PSYCHOMETRIKA—VOL. 53, NO. 1, 71–92
MARCH 1988

TEST THEORY WITHOUT AN ANSWER KEY

WILLIAM H. BATCHELDER
A. KIMBALL ROMNEY

UNIVERSITY OF CALIFORNIA, IRVINE

A general model is presented for homogeneous, dichotomous items when the answer key is not known a priori. The model is structurally related to the two-class latent structure model with the roles of respondents and items interchanged. For very small sets of respondents, iterative maximum likelihood estimates of the parameters can be obtained by existing methods. For other situations, new estimation methods are developed and assessed with Monte Carlo data. The answer key can be accurately reconstructed with relatively small sets of respondents. The model is useful when a researcher wants to study objectively the knowledge possessed by members of a culturally coherent group that the researcher is not a member of.

Key words: test theory, dichotomous responses, signal detection, latent class model, parameter estimation, information pooling.

Introduction

In many situations in the social sciences, researchers attempt to gain objective knowledge about some culture by asking questions to informants. In some of these situations, the informants are assumed to come from a coherent cultural unit, and they are thought to possess knowledge about their common culture. Of course the informants may vary in their competence, so some of their responses may differ, and the problem the researcher faces is to aggregate rationally the data to reach normative conclusions about the culture.

Romney, Weller, and Batchelder (1986) survey the literature of anthropologists who have discussed such problems. Similar problems occur for a linguist studying the syntax of an exotic language, a social psychologist studying the beliefs of a deviant group, or in the pooling of expert opinion. In all these cases, it may be reasonable to assume that the researcher is able to write relevant objective questions: however, the researcher cannot be expected to know the culturally correct answers a priori. This assumption is most likely to be satisfied for dichotomous items of the "yes-no" variety, so this paper develops models and methods for dealing with dichotomous test items when the answer key is unknown.

In the first section, a model is formalized for the case of homogeneous items. Next some estimation theory for the model is presented including an interpretation of the model as a latent class model that permits maximum likelihood estimates for small numbers of informants. In the third section, new estimation theory for a special case of the model is discussed in detail. A method for estimating both the informants' competencies and the missing answers is presented that applies to any number of informants. Section four discusses some applications of the methods to both simulated and empirical

Requests for reprints should be sent to W. H. Batchelder, School of Social Sciences, University of California, Irvine, CA 92717.

This research was supported by NSF Grant No. SES-8320173 to the authors. We gratefully acknowledge comments and suggestions from John Boyd, Tarow Indow, and Kathy Maher as well as the editor and several anonymous referees.

0033-3123/88/0300-8033$00.75/0

data. Section five considers the important question of how many informants are required to obtain accurate estimates of the answers. Section six discusses some limitations of the approach and proposes solutions for later work.

1. The General Model

Assume each of N informants answers each of M dichotomous questions in some domain of their common knowledge. In this paper, the "yes-no" format is analyzed because it may involve response bias, so the case of "unvalenced" response alternatives becomes the special case with no response bias. In typical test theory models for yes-no items, it is assumed that the theorist has the answer key and each respondent's data are scored as corrects or errors. Then the respondent by item *performance data* are analyzed by the model, and the result is the estimation of each respondent's ability (and perhaps response bias parameters), as well as parameters characterizing each item (Lord & Novick, 1968). In our approach, it is not assumed that the answer key is known a priori, so the only data available to analyze are the informant by item *response data.*

To formalize the situation, define three classes of random variables as follows:

1. Response profile data. $\mathbf{X} = (X_{ik})_{N \times M}$, where

$$X_{ik} = \begin{cases} 1 & \text{if informant } i \text{ answers "yes" to item } k \\ 0 & \text{if informant } i \text{ answers "no" to item } k. \end{cases} \tag{1}$$

2. Answer key. $\mathbf{Z} = (Z_k)_{1 \times M}$, where

$$Z_k = \begin{cases} 1 & \text{if correct answer to item } k \text{ is "yes"} \\ 0 & \text{if correct answer to item } k \text{ is "no".} \end{cases} \tag{2}$$

3. Performance profile data. $\mathbf{Y} = (Y_{ik})_{N \times M}$, where

$$Y_{ik} = \begin{cases} 1 & \text{if informant } i \text{ is correct on item } k \\ 0 & \text{if informant } i \text{ is wrong on item } k. \end{cases} \tag{3}$$

The following observation shows how the three classes of variables are interrelated. It is easy to verify by elementary truth tables.

Observation 1. Given any two of **X**, **Z**, and **Y**, we can determine the third by the relationships:

$$\begin{aligned} X_{ik} &= Y_{ik} Z_k + (1 - Y_{ik})(1 - Z_k) \\ Y_{ik} &= X_{ik} Z_k + (1 - X_{ik})(1 - Z_k) \\ Z_k &= X_{ik} Y_{ik} + (1 - X_{ik})(1 - Y_{ik}). \end{aligned} \tag{4}$$

The models that are analyzed are all instances of the "general Condorcet model" (GCM) proposed by Batchelder and Romney (1986). The model assumes that the items are homogeneous in difficulty; however, in section 6 the case of inhomogeneous items is discussed. The GCM characterizes each informant by a *hit rate* given by

$$P_{1i} = \Pr(X_{ik} = 1 \mid Z_k = 1) \tag{5}$$

and a *false alarm rate*

$$P_{0i} = \Pr(X_{ik} = 1 \mid Z_k = 0), \tag{6}$$

where $0 \leq P_{1i}, P_{0i} \leq 1$, $i = 1, 2, \ldots, N$, and $k = 1, 2, \ldots, M$. The GCM is formalized in the next definition.

Definition 1. The general Condorcet model is characterized by the following three axioms:

Axiom 1: Common truth. There is a fixed answer key "applicable" to all informants; that is, each item k has a correct answer, $z_k = 1$ or $z_k = 0, k = 1, 2, \ldots, M$.

Axiom 2: Local independence. The informant-item response random variables satisfy conditional independence given by

$$\Pr\left[(X_{ik})_{N\times M} = (x_{ik}) \mid (Z_k)_{1\times M} = (z_k)\right] = \prod_{k=1}^{M}\prod_{i=1}^{N} \Pr\left(X_{ik} = x_{ik} \mid Z_k = z_k\right), \tag{7}$$

where (z_k) is the answer key.

Axiom 3: Homogeneity of items. Each informant i has a fixed "hit rate" P_{1i} and a fixed "false alarm rate" P_{0i} such that

$$\Pr\left(X_{ik} = 1 \mid Z_k = z_k\right) = \begin{cases} P_{1i} & \text{if} \quad z_k = 1 \\ P_{0i} & \text{if} \quad z_k = 0. \end{cases} \tag{8}$$

If the Z_k are assumed to be independent and identically distributed Bernoulli random variables with $\pi = \Pr(Z_k = 1)$, $0 \leq \pi \leq 1$, then the model can be viewed as a general forced-choice, signal detection model, (Green & Swets, 1966), with a Bernoulli signal presentation schedule. For this interpretation, it is necessary to think of items with $Z_k = 1$ as trials where the signal is in the first interval and items with $Z_k = 0$ as trials where the signal is in the second interval.

A special case of the GCM is the *high threshold model* (HTM) (Blackwell, 1963) given by the following equations:

$$P_{1i} = D_i + (1 - D_i)g_i \tag{9}$$

$$P_{0i} = (1 - D_i)g_i, \tag{10}$$

where $0 \leq D_i \leq 1$ and $0 \leq g_i \leq 1$ are interpreted as informant i's *competence* and *bias*, respectively. Of course in the signal detection case, one knows the answer key a priori.

2. Estimation Theory for the Model

The next observation is obvious, and it shows that estimation theory is elementary if the answer key is known.

