The Analysis of Multivariate Misclassified Data With Special Attention to Randomized Response Data

ARDO VAN DEN HOUT
PETER G. M. VAN DER HEIJDEN
Utrecht University, The Netherlands

This article discusses log-linear analysis of misclassified categorical data when conditional misclassification probabilities are known. This kind of misclassification occurs when data are collected using a randomized response design. The authors describe the misclassification by a latent class model. Since a latent class model is a log-linear model with one or more categorical latent variables, it is possible to investigate relations between misclassified variables. Methods to fit log-linear models for the latent table are discussed, including an EM algorithm. Attention is given to problems with boundary solutions. The results can also be used in statistical disclosure control when the post-randomization method is applied to protect the privacy of respondents, in epidemiology when specificity and sensitivity are known, and in data mining when privacy is protected by intentional statistical perturbation. Examples are given using randomized response data from a research into social benefit fraud.

Keywords: *latent class analysis; log-linear model; misclassification; randomized response; statistical disclosure control*

INTRODUCTION

Randomized response (RR) is an interview technique that can be used when sensitive questions have to be asked (Warner 1965; Chaudhuri and Mukerjee 1988). RR data can be seen as misclassified data where conditional misclassification probabilities are known. The main purpose of this article is to show how research questions concerning association patterns in multivariate RR data can be assessed using latent

AUTHORS' NOTE: *The authors would like to thank Jacques Hagenaars for the idea to describe randomized response designs by latent class models and Jeroen Vermunt for his advice with respect to* ℓEM.

SOCIOLOGICAL METHODS & RESEARCH, Vol. 32, No. 3, February 2004 384-410
DOI: 10.1177/0049124103257440

class models (Haberman 1979; Hagenaars 1993). Describing the RR design by a latent class model (LCM) is an advantage in practice since software to assess LCMs is widely available (e.g., the program ℓEM, Vermunt 1997).

In addition, this article considers problems with respect to boundary solutions in the log-linear models that we want to fit. As far as we know, these problems are not discussed in the literature. As will be shown by examples, boundary solutions can occur when analyzing RR data, and in that situation, one should take care regarding the formulation of the EM algorithm. We review the discussion in Kuha and Skinner (1997) and Chen (1989), in which the EM algorithm is used for log-linear analysis of misclassified data.

As an example, RR data concerning violations of regulations for social benefit are used. Sensitive items were binary: Respondents were asked whether they had violated certain regulations (Van Gils, Van der Heijden, and Rosebeek 2001).

The outline of the article is as follows. Section 2 introduces the RR design and misclassification designs that are closely related. Section 3 discusses the χ^2 test of independence and introduces some of the techniques that are used in the following sections. In Section 4, the RR design is described by a latent class model, and consequently log-linear models for RR data are presented and an example is given. Section 5 presents techniques for fitting the log-linear models to RR data. The likelihood is given, and the EM algorithm in the literature is discussed. In Section 6, attention is paid to boundary solutions and bias in RR data. Section 7 concludes.

THE RANDOMIZED RESPONSE DESIGN

The research by Van Gils et al. (2001) used the RR design introduced by Kuk (1990). In this design, there are two stacks of cards, each containing black and red cards. The proportion of red cards is 8/10 in the right stack and 2/10 in the left stack. The respondent is asked to draw one card from each stack and to keep the color of the cards hidden from the interviewer. Next, the sensitive question is asked. Instead of answering the question directly with yes or no, the respondent names the color of the card he took from the related stack (i.e., when the

answer is yes, the respondent names the color of the card he or she took from the right stack, and when the answer is no, he or she names the color of the card from the left stack).

RR data can be described as misclassified data. We associate violations with the color red. In this way, the probability to be correctly classified is 8/10 both for respondents who violated regulations and for those who did not. The RR matrix that contains the conditional misclassification probabilities

$$p_{ij} = I\!P(\text{category } i \text{ is observed}|\ \text{true category is } j) \tag{1}$$

is therefore given by

$$\boldsymbol{P} = \begin{pmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{pmatrix} = \begin{pmatrix} 8/10 & 2/10 \\ 2/10 & 8/10 \end{pmatrix}. \tag{2}$$

The main idea behind RR is that the perturbation induced by the misclassification design (in this case, the red and black cards) protects the privacy of the respondent and that insight into the misclassification design (in this case, the knowledge of the proportions red/black) can be used to analyze the observed data.

It is possible to create RR designs in which questions are asked to get information about a variable with $K > 2$ categories (see, e.g., Chaudhuri and Mukerjee 1988, chap. 3). The general form of the RR designs we discuss is

$$\boldsymbol{\theta}^* = \boldsymbol{P}\boldsymbol{\theta}, \tag{3}$$

where $\boldsymbol{\theta}^* = \left(\theta_1^*, \ldots, \theta_K^*\right)^t$ is a vector denoting the probabilities of the observed answers with categories $1, \ldots, K$; $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^t$ is the vector of the probabilities of the true answers; and $\boldsymbol{P}$ is the $K \times K$ transition matrix of conditional misclassification probabilities p_{ij}, as given in (1). Note that this means that the columns of $\boldsymbol{P}$ add up to 1. Further background and more complex randomized response designs can be found in Chaudhuri and Mukerjee (1988) and Fox and Tracy (1986).

Since we are dealing with the general form of misclassification as given in (3), the methods discussed in this article can also be used

in categorical data analysis when misclassification occurs and the probabilities given by (1) are known (e.g., instruments with known sensitivity and specificity in epidemiologic research) (see, e.g., Magder and Hughes 1997).

