a(\theta - b_j).$$

In addition, the ability distribution, $F(\theta)$, is usually unspecified in most discussions of the Rasch model. In Cressie and Holland (1983) it is shown that for the Rasch model the manifest probabilities, $\{p(x)\}$, have the following loglinear form:

$$\log p(x) = \alpha + \sum_j x_j \beta_j + \sum_k \gamma_k \delta(k, x_+),$$

where $x_+ = \sum x_j$ and $\delta(k, x_+)$ is defined in Example 2 of section 3.1. The parameters $\{\beta_j\}$ are unconstrained and each $\beta_j$ may vary over $(-\infty, \infty)$. The $\{\gamma_k\}$ are the logarithms of a moment sequence, that is,

$$\gamma_k = \log [E(U^k)],$$

for an arbitrary positive random variable $U$. Thus, the $\gamma_k$ are subject to a system of inequalities described in detail in Cressie and Holland and in de Leeuw and Verhelst (1986).

The main tool used by Cressie and Holland (1983) to establish (39) is a form of the Dutch Identity with $y = 0$. We may obtain an alternative formulation of (39) using the general Dutch Identity. This is given in Theorem 2.
Theorem 2. If \( p(x) \) satisfies an item response model with one-parameter logistic IRFs (i.e., (38)) and general \( F \), then, for any choice of \( y \), \( p(x) \) satisfies the loglinear model

\[
\log p(x) = \alpha + \sum_j (x_j - y_j)\beta_j + \sum_k \gamma_k \delta(k, x_+),
\]

where the \( \beta_j \) vary over \((-\infty, \infty)\) and the \( \gamma_k \) have the form

\[
\gamma_k = \log [E(U^k) - y^k]
\]

for \( k = 0, 1, \ldots, J \), and \( U \) is an arbitrary positive random variable.

Proof. From the Dutch Identity we have

\[
p(x) = p(y) E\left[ \exp\left\{ \sum_j (x_j - y_j)\alpha - \beta_j \right\} | X = y \right] = p(y) \exp\left\{ \sum_j (-ab_j)(x_j - y_j) \right\} E\left[ \exp\left\{ a\theta(x_+ - y_+) \right\} | X = y \right].
\]

Now let \( U = e^{a\theta} \) and take logs. This yields

\[
\log p(x) = \alpha + \sum_j (x_j - y_j)\beta_j + \log E[U^{x_+ - y_+} | X = y],
\]

where

\[
\alpha = \log p(y), \quad \text{and} \quad \beta_j = -ab_j.
\]

Then set

\[
\gamma_k = \log E[U^k | X = y]. \quad QED.
\]

Cressie and Holland (1983) show that the total number of nonredundant parameters in (39) is \( 2J - 1 \)—there are \( J \) \( \beta \)'s and \( J - 1 \) \( \gamma \)'s. However, the \( \gamma \)'s are not freely varying parameters and are subject to a system of inequalities. While these inequalities do not restrict the \( \gamma \)'s in a functional way, they do have an interesting impact on the values that the \( \gamma \)'s can take on as the next corollary shows.

Corollary 2. If \( p(x) \) satisfies the hypothesis of Theorem 2 and if \( y \) is such that \( F(\theta | X = y) \) is \( N(\mu_y, \sigma_y^2) \) then \( \gamma_k \) defined in (41) satisfies

\[
\gamma_k = a\mu_y (k - y_+) + 1/2(a\sigma_y)^2(k - y_+)^2
\]

so that \( \gamma_k \) lies on a quadratic curve as a function of \( k \).

Proof. Simply evaluate \( E[e^{a\theta(x_+ - y_+)} | X = y] \) from the proof of Theorem 2 using the moment generating function of an univariate normal distribution. \( QED. \)

Observe that the existence of a \( y \) for which \( \theta | X = y \) has a nearly normal distribution is another example of approximate posterior normality and should hold if \( J \) is large and \( F(\theta) \) is smooth. Hence, it is a natural conjecture that when \( J \) is large, rather than having \( 2J - 1 \) parameters, the Rasch model should behave as if there were only \( J + 1 \) parameters—\( J \) \( \beta \)'s and one \( \gamma \) (i.e., the coefficient of \( (k - y_+)^2 \) in (42)).
4. Discussion

The Dutch Identity is a useful (or, at least, suggestive) tool for the analysis of item response models. The very fact it suggests that long tests should exhibit very little third and higher-order interactions in their manifest probabilities, \( \{ p(x) \} \), is remarkable and not well-known. I have begun Monte Carlo simulation work to investigate how large \( J \) must be in order for SOE models to fit data generated by a model of the form (3). For ten items with Rasch IRFs the fit based on 30,000 simulated examinees is quite good—likelihood ratio chi-squares of 965 on 968 degrees of freedom. For non-Rasch IRFs (either linear logit functions with different slopes or 3PL IRFs) the fit on ten items is not as good. These results are in agreement with the theory in this paper, but there is clearly more work to be done.

A second remarkable fact that the Dutch Identity suggests concerns the number of parameters that can be estimated in a long test. The discussion in section 3.2 suggests that all "smooth" unidimensional item response models ought to approximate a model of the form (26) when \( J \) is large. The model in (26) has only two parameters per item which may be interpreted as the value of \( \lambda_j \) and of its first \( \theta \)-derivative at a single point. This suggests that models that contain three or more parameters per item can only estimate these parameters successfully for one of two reasons; either they are not applied to a large enough item set or the test is not unidimensional. This analysis suggests that there is a "conservation law" for item parameters of the form: a \( D \)-dimensional set of \( J \) items can only support a total of \( (D + 1)J \) parameters when \( J \) is large. Individual items may be able to have more than \( D + 1 \) parameters estimated for them, but only at the expense of fewer estimable parameters for some other items. The total cannot exceed \( (D + 1)J \). It will be very interesting to see if this type of result can explain the well-known problems of item parameter estimation in item response models.

Item response models, that is (3), and SOE models, that is (15), differ in one respect that deserves further comment. The marginal distributions of an item response model are also item response models with the same IRFs and ability distribution, but the marginal distributions of SOE models are not necessarily SOE models. The former is obvious from (3) while the latter is well-known under the rubric of "collapsibility" of loglinear models (Bishop, Fienberg, & Holland, 1975). These two facts might appear to contradict the conjecture that when \( J \) is large item response models will behave like SOE models, but this is not the case. The 3- and higher-way interactions observed in item response models when \( J \) is small may be viewed as only there because a marginal distribution of a larger joint distribution is being considered, that is, when the test is considered as part of a larger test, the 3- and higher-way interaction will disappear in the joint distribution of the responses to the larger test. Thus, there is no contradiction between the marginalization properties of item response models and its absence in loglinear models.

The models that emerge from the Dutch Identity—the log-quadratic model in (26), its generalization in (28), and the model in (32)—are interesting in their own right and efficient methods of fitting them when \( J \) is large ought to be investigated. In particular we may combine (26) and (32) into a general form

\[
\log p(x) = \alpha + \sum_j \beta_j x_j + f(\gamma + \sum_j \delta_j x_j),
\]

(43)

where \( f \) is a specified smoothed function. The class of models in (43) can be analysed for specific choices of \( f \) and may have some other useful special cases.

Finally, in order to use the ideas of section 3.3 to investigate test dimensionality,
methods for fitting the SOE model (15) when $J$ is large need to be developed. The use of sufficient statistics and the exponential parameters ($\beta_j$ and $\gamma_{rs}$ in (15)) may provide ways of avoiding computations that involve the potentially enormous $2^J$ table.

References


Manuscript received 3/18/87

Final version received 7/27/88