Observation 2. If the answer key $\mathbf{Z}$ is known, then the maximum likelihood estimators (MLEs) for the GCM are given by

$$\hat{P}_{1i} = \frac{\sum_{k=1}^{M} X_{ik} Z_k}{\sum_{k=1}^{M} Z_k} \tag{11}$$

$$\hat{P}_{0i} = \frac{\sum_{k=1}^{M} X_{ik}(1 - Z_k)}{M - \sum_{k=1}^{M} Z_k}, \tag{12}$$

where

$$0 < \sum_{k=1}^{M} Z_k < M.$$

If the answer key is unknown, then the likelihood function is given in the next observation.

Observation 3. If $\mathbf{P}_1 = (P_{11}, \ldots, P_{1N})$, $\mathbf{P}_0 = (P_{01}, \ldots, P_{0N})$, and $\mathbf{Z} = (Z_1, \ldots, Z_M)$ are regarded as parameters of the GCM, then the likelihood equation is given by

$$L(\mathbf{X}; \mathbf{P}_1, \mathbf{P}_0, \mathbf{Z}) = \prod_{k=1}^{M} \prod_{i=1}^{N} \left[\frac{P_{1i}(1 - P_{0i})}{P_{0i}(1 - P_{1i})}\right]^{X_{ik}Z_k} \left[\frac{1 - P_{1i}}{1 - P_{0i}}\right]^{Z_k} \left[\frac{P_{0i}}{1 - P_{0i}}\right]^{X_{ik}} [1 - P_{0i}]. \tag{13}$$

Proof. Equation (13) is obtained by rearranging the terms of (7) when (8) is substituted. To verify this, one can consider the truth table for the four cases of possible values of X_{ik} and Z_k. □

It is not straightforward to obtain maximum likelihood estimators of $\mathbf{P}_1$, $\mathbf{P}_0$, and $\mathbf{Z}$ from (13) because each Z_k has a dichotomous parameter space. One algorithm to maximize (13) would be to consider each of the 2^M possible answer keys, use (11) and (12) to obtain $\mathbf{P}_1$ and $\mathbf{P}_0$, and then plug back into (13) to see which combinations maximize the likelihood function. This algorithm is not polynomial-bounded (Carre, 1979, p. 79) since 2^M grows faster than any polynomial function of M. We conjecture that the problem of maximizing (13) is not polynomial-bounded in M, and, in any event, we have not been able to find a computationally feasible solution for large M.

The GCM is not identifiable as Observation 4 shows. This is due to the fact that the model does not contain the restriction that $P_{0i} \leq P_{1i}$.

Observation 4. If $\hat{\mathbf{P}}_1 = (\hat{P}_{1i})_{1 \times N}$, $\hat{\mathbf{P}}_0 = (\hat{P}_{0i})_{1 \times N}$, $\hat{\mathbf{Z}} = (\hat{Z}_k)_{1 \times M}$ are M.L.E.s for the GCM, then so are $\hat{\mathbf{P}}_1^* = (\hat{P}_{0i})$, $\hat{\mathbf{P}}_0^* = (\hat{P}_{1i})$, $\hat{\mathbf{Z}}^* = (\mathbf{I} - \hat{\mathbf{Z}})$, where $\mathbf{I}$ is a $1 \times M$ vector of 1s.

Proof. The result follows by substituting P_{1i} for P_{0i}, P_{0i} for P_{1i}, and $1 - Z_k$ for Z_k in (13). A simple rearrangement of terms verifies that (13) is unchanged. Therefore, if $\hat{\mathbf{P}}_1$, $\hat{\mathbf{P}}_0$, and $\hat{\mathbf{Z}}$ maximize (13), then so do $\hat{\mathbf{P}}_1^*$, $\hat{\mathbf{P}}_0^*$, and $\hat{\mathbf{Z}}^*$. □

Despite the nonidentifiability of the GCM, the problem is not severe since the possible MLEs form duos, and this means that the answer key can be estimated up to a complete reversal. A practical approach would be to select the solution from the duo where $P_{1i} > P_{0i}$ for the majority of informants.

The GCM, and hence the general two-interval, forced-choice signal detection model, bears an interesting relationship to the two-class latent structure model of Lazarsfeld (Clogg, 1981; Lazarsfeld & Henry, 1968, chap. 2).

The two-class latent structure model is a model for response profile data $\mathbf{X}$ in (1). Each row of $\mathbf{X}$ can be regarded as a respondent "signature" over the M questions, and it is a member of the set of 2^M signatures given by

$$S = \{\mathbf{x} = (x_1, \ldots, x_M) \mid x_k \in \{0, 1\}, k = 1, \ldots, M\}. \tag{14}$$

The model attempts to account for the frequencies of each of the 2^M members of S, namely, $f(\mathbf{x})$, that, of course, are nonnegative integers that sum to N.

The model assumes that each respondent falls into one of two latent classes, C_1 and C_2, and respondents in class C_t respond independently of each other and independently

over items with marginal probabilities p_{tk} of responding "yes" to item k in latent class C_t, $t = 1, 2$; $k = 1, 2, \ldots, M$. Finally, the model assumes that π_t is the a priori probability a respondent is in class C_t, where of course $\pi_1 + \pi_2 = 1$.

With these assumptions, it is easy to see that for all $\mathbf{x} \in S$,

$$\begin{aligned}\Pr(\mathbf{x}) &= \pi_1 \Pr(\mathbf{x} \mid C_1) + (1 - \pi_1) \Pr(\mathbf{x} \mid C_2) \\ &= \pi_1 \prod_{k=1}^{M} p_{1k}^{x_k}(1 - p_{1k})^{1 - x_k} + (1 - \pi_1) \prod_{k=1}^{M} p_{2k}^{x_k}(1 - p_{2k})^{1 - x_k}. \end{aligned} \tag{15}$$

The two class model has $2M + 1$ parameters, namely, π_1 and the p_{tk}, and $2^M - 1$ degrees of freedom in the M-way contingency table of the observed $f(\mathbf{x})$. So as long as $M > 3$, the model implies restrictions on the general M-way table. A variety of iterative algorithms are available that provide maximum likelihood estimates of the parameters of the model (see review by Clogg, 1981). Once parameter estimates are obtained, one can estimate respondent recruitment probabilities, namely, the conditional probability that a respondent with signature $\mathbf{x}$ falls into class C_t, and these can be used to classify the respondents. Further, as long as there are sufficiently many respondents, a likelihood ratio chi-square statistic can be computed to assess the goodness of fit of the model.

Observation 5 shows that the GCM is structurally isomorphic to the two class latent structure model if the roles of respondents and items are interchanged, that is, that the data are given by the transpose of $\mathbf{X}$ and the GCM is viewed as a model for accounting for the 2^N item signatures taken over the M items.

Observation 5. If the answer key is unknown for the GCM, then the GCM is structurally isomorphic to the two-class latent structure model with the role of respondents and items interchanged. The two "latent" classes of items are the items where the correct answer is "yes" and the class of "no" items, respectively.

Proof. Axiom 1 of the GCM assumes that the M items fall into two classes, namely, $C_1 = \{k \mid Z_k = 1\}$ and $C_0 = \{k \mid Z_k = 0\}$. Corresponding to each item k is an item-response signature $\mathbf{x}$ corresponding to the k-th column of $\mathbf{X}_{N \times M}$. If π is regarded as the a priori probability that item k falls into class C_1, then Axioms 2 and 3 require that

$$\begin{aligned}\Pr(\mathbf{x}) &= \pi \Pr(\mathbf{x} \mid Z_k = 1) + (1 - \pi) \Pr(\mathbf{x} \mid Z_k = 0) \\ &= \pi \prod_{i=1}^{N} P_{1i}^{x_i}(1 - P_{1i})^{1 - x_i} + (1 - \pi) \prod_{i=1}^{N} P_{0i}^{x_i}(1 - P_{0i})^{1 - x_i} \end{aligned} \tag{16}$$

Equation (16) is structurally identical to (15), where π corresponds to π_1, P_{1i} corresponds to p_{1k} and P_{0i} corresponds to p_{2k}. □

Observation 5 permits us to use standard estimation theory for the two-class latent structure model to estimate the parameters of the GCM. Further, the special cases of the model can be seen as natural restrictions on the parameters of the two class model. For example, the HTM in (9) and (10) is equivalent to the GCM with the restriction $P_{1i} \geq P_{0i}$, and the HTM with no bias ($g_i = \frac{1}{2}$) requires the additional restriction $P_{0i} = 1 - P_{1i}$.