There is also a similarity between RR designs and the post randomization method (PRAM), introduced by Kooiman, Willenborg, and Gouweleeuw (1997) as a method for statistical disclosure control of data matrices. Statistical disclosure control aims at safeguarding the identity of respondents. Because of the privacy protection, data producers, such as national statistical institutes, are able to pass on data to a third party. The PRAM procedure yields a new data matrix in which the values of certain categorical variables in the original matrix may be misclassified into different values according to a given probability mechanism. In this way, PRAM introduces uncertainty in the data: The user of the data cannot be sure whether the information in the matrix is original or perturbed due to PRAM. As with RR, the misclassification scheme is given by means of a $K \times K$ transition matrix $\boldsymbol{P}$ of conditional probabilities p_{ij}, where

$$p_{ij} = I\!P(\text{category } i \text{ is released}|\text{true category is } j).$$

The role of the transition matrix in the analysis of PRAM data is the same as the role of the transition matrix in the analysis of RR data. More about PRAM and the similarity with RR can be found in Van den Hout and Van der Heijden (2002).

A third field that may benefit from results regarding misclassification with known misclassification probabilities is data mining. In this field, huge amounts of data are collected from surfers on the Web, and privacy concerns have initiated research into ways to protect the privacy of surfers by intentional statistical perturbation (see IBM scientists 2002).

Specific to the misclassification induced by RR is that it is nondifferential and independent. Let A and B denote two categorical variables, where A has I categories and B has J categories. Let A^* and B^* be the misclassified versions of A and B. Misclassification of A is called nondifferential with respect to B if

$$I\!P(A^* = k|\ A = i,\ B = j) = I\!P(A^* = k|\ A = i), \qquad (4)$$

where $k, i \in \{1, 2, \ldots, I\}$ and $j \in \{1, 2, \ldots, J\}$ (see Kuha and Skinner 1997). The notion of independence is used when there are more than two misclassified variables. The misclassification is independent if

$$\begin{aligned} &I\!P(A^* = k,\ B^* = l|\ A = i,\ B = j) \\ &\quad = I\!P(A^* = k|\ A = i,\ B = j) I\!P(B^* = l|\ A = i,\ B = j), \end{aligned} \tag{5}$$

where $k, i \in \{1, 2, \ldots, I\}$ and $l, j \in \{1, 2, \ldots, J\}$.

If $\boldsymbol{P}$ in (3) is nonsingular and we have an unbiased estimate $\widehat{\boldsymbol{\theta}}^*$ of $\boldsymbol{\theta}^*$, we can estimate $\boldsymbol{\theta}$ by the unbiased moment estimator

$$\widehat{\boldsymbol{\theta}} = \boldsymbol{P}^{-1}\widehat{\boldsymbol{\theta}}^* \tag{6}$$

(see Chaudhuri and Mukerjee 1988; Kuha and Skinner 1997). In practice, assuming that $\boldsymbol{P}$ in (3) is nonsingular does not impose much restriction on the choice of the misclassification design. Matrix $\boldsymbol{P}^{-1}$ exists when the diagonal of $\boldsymbol{P}$ dominates—that is, $p_{ii} > 1/2$ for $i \in \{1, \ldots, K\}$—and this is reasonable since these probabilities are the probabilities that the classification is correct.

Due to the fact that the misclassification in a RR design is independent and nondifferential, the generalization to an m-dimensional contingency table with $m > 1$ is straightforward. First, the m-dimensional contingency table is structured as an one-dimensional table of a compounded variable. For instance, when we have three binary variables, we obtain a one-dimensional table with rows indexed by 111, 112, 121, 122, 211, 212, 221, and 222 (the last index changes first). Second, due to the properties (4) and (5), it is possible to create the transition matrix of the compounded variable using the transition matrices of the underlying variables. Given the observed compounded variable and its transition matrix, we can use the moment estimator as described above.

CHI-SQUARE TEST OF INDEPENDENCE

This section discusses testing the independence between two categorical variables when one variable or both variables are subject to misclassification due to a RR design such as (3).

Consider the cross-tabulation of the variables A and B, which are defined in the previous section. Let $\pi_{ij} = I\!P(A = i, B = j)$ for each $i \in \{1, 2, \ldots, I\}$ and $j \in \{1, 2, \ldots, J\}$. The data are assumed to be distributed multinomially. The null hypothesis of independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, where the plus sign denotes summation over the related index (e.g., $\pi_{i+} = \pi_{i1} + \cdots + \pi_{iJ}$). In the standard situation without misclassification, the expected frequencies in the (i, j) cell under H_0 are estimated by $\widehat{m}_{ij} = n_{i+}n_{+j}/N$, where n_{ij} denotes the observed frequencies in the (i, j) cell of the cross-tabulation of A and B, and N is the sample size. The test statistic is the standard chi-square test of independence.

When one or two variables are misclassified and the misclassification is nondifferential and independent, the collapsibility theorem (Bishop, Fienberg, and Holland 1975) can be used to show that the standard chi-square test of independence can be applied to the observed table (see Korn 1981). As a result, when the misclassification is due to RR, and A^* and B^* denote the misclassified versions of A and B, it is possible to make inference about the independence between A and B by applying the chi-square test to the observed cross-classification of A^* and B^*. The test has the correct significance level, but power is reduced compared to the situation without misclassification. Several other authors discussed the chi-square test when one or more variables are misclassified—see, for example, Mote and Anderson (1965) and Assakul and Proctor (1967), who give attention to the reduction of power, and Rosenberg (1979).

An example shows how this works with RR data. In Table 1a, two variables are cross-classified that come from research into violating regulations of social benefit (Van Gils et al. 2001).

The variable G denotes gender. The observed red