Clogg (1981) discusses his MLLSA program that obtains maximum likelihood estimates of the parameters for a two-class model for dichotomous items (and other generalizations) by using Goodman's (1974) algorithm based on iterative proportional scaling of parameter estimates. MLLSA permits a variety of restricted models to be estimated, it allows for routine examination of the identifiability of model parameters, and it computes likelihood ratio chi-square tests of goodness of fit for various nested models.

Unfortunately for our purposes MLLSA and other related algorithms are restricted to only a few items (six or less for MLLSA), so it can be used to estimate parameters for the GCM and its nested submodels only in cases of a few informants. The reason for this restriction is the nonpolynomial explosion of the signature set in (14) (also noted in the context of Observation 3).

In most social science applications of latent structure models, there are six or fewer questions and many respondents (often several thousand), so it has not been necessary to develop estimation theory when there are many questions. In applications of our model, there will be a small number of informants (the counterpart to questions) though sometimes greater than six. However, a severe restriction on the application of MLLSA (and other approaches based on latent structure analysis) for our model is that it is practically impossible to deal with more than forty or fifty questions (analogous to forty or fifty informants for the latent class models). For example, with forty questions and four informants, there will be a ratio of 2.5 items per signature ($40/2^4$), so perhaps MLLSA can be used. However, with larger numbers of informants, the parameter estimation and especially the goodness of fit tests that MLLSA and other algorithms provide are not usable. Consequently we have developed new estimation theory for the HTM to deal with these cases discussed next.

3. Estimation Theory For The High Threshold Model

In this section, we describe the details of a two step estimation scheme for the HTM. The first step obtains estimates of the informants' competencies ($\hat{D}_i$ s), and the second step uses these to estimate the answer key.

Estimating Respondent Competencies

Define C_{ij} to be an unbiased estimator of the empirical covariance and M_{ij} to be the proportion of matches between informants i and j. Then in terms of Table 1,

$$C_{ij} = \frac{ad - bc}{M(M-1)}, \tag{17}$$

and

$$M_{ij} = \frac{a+d}{M}. \tag{18}$$

Observation 6 presents $E(C_{ij})$ and $E(M_{ij})$ for the HTM in the case where the answer key is governed by a Bernoulli process.

Observation 6. Assume the HTM of (9) and (10), and suppose $\mathbf{Z}$ is governed by a Bernoulli process with $\Pr(Z_k = 1) = \pi$, $0 < \pi < 1$. Then

i. Irrespective of the biases g_i and g_j,

$$E(C_{ij}) = \pi(1-\pi)D_iD_j, \tag{19}$$

and

ii. If $g_i = g_j = \frac{1}{2}$,

$$E(M_{ij}) = \frac{D_iD_j + 1}{2}, \tag{20}$$

regardless of the value of π.

TABLE 1

Joint Response Data on Two Informants

		j = 1	j = 0	
i	1	a	b	a+b
	0	c	d	c+d
		a+c	b+d	M

$$a = \sum_{k=1}^{M} X_{ik}X_{jk}$$

$$b = \sum_{k=1}^{M} X_{ik}(1-X_{jk})$$

$$c = \sum_{k=1}^{M} (1-x_{ik})\, x_{jk}$$

$$d = \sum_{k=1}^{M} (1-x_{ik})(1-x_{jk})$$

Proof. i. Note that

$$ad = \sum_{k=1}^{M} X_{ik}X_{jk} \sum_{\ell=1}^{M} (1 - X_{i\ell})(1 - X_{j\ell})$$

and

$$bc = \sum_{k=1}^{M} X_{ik}(1 - X_{jk}) \sum_{\ell=1}^{M} X_{j\ell}(1 - X_{i\ell}).$$

Putting these together yields

$$\begin{aligned} ad\text{–}bc &= M \sum_{k=1}^{M} X_{ik}X_{jk} - \sum_{k=1}^{M} X_{ik} \sum_{\ell=1}^{M} X_{j\ell} \\ &= (M-1) \sum_{k=1}^{M} X_{ik}X_{jk} - \sum_{k} \sum_{\ell} X_{ik}X_{j\ell}. \end{aligned}$$

Since X_{ik} and X_{jk} are 1–0 random variables, we have

$$E(C_{ij}) = \frac{\sum_{k} \Pr(X_{ik} = 1,\ X_{jk} = 1)}{M} - \frac{\sum\sum_{k \neq \ell} \Pr(X_{ik} = 1,\ X_{j\ell} = 1)}{M(M-1)}.$$

From (9) and (10), let $A_i = D_i + (1 - D_i)g_i$, $B_i = (1 - D_i)g_i$, and A_j, B_j be defined similarly. Further, if $k \neq \ell$, Axiom 2 implies

$$\Pr(X_{ik} = 1,\ X_{j\ell} = 1) = \Pr(X_{ik} = 1) \Pr(X_{j\ell} = 1).$$

Using Axioms 2 and 3 and conditioning on the values of Z_k and Z_ℓ, we have

$$E(C_{ij}) = [\pi A_i A_j + (1-\pi)B_i B_j] - [\pi A_i + (1-\pi)B_i][\pi A_j + (1-\pi)B_j]$$
$$= \pi(1-\pi)(A_i - B_i)(A_j - B_j) = \pi(1-\pi)D_i D_j.$$

ii. Define the random variables

$$M_{ij,k} = \begin{cases} 1 & \text{if } X_{ik} = X_{jk} \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E(M_{ij}) = \left(\frac{1}{M}\right) \sum_{k=1}^{M} \Pr(M_{ij,k} = 1).$$

Further, it is easy to compute

$$\Pr(M_{ij,k} = 1) = D_i D_j + D_i(1-D_j)[\pi g_j + (1-\pi)(1-g_j)]$$
$$+ D_j(1-D_i)[\pi g_i + (1-\pi)(1-g_i)]$$
$$+ (1-D_i)(1-D_j)[g_i g_j + (1-g_i)(1-g_j)].$$

In case $g_i = g_j = \frac{1}{2}$, the above equation becomes

$$\Pr(M_{ij,k} = 1) = \frac{D_i D_j + 1}{2};$$

therefore

$$E(M_{ij}) = \frac{D_i D_j + 1}{2}. \qquad \square$$

We can observe the values of C_{ij} and M_{ij} for each pair of informants. So if we know the value of π, a point estimator of the product $D_i D_j$ is given by

$$\widehat{D_i D_j} = \frac{C_{ij}}{\pi(1-\pi)} = C_{ij}^*. \tag{21}$$

Further, if we do not know the value of π, but if we assume $g_i = g_j = \frac{1}{2}$, a point estimator of $D_i D_j$ is given by

$$\widehat{D_i D_j} = 2M_{ij} - 1 = M_{ij}^*. \tag{22}$$

Neither (21) nor (22) are MLEs of the product $D_i D_j$ since only the summary data in Table 1 are used, and they are not sufficient statistics for $D_i D_j$.

Equations (21) and (22) do suggest two natural ways that the competencies of the HTM can be estimated. From (19), if we know the value of π, we can write the equation

$$\begin{bmatrix} D_1^2 & E(C_{12}^*) & \cdots & E(C_{1N}^*) \\ E(C_{21}^*) & D_2^2 & \cdots & E(C_{2N}^*) \\ \vdots & & & \\ E(C_{N1}^*) & \cdots & \cdots & D_N^2 \end{bmatrix} = \begin{bmatrix} D_1 \\ D_2 \\ \vdots \\ D_N \end{bmatrix} [D_1, \ldots, D_N]. \tag{23}$$

Further, in case we assume no bias ($g_i = \frac{1}{2}$), an identical equation holds with $E(M_{ij}^*)$ replacing $E(C_{ij}^*)$ in (23).

Equation (23) suggests a method of estimating competence; namely, replace the off-diagonal expectation terms with empirical observations C_{ij}^* and seek the vector of

competency estimates $\hat{\mathbf{D}} = \langle \hat{D}_i \rangle_{1 \times N}$ that gives a "best" fit to the equation. This method is dubbed the *covariance method*, and the counterpart with M^*_{ij} replacing C^*_{ij} is called the *matching method*.

A method of dealing with (23) is to use the minimum residual method of factor analysis first described by Comrey (1962). This method ignores the main diagonal. A version of this program is available on SPSS in the PA2 option (see Nie, Hull, Jenkins, Steinbrenner, Bent, 1975). By specifying just one factor, $\hat{\mathbf{D}}$ can be obtained. This is just a trick to get a least squares solution to (23).

An additional advantage of using the minimal residual method is that it gives one a check on the accuracy of the model. If the model holds, then one would expect that additional eigenvalues extracted from the matrix of M^*_{ij} (or C^*_{ij}) values to be small with respect to the first eigenvalue. This check can be regarded as a mild goodness of fit test of the model.

Estimating the Answer Key

Suppose that we know the true values of the competencies, $\mathbf{D} = (D_i)_{1 \times N}$ and biases $\mathbf{g} = (g_i)_{1 \times N}$, and let $\mathbf{z}_{1 \times M}$ be any one of the 2^M possible answer keys. Then Bayes' theorem can be used to calculate the probability that $\mathbf{z}$ is the correct answer key given the response profile data $\mathbf{X}_{N \times M}$. The result is

$$\Pr(\mathbf{Z} = \mathbf{z} \mid \mathbf{X}) = \frac{\Pr(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}) \Pr(\mathbf{Z} = \mathbf{z})}{\Pr(\mathbf{X})}, \tag{24}$$

where $\Pr(\mathbf{X})$ is the value of the numerator summed over $\mathbf{z}$.

Equation (24) can be interpreted as a posterior distribution over the possible answer keys if $\Pr(\mathbf{Z} = \mathbf{z})$ is interpreted as a prior distribution. A reasonable prior distribution is to assume that each answer key is equally likely (this might be achieved by randomizing whether each question is put in a positive or negative mode), and a reasonable policy to obtain a point estimate $\hat{\mathbf{Z}}$ is to select the $\mathbf{z}$ that maximizes the posterior distribution in (24). This policy is analogous to selecting the mode of the posterior distribution in Bayesian estimation of a continuous parameter (Hogg & Craig, 1978, p. 229), and it often yields estimators that coincide with MLEs obtained by classical estimation theory.

Note that under the above assumptions, the $\mathbf{z}$ that maximizes the posterior distribution is also the $\mathbf{z}$ that maximizes $\Pr(\mathbf{X} \mid \mathbf{Z} = \mathbf{z})$. Observation 7 provides a rule for obtaining $\hat{\mathbf{Z}}$.

Observation 7. Consider the HTM and assume we know the competency and bias parameters $\mathbf{D}$ and $\mathbf{g}$. Then the value of $\hat{\mathbf{z}} = (\hat{\mathbf{z}}_k)_{1 \times M}$ that maximizes $\Pr(\mathbf{X} \mid \mathbf{Z} = \mathbf{z})$ is given by $\hat{z}_k = 1$ iff

$$G_k = \sum_{i=1}^{N} \left\{ X_{ik} \ln \left[\frac{[D_i + (1 - D_i)g_i][1 - (1 - D_i)g_i]}{(1 - D_i)^2 g_i (1 - g_i)} \right] - \ln \left[\frac{[1 - (1 - D_i)g_i]}{(1 - D_i)(1 - g_i)} \right] \right\} \geq 0. \tag{25}$$

In particular, if $g_i = \frac{1}{2}$ for all respondents, $\hat{z}_k = 1$ iff

$$G_k = \sum_{i=1}^{N} (2x_{ik} - 1) \ln \left[\frac{1 + D_i}{1 - D_i} \right] \geq 0. \tag{26}$$

Proof: By local independence (Axiom 2)

$$\Pr(\mathbf{X} \mid \mathbf{Z} = \mathbf{z}) = \prod_{k=1}^{M} \prod_{i=1}^{N} \Pr(X_{ik} = x_{ik} \mid Z_k = z_k).$$

The $\hat{z}_k$ can be obtained on a question by question basis by determining whether $z_k = 1$ or $z_k = 0$ maximizes

$$\sum_{i=1}^{N} \ln \Pr (X_{ik} = x_{ik} \mid z_k).$$

Substituting **D** and **g** into the above equation using (9) and (10) and rearranging terms yields the conclusion in (25), and the additional requirement that $g_i = \frac{1}{2}$ yields (26). □

Formulas like (25) and (26) were obtained in Batchelder and Romney (1986) using the classical Neyman-Pearson theorem, and related formulas are well known in the literature on information pooling (Grofman, Feld, and Owen, 1984; Nitzan and Paroush, 1982). One interpretation of (26) is that the criterion for determining the correct answer to a question is based on linear, log odds aggregation since $(1 + D_i)/(1 - D_i)$ can be interpreted as the odds ratio, namely, the probability of a correct response by informant i divided by the probability of an error.

One interesting point about (26) is that the rule differs from a majority rule, that is, pick $\hat{z}_k = 1$ iff a majority of informants give a "yes" response to question k. For large numbers of informants, the two rules lead to the same results in most cases; however, for small N and heterogeneous competencies, it can easily occur that a minority of informants with higher competencies will outweigh a majority of informants with lower competencies.

In practice, we do not have the true values of the D_i. However, if we assume that $g_i = \frac{1}{2}$ for each informant, we can obtain estimates $\hat{D}_i$ using the matching method and insert these in (26). While this approach is not as optimal as obtaining the MLEs from (13) or using MLLSA, it is computationally efficient, and as seen in the next section, it works well in practice.

In empirical applications it may not be desirable to estimate the answer to a question unless the evidence is sufficiently strong to justify a particular choice. In this case, the evidence dimension $(-\infty, \infty)$ for (25) can be partitioned into three regions, where the middle region corresponds to no estimate of Z_k. In section 5, we develop the properties of this approach in the (conservative) case of homogeneous informants.

4. Applications of the Approach

In this section we present two studies of the matching method; the first is a Monte Carlo study and the second uses real data when the number of informants is small enough so that it is computationally efficient to use MLLSA as a comparison baseline.

Monte Carlo Study

In the Monte Carlo study we investigated several cases involving $N = 10$ stat-informants and $M = 50$ questions. In all cases, exactly half of the questions were correctly answered "yes". The true competencies studied were in the set $T = \{0.50, 0.60, 0.70, 0.80, 0.90\}$ and a given case involved coupling five informants with competency D_1 with five with competency D_2.

The fifteen cases investigated comprised all ways of coupling D_1 with D_2 $(D_1 \leq D_2)$. In each case, 100 Monte Carlo runs were conducted from the HTM with no bias, where each run led to a 10×50 response-profile matrix. For each run, the corrected matching matrix from (23) was obtained and submitted to a PC version of the minimum residual method, and the output was an estimate of each of the ten stat-informant's competence. Table 2 reports means and standard deviations of the pool of 500 (5×100) estimated $\hat{D}$ in

TABLE 2

Means and Standard Deviations of Estimates of Competence

	$D_2 = 0.50$	$D_2 = 0.60$	$D_2 = 0.70$	$D_2 = 0.80$	$D_2 = 0.90$
$D_1 = 0.50$	$\bar{D}_1 = 0.50$ $S_1 = 0.15$	$\bar{D}_1 = 0.49$ $S_1 = 0.13$	$\bar{D}_1 = 0.52$ $S_1 = 0.13$	$\bar{D}_1 = 0.51$ $S_1 = 0.13$	$\bar{D}_1 = 0.50$ $S_1 = 0.12$
	$\bar{D}_2 = 0.50$ $S_2 = 0.14$	$\bar{D}_2 = 0.60$ $S_2 = 0.12$	$\bar{D}_2 = 0.70$ $S_2 = 0.11$	$\bar{D}_2 = 0.80$ $S_2 = 0.09$	$\bar{D}_2 = 0.90$ $S_2 = 0.06$
$D_1 = 0.60$		$\bar{D}_1 = 0.60$ $S_1 = 0.12$	$\bar{D}_1 = 0.59$ $S_1 = 0.12$	$\bar{D}_1 = 0.60$ $S_1 = 0.12$	$\bar{D}_1 = 0.60$ $S_1 = 0.12$
		$\bar{D}_2 = 0.60$ $S_2 = 0.12$	$\bar{D}_2 = 0.70$ $S_2 = 0.11$	$\bar{D}_2 = 0.80$ $S_2 = 0.09$	$\bar{D}_2 = 0.90$ $S_2 = 0.06$
$D_1 = 0.70$			$\bar{D}_1 = 0.71$ $S_1 = 0.11$	$\bar{D}_1 = 0.70$ $S_1 = 0.10$	$\bar{D}_1 = 0.70$ $S_1 = 0.10$
			$\bar{D}_2 = 0.70$ $S_2 = 0.10$	$\bar{D}_2 = 0.80$ $S_2 = 0.09$	$\bar{D}_2 = 0.90$ $S_2 = 0.06$
$D_1 = 0.80$				$\bar{D}_1 = 0.80$ $S_1 = 0.09$	$\bar{D}_1 = 0.80$ $S_1 = 0.09$
				$\bar{D}_2 = 0.80$ $S_2 = 0.09$	$\bar{D}_2 = 0.90$ $S_2 = 0.07$
$D_1 = 0.90$					$\bar{D}_1 = 0.90$ $S_1 = 0.06$
					$\bar{D}_2 = 0.90$ $S_2 = 0.06$

Note: Five informants have true competence given by row value (D_1), and five have true competence given by column value (D_2). Each set is based on a sample of 500 estimates.

each of the cells. Each cell reports on two true D values, D_1 given by the row label and D_2 given by the column label.

The first fact about the Monte Carlo data is that the means of the estimated $\hat{D}s$ from the matching method are very near to the true values (within 0.02 in all cases). Further,

this is true regardless of whether the other five informants have the same true competence (main diagonal) or different true competence (off-diagonal). Thus we are led to conjecture that estimators based on the matching method have at most a small bias; however, we have been unable to establish this conjecture analytically.

To obtain a base line for evaluating the matching method, suppose we knew the answer key. Let $\hat{P}_i$ be the proportion of correct for the i-th informant. Then if $\hat{P}_i \geq 0.50$, $\hat{D}_i = 2\hat{P}_i - 1$ is an MLE (and a best estimator) of D_i, with

$$\sigma(\hat{D}) = \left[\frac{(1 + D)(1 - D)}{50}\right]^{1/2}.$$

Plugging in the values of D used in the Monte Carlo study yields $\sigma(.5) = 0.122$, $\sigma(.6) = 0.113$, $\sigma(.7) = 0.101$, $\sigma(.8) = 0.085$, $\sigma(.9) = 0.062$.

When the standard deviations of the estimated D's in the Monte Carlo runs in Table 2 are compared with these MLE lower bounds, it is seen that the correspondence is very close (within 0.03 in all cases). Thus we conclude that the matching method yields estimates of D (based on response profile data) that are almost as efficient as the MLE estimates based on the performance profile data. At least this conclusion is warranted for the values used, $N = 10$ and $M = 50$, that are realistic for field studies. In other runs not reported here, similar conclusions were reached for smaller values of M and N.

We computed measures of skewness and kurtosis of the samples to see if the estimates were approximately normally distributed. We used the SPSS updated formulas designed to conform to Fisher's G statistics (Bliss, 1967, p. 145). The results showed that the sampling distribution of the estimates based on the matching method were not normal in the great majority of cases. The problem was that the distributions were significantly skewed, probably because the probability of correct was 0.75 or larger on only 50 questions.

The Monte Carlo study did not permit a stringent test of the ability of the Bayesian method to reconstruct the answer key. The reason is that with ten informants whose probabilities of correct are so high, very few mistakes are made by the Bayes method in most of the cells investigated. Batchelder, Maher, and Romney (1987) have studied runs with fewer informants, and these tend to validate the Bayes method as a viable means of reconstructing the answer key, and in particular, it outperformed the majority rule in all cases of heterogeneous competencies (also see Maher, 1987).

Examination of Real Data

In Romney, et al (1986), the matching method for estimating competencies along with the answer key estimation rule in (23) was applied to data from $N = 41$ undergraduate students each answering a true-false general information test with $M = 40$ questions. The answer key was known, and the matching method developed in this paper worked quite well in the sense that the key was accurately reconstructed and the estimates of the competencies correlated quite high with the estimates obtained from the key.

Here we report new results from that data set by randomly dividing the informants into ten disjoint sets of four informants each. Four informants per set allows us to use the MLLSA program which generates iterative MLEs based on our modification of the two-class latent structure model in Observation 5. Our adaption of MLLSA also classifies questions in the same way that respondent recruitment probabilities are obtained in its usual application.

The ten data sets were analyzed by MLLSA in three different ways. First, the unconstrained GCM was used to estimate π, P_{1i}, and P_{0i} for each of the four informants (9 parameters estimated) by using the result in Observation 5. Since there are 16 possible

item signatures, there are 6 (16-1-9) degrees of freedom to test the goodness of fit of the GCM. Second, the HTM model with $g = 0.50$ was fit to the data. This restriction led to 5 estimated parameters and 10 degrees of freedom for test. Finally, the HTM with $g = 0.50$ and $\pi = 0.475$ was fit (4 estimated parameters, 11 degrees of freedom for test). The restriction on π was motivated by the fact that the correct answer key had this proportion of true items. A starting configuration with $P_{1i} > P_{0i}$ was used to try and avoid the identifiability problem of the GCM (see Observation 4), and this was successful and yielded estimates satisfying $\hat{P}_{1i} \geq \hat{P}_{0i}$ in all cases. Thus the GCM estimated by MLLSA can be regarded as the unconstrained HTM.

Table 3 reports likelihood ratio chi-square statistics on the three nested versions of the HTM, where $H(2)$ is nested in $H(1)$, and $H(1)$ is nested in $H(0)$. Each data set provided an average of 2.5 observations per signature, so we decided to treat these as asymptotic chi-square values. In no case was the unconstrained HTM rejected at the $\alpha = 0.05$ level, and summing the chi-square values over the ten (independent) sets yielded a value $\chi^2(63) = 64.88$ which does not even approach significance. Thus the unconstrained HTM does a good job of accounting for the general information data.

The two restricted versions of the HTM also fit the results adequately except for set 6, although the overall fit is not as good as for the unconstrained HTM (see the chi-square values summed over the ten sets). The unconstrained HTM gives an improvement

TABLE 3

Goodness of Fit Analyses

Data Set	H(0)	df	H(1)	df	H(2)	df	H(1)-H(0)	df	H(2)-H(1)	df
1	1.44	6	6.71	10	7.04	11	5.27	4	0.33	1
2	5.21	6	10.05	10	11.47	11	4.84	4	1.42	1
3	7.69	7	11.38	10	11.39	11	3.69	3	0.01	1
4	6.98	6	10.61	10	11.11	11	3.63	4	0.50	1
5	6.06	6	8.93	10	9.84	11	2.87	4	0.91	1
6	13.45	7	27.70**	10	27.72**	11	14.25**	3	0.02	1
7	1.62	6	11.94	10	12.02	11	10.32*	4	0.08	1
8	7.62	6	13.43	10	15.15	11	5.81	4	1.72	1
9	11.96	7	16.77	10	23.36*	11	4.81	3	6.59*	1
10	2.85	6	5.34	10	8.09	11	2.49	4	2.75	1
TOTAL	64.88	63	122.86	100	137.19	110	57.98*	37	14.33	10

* $p < 0.05$; ** $p < 0.01$

Note: Likelihood ratio chi-square statistic for MLLSA fits to unconstrained HTM (H(0)), HTM with $g = 0.500$ (H(1)), and HTM with $g = 0.500$ and $\pi = 0.475$ (H(2)), along with degrees of freedom (df). The df = 7 in the third column mean only 8 parameters were identified (see Clogg, 1981).

in fit over the HTM with $g = 0.50$ in Sets 6, 7, and overall ($\chi^2(37) = 57.98$, $p = 0.013$). There is no evidence that the restriction to $\pi = 0.475$ represents a significant improvement ($\chi^2(10) = 14.33$, $p = 0.16$).

We obtained estimates with the matching method to compare with the MLEs provided by MLLSA. The matching method assumes $g = 0.50$ but leaves π unconstrained. The results are reported in Table 4.

Table 4 reports means and standard deviations of all the estimates of the unconstrained parameters for the matching method and the MLEs of the three nested versions of the HTM. In the only case where the bias parameter g is unconstrained, the observed mean of 0.54 is close to the constrained value of 0.50; however, the standard deviation is quite large. Thus the close fit of the unconstrained HTM (see Table 3) is due in large part to allowing variability in g. On the other hand, the acceptable fit of the constrained versions of the HTM suggest that the model is fairly robust under violations of the $g = 0.50$ assumption.

Looking at the estimates of the D and π parameters, the matching method gives aggregate results that are comparable to the MLEs provided by MLLSA. To further see this, we correlated the 40 competence estimates from the matching method (MM), the HTM with $g = 0.50$, and π free ($H(1)$), and the HTM with $g = 0.50$ and π constrained to 0.475 ($H(2)$). The results were r[MM, $H(1)$] = 0.91, r[MM, $H(2)$] = 0.96, and r[$H(1)$, $H(2)$] = 0.825. Thus the estimates from MM correlate highly with the MLE estimates, and in fact the lowest correlation is between the two versions of the HTM using MLLSA.

These results are surprising because the two MLE methods look at all 16 signatures; hence they look at four-way interactions among informants, whereas the matching

TABLE 4

Descriptive statistics for Estimates

Method	$\bar{\hat{D}}$	$S_{\hat{D}}$	$\bar{\hat{g}}$	$s_{\hat{g}}$	$\bar{\hat{\pi}}$	$S_{\hat{\pi}}$
MM	0.54	0.20	--	--	0.55	0.08
H(0)	0.56	0.19	0.54	0.28	0.52	0.08
H(1)	0.53	0.24	--	--	0.56	0.08
H(2)	0.54	0.22	--	--	--	--

Note: The table reports means and standard deviations for estimates of competence, bias, and π for the matching method, the unconstrained HTM (H(0)), HTM with g = 0.50 (H(1)), and HTM with g = 0.50 and π = 0.475 (H(2)). There are 40 observations on D and g and 10 for π.

method only looks at interactions among pairs of informants. Thus, if the model is wrong, data values could easily exist where the two methods would yield quite different sets of estimates. The fact that they yield very similar estimates for the data under study is additional confirmation that the HTM does a good job of fitting these data.

The MLE methods have built-in goodness of fit tests by examining the likelihood ratio chi-square statistics. A mild goodness of fit test for the matching method is to examine the latent roots of the minimum residual factoring method (used iteratively on residuals from earlier factors) to see if the corrected matching matrix has a one factor solution. When this was examined for our ten sets of data, all but sets 6 and 9 had a strong one factor structure. It is interesting that these sets are the only two that are not fit well by the HTM with $g = 0.50$ (see Table 3). Thus for more than six informants, when MLLSA and other iterative MLE programs cannot be used, the mild goodness-of-fit criterion proposed for the matching method seems to have merit. It would be useful to obtain an approximate distribution of the latent roots under the HTM from the minimum residual solution; however, we have not been able to make progress so far.

Next the answer key was reconstructed for each of the four methods of obtaining estimates. None of the ten sets were expected to do a spectacular job of reconstructing the answer key because each set is based on only four informants. Table 5 (see next section) shows that when the Ds are around 0.50 (as in our ten data sets), one must use 9 or more informants to achieve high degrees of accuracy. Instead, the major point of interest was to compare the Bayesian component of the matching method with the recruitment method of MLLSA based on the MLEs of the D_i. The comparison version of the HTM was $H(1)$ since it constrains only the parameter $g = 0.50$ as does the matching method.

In seven of the ten data sets, there was *no difference* between $H(1)$ and the matching method in their classification of the questions. In set one, there were four (out of 40 possible) disagreements; however, in all four cases the a posteriori probability of classification by the Bayes method was not extreme. There was one difference in set 6 with an unextreme value of the a posteriori probability. Set 9 was the only set that had considerable disagreements. There were 9 questions (out of 40) where $H(1)$ differed from the matching method. The cause of the disagreement was that MLLSA found a configuration of the answer key where the competencies of the four informants were quite different from the solution given by the matching method. In all the other nine data sets, the only disagreements (5 of 450) were cases of close calls. This is additional confirmation that MLLSA and the matching method give similar parameter estimates.

Conclusions of Estimation Studies

With four informants and 40 to 50 questions, MLLSA can be used to obtain MLEs and goodness-of-fit tests for the GCM and its various nested versions. In this case, the matching method appears to give very similar results when the model is true. For six or more informants, MLLSA cannot be used, and Monte Carlo studies suggest that the matching method yields estimates of competency that are not strongly biased and are almost as efficient as the MLEs based on knowing the answer key. For very large numbers of informants, one also can use the majority rule to obtain the answer key and then use Observation 2 to obtain (conditional) MLEs of the D_i. The resulting Ds can be input to the Bayes component of the matching method to obtain a measure of confidence or evidence strength for each question.

Overall, for small, intermediate, and large numbers of informants, the matching method seems to yield very efficient and at worst only slightly biased estimates of the competence parameters.

5. How Many Informants Are Needed

In order to determine the minimal number of informants needed to estimate accurately the answer key, we consider only the conservative case of homogeneous informants and no bias for the HTM. In case $D_i = D$ for all informants, the evidence for $\hat{Z}_k = 1$ in (26) becomes

$$F_k = (2S_k - N) \ln\left(\frac{1+D}{1-D}\right), \tag{27}$$

where

$$S_k = \sum_{i=1}^{N} X_{ik}.$$

The homogeneous informant case is conservative as the next observation shows.

Observation 8. Suppose $\langle D_i \rangle_{i=1}^{N}$ are competencies of N informants with $(1/N) \sum D_i = D$, and $g_i = \frac{1}{2}$. Then $E(G_k \mid Z_k = 1)$ in (26) is at least as large as $E(F_k \mid Z_k = 1)$, where the latter assumes $D_i = D$, for all $i = 1, 2, \ldots, N$.

Proof. In case $D_i = D$,

$$E(F_k \mid Z_k = 1) = [2E(S_k \mid Z_k = 1) - N] \ln\left(\frac{1+D}{1-D}\right)$$

$$= ND \ln\left(\frac{1+D}{1-D}\right).$$

In the heterogeneous case

$$E(2X_{ik} - 1 \mid Z_k = 1) = D_i,$$

so

$$E(G_k \mid Z_k = 1) = \sum_{i=1}^{N} D_i \ln\left(\frac{1+D_i}{1-D_i}\right)$$

$$\geq \sum_{i=1}^{N} D_i \left(\frac{1}{N}\right) \sum_{i=1}^{N} \ln\left(\frac{1+D_i}{1-D_i}\right)$$

since $\ln [(1 + D_i)/(1 - D_i)]$ is an increasing function of D_i.

Since $\sum D_i = ND$, the desired conclusion will obtain if it can be shown that

$$\left(\frac{1}{N}\right) \sum_{i=1}^{N} \ln\left(\frac{1+D_i}{1-D_i}\right) \geq \ln\left(\frac{1+D}{1-D}\right). \tag{28}$$

Equation (28) is an immediate consequence of the lemma that if $\mathbf{W}$ is any random variable with finite expectation and g is convex on $(0, \infty)$, then

$$E[g(\mathbf{W})] \geq g[E(\mathbf{W})] \tag{29}$$

Loéve (1963, p. 159). To apply the lemma, note that $g(D) = \ln [(1 + D)/(1 - D)]$ is convex on $(0, \infty)$ since

$$\frac{d^2 g(D)}{d^2 D} = \frac{4D}{(1 - D^2)^2} > 0.$$

Further, from (28) $\mathbf{D}$ can be interpreted as a random variable with equally probable values $\langle D_i \rangle_{i=1}^{N}$ and, $E(\mathbf{D}) = D$ by hypothesis. With these interpretations, (28) follows immediately from (29). □

Observation 8 shows that if we can estimate the answer key adequately with a group of N homogeneous informants, then we can expect to do no worse with a heterogeneous group of the same average competence. This result is true because an identical argument shows that $E(G_k \mid Z_k = 0)$ is at least as small as $E(F_k \mid Z_k = 0)$ under the conditions of Observation 8.

In the homogeneous case, we decide the answer to a question k by observing the value of S_k. Clearly S_k is integer valued and $0 \leq S_k \leq N$. Suppose we are willing to suspend judgment on the value of Z_k unless the evidence is sufficiently strong. In the case of no response bias, the situation is symmetric, so estimation rules involving a "no decision option" amount to picking an integer decision threshold $c \geq N/2$ and estimating Z_k by the rule

$$\hat{Z}_k = \begin{cases} 1 & \text{if} \quad c \leq S_k \\ \text{No Decision} & \text{if} \quad N - c < S_k < c \\ 0 & \text{if} \quad S_k \leq N - c. \end{cases} \tag{30}$$

The choice of the decision threshold c should depend on the selection of some desired confidence level τ satisfying

$$\Pr(Z_k = 1 \mid S_k \geq c) \geq \tau, \tag{31}$$

then c is selected to be the smallest integer satisfying (31). This assumes that for the selected c, the posterior probability that $Z_k = 1$ must be larger than τ if the evidence exceeds the threshold c. Observation 9 shows how to calculate c as a function of the N, D, and τ.

Observation 9. Assume the HTM with no bias, N informants of common competency D, and τ in the interval $(\frac{1}{2}, 1)$. Let

$$c^* = \frac{\ln\left(\frac{\tau}{1-\tau}\right) + N \ln\left(\frac{1+D}{1-D}\right)}{2 \ln\left(\frac{1+D}{1-D}\right)}. \tag{32}$$

Then if $N/2 < c^* \leq N$, the smallest integer c satisfying (31) exists and is given by smallest integer not exceeded by c^*. If $c^* > N$, then no c satisfies (31).

Proof. For $x = 0, 1, \ldots, N$,

$$\Pr(Z_k = 1 \mid S_k = x) = \frac{\Pr(S_k = x \mid Z_k = 1)\Pr(Z_k = 1)}{\Pr(S_k = x)}$$

$$= \frac{\binom{N}{x}\left(\frac{1+D}{2}\right)^x\left(\frac{1-D}{2}\right)^{N-x}\frac{1}{2}}{\frac{1}{2}\binom{N}{x}\left(\frac{1+D}{2}\right)^x\left(\frac{1-D}{2}\right)^{N-x} + \frac{1}{2}\binom{N}{x}\left(\frac{1-D}{2}\right)^x\left(\frac{1+D}{2}\right)^{N-x}}$$

$$= \frac{1}{1 + \left[\frac{1-D}{1+D}\right]^{2x-N}}.$$

Since the above is an increasing function of x, it is easy to see that $\Pr(Z_k = 1 \mid S_k \geq c) \geq \tau$ iff $c \geq c^*$ in (32). If $c^* > N$, it is clear that $\Pr(Z_k = 1 \mid S_k = x) < \tau$ for all $x = 0, 1, \ldots, N$. □

Given N, D, and a choice of τ, the probability ℓ that the response data will lead to a decisive classification in (30) is given by

$$\begin{aligned} \ell &= 1 - \Pr(N - c < S_k < c) \\ &= 1 - \sum_{n=N-c+1}^{c-1} \binom{N}{n} \left(\frac{1+D}{2}\right)^n \left(\frac{1-D}{2}\right)^{N-n}. \end{aligned} \tag{33}$$

Other things being equal, one would like to select a threshold c that yielded a large value of ℓ.

Now we are in a position to answer an interesting question; namely, given D, what is the smallest value of N that attains prespecified values of τ (confidence) and ℓ (decisiveness)? Table 5 provides a partial answer to this question. It is clear that very high levels of τ and ℓ can be attained with relatively small numbers of informants.

In practice we do not know the respondents' competencies a priori, so the minimal Ns in Table 5 are somewhat under estimated. The larger the number of questions, the more accurate is the competence estimation, so Table 5 can be regarded as valid if M is sufficiently large.

6. Generalizations of the GCM

The GCM is restricted in two salient ways. First, it applies only when there are two response alternatives; and second, it assumes the items are homogeneous. The next two subsections show how these limitations may be removed.

Other Response Spaces

In typical testing situations, there is a general response space and a performance rule that maps responses into graded performance scores such as "pass" or "fail". In this case, absence of the answer key amounts to lack of knowledge of the performance rule, and without more specification, little progress can be made.

Consider a multiple choice situation, where each item has L possible response alternatives and one correct response. The GCM and its special cases can be generalized for this situation, and many of our results and methods still apply. For example, the HTM with no bias can be generalized by

$$\Pr(X_{ik} = \ell) = \begin{cases} \dfrac{D_i(L-1)+1}{L} & Z_k = \ell \\ \dfrac{1-D_i}{L} & Z_k \neq \ell, \end{cases} \tag{34}$$

$\ell = 1, 2, \ldots, L$ and i and k as before. This model is briefly discussed in Romney, et al. (1986).

It is easy to develop the analogue of (20), namely,

$$E(M_{ij}) = \frac{(L-1)D_iD_j + 1}{L}, \tag{35}$$

and of (20), namely,

$$\widehat{D_iD_j} = \frac{LM_{ik} - 1}{L-1}. \tag{36}$$

Table 5

Minimal Number of Informants Needed

Proportion of Questions (ℓ)	Average Level of Cultural Competence (D) 0.50	0.60	0.70	0.80	0.90
.90 Confidence Level (τ)					
.80	9	4	4	4	4
.85	11	6	4	4	4
.90	13	6	6	4	4
.95	17	10	6	6	4
.99	25	16	10	8	4
.95 Confidence Level					
.80	9	7	4	4	4
.85	11	7	4	4	4
.90	13	9	6	4	4
.95	17	11	6	6	4
.99	29	19	10	8	4
.99 Confidence Level					
.80	15	10	5	4	4
.85	15	10	7	5	4
.90	21	12	7	5	4
.95	23	14	9	7	4
.99	*	20	13	8	6
.999 Confidence Level					
.80	19	11	7	6	4
.85	21	13	8	6	4
.90	23	13	10	8	5
.95	29	17	10	8	5
.99	*	23	16	12	7

* Well Over 30 Informants Needed

Note: Table reports the smallest number of informants needed to classify a desired proportion of questions ℓ with a specified confidence level τ when average cultural competence is known.

Thus competencies can be estimated by the matching method used for the HTM with no bias. Further, the Bayesian approach to estimating the answer key is easily extended if it is assumed that each possible alternative is equally likely a priori.

Maher (1987) has applied the approach to multiple choice exam data where the answer key is known, and estimated competencies correlate in the high 90s with those

obtained from the answer key. Also her empirical work shows that the answer key can be estimated from very few informants.

Inhomogeneous Items

One of the restrictive assumptions of the GCM is the item homogeneity axiom (Axiom 3). In most test theory models, each item is allowed to have its own performance characteristics (Lord, 1974). These characteristics are reflected in the item characteristic curve (ICC) which gives the probability of a correct response to an item as a function of one or more latent ability or competence parameters. Each model is characterized by a class of possible ICCs. It is not straightforward to extend these models to our situation since they are developed to deal with performance profile data. However, if the informant-item competencies, D_{ik}, are modeled, progress is possible.

An example of a test theory model with separate competence and item difficulty parameters is the Rasch model (Rasch, 1960). The model is developed for performance profile data, and it postulates respondent "ability" parameters, $\delta_i > 0$ and item "easiness" parameters, $\varepsilon_k > 0$. The model relates the parameters to performance data by

$$\Pr(Y_{ik} = 1) = \frac{\delta_i \varepsilon_k}{1 + \delta_i \varepsilon_k}. \tag{37}$$

Equation (37) is strictly increasing in both δ_i and ε_k, and it satisfies a desirable *monotonicity* condition, namely, for any respondents i and j, if there is an item k, where $\Pr(Y_{ik} = 1) \geq \Pr(Y_{jk} = 1)$, then for all items ℓ, $\Pr(Y_{i\ell} = 1) \geq \Pr(Y_{j\ell} = 1)$.

The Rasch model can be adapted to the HTM by allowing competence, D_{ik}, to depend on both informants and items by

$$D_{ik} = \frac{D_i \varepsilon_k}{1 - D_i + D_i \varepsilon_k}, \tag{38}$$

where $\varepsilon_k > 0$ is an item easiness parameter and, as before, D_i is the informant's competence. Equation 38 is easily seen to be consistent with (37) when the range restriction $0 \leq D_i \leq 1$ is imposed.

In case $\varepsilon_k = 1$ for all items, then $D_{ik} \equiv D_i$. Happily a monotonicity condition is satisfied by (38), since if $D_{ik} \geq D_{jk}$, for any item k, then $D_{i\ell} \geq D_{j\ell}$ for all items ℓ.

The inhomogeneous HTM has informant-item hit and false alarm rates given by

$$P_{1ik} = \frac{[D_i \varepsilon_k + (1 - D_i) g_i]}{[1 - D_i + \varepsilon_k D_i]} \tag{39}$$

$$P_{0ik} = \frac{(1 - D_i) g_i}{[1 - D_i + \varepsilon_k D_i]}, \tag{40}$$

respectively. Further, the model has $2N$ informant parameters, M item easiness parameters, and M dichotomous answer key parameters, so in principle it is estimatable in case $2(N + M) < NM$.

The two extensions of the model reported here are currently under study (see Batchelder, Maher, and Romney, 1986, for preliminary results).

Conclusions

Psychometricians have developed many models and methods for dealing with test data where a unitary answer key is known. This work has made possible the estimation of latent ability parameters of a respondent as well as parameters characterizing an item. These models have been applied extensively in the area of educational testing.

When there is no unitary answer key, as in the case of subjective items, methods such as latent structure analysis have been developed for disaggregating respondents into homogeneous latent groups and characterizing the opinion structure of each group. These methods have been applied by sociologists and public opinion pollsters as well as professionals in other areas of social science.

On the other hand, there has not been much work on the case of inhomogeneous respondents where there is reason to feel that a unitary key exists but it is not known a priori to the researcher. This situation is one faced by social scientists attempting to investigate objective knowledge in cultures different from their own. As pointed out in the introduction, anthropologists, linguists, and social psychologists face this problem routinely.

This paper shows that standard models and methods from test theory can be adapted to handle the case of an unknown answer key. Our approach is to develop parametric models and methods for simple testing situations because they are relatively easy to run on other cultural groups. The methods work rather well even with small groups of informants.

Typical test theory models analyze performance profile data and work by exploiting structure in the item by item association matrix. In the early history of test theory, several psychometricians tried to develop methods for analyzing respondent by respondent association measures; however, the conclusion was that relatively little can be obtained from this approach (for example, Guilford, 1954). This conclusion is true if performance profile data are used, because the models postulate independence over items. However, if response profile data are used, there is large amount of information in respondent by respondent association measures as shown by such results as (19) and (20). This structure turns out to provide the information for models and methods that develop test theory without an answer key.

References

Batchelder, W. H., Maher, K. M., & Romney, A. K. (1986). *Extensions of a model for test theory without an answer key.* Manuscript submitted for publication. (Prepublication draft available upon request).

Batchelder, W. H., & Romney, A. K. (1986). The statistical analysis of a general Condorcet model for dichotomous choice situations. In B. Grofman & G. Owen (Eds.), *Information pooling and group decision making* (pp. 103–112). Greenwich, CT: JAI Press.

Blackwell, H. R. (1963). Neural theories of simple visual discriminations. *Journal of the Optical Society of America, 53*, 129–160.

Bliss, C. I. (1967). *Statistics in biology* (Vol. I). New York: McGraw-Hill.

Carré, B. (1979). *Graphs and networks.* Oxford, England: Oxford University Press.

Clogg, Clifford C. (1981). New developments in latent structure analysis. In D. M. Jackson & E. F. Borgatta (Eds.), *Factor analysis and measurement in sociological research.* Beverly Hills, CA: Sage.

Comrey, A. L. (1962). The minimum residual method of factor analysis. *Psychological Reports, 11*, 15–18.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61*, 219–231.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics.* New York: Wiley.

Grofman, B., Feld, S. L., & Owen, G. (1984). Group size and the performance of a composite group majority: Statistical truths and empirical results. *Organizational Behavior and Human Performance, 33*, 350–359.

Guilford, J. P. (1954). *Psychometric methods* (2nd ed.). New York: McGraw-Hill.

Hogg, R. V., & Craig, A. T. (1978). *Introduction to mathematical statistics* (4th ed.). New York: Macmillan.

Lazarsfeld, P. F., & Henry, N.W. (1968). *Latent structure analysis.* New York: Houghton Mifflin.

Loéve, M. (1963). *Probability theory.* Princeton: Van Nostrand.

Lord, F. M. (1974). Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Luce, & P. Suppes (Eds.), *Contemporary developments in mathematical psychology* (Vol. 2). San Francisco: Freeman.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley.

Maher, K. M. (1987). A multiple choice model for aggregating group knowledge and estimating individual competencies. Unfinished doctoral dissertation, University of California, Irvine. (Available upon request.)

Nie, N. H., Hull, J. G., Jenkens, J. G., Steinbrenner, K., & Bent, D. M. (1975). *Statistical package for the social sciences.* New York: McGraw Hill.

Nitzan, S., & Paroush, J. (1982). Optimal decision rules in uncertain dichotomous choice situations. *International Economics Review, 23,* 289–297.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

Romney, A. K., Weller, S. C., & Batchelder, W. H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist, 88,* 313–338.

Manuscript received 10/1/85
Final version received 2/11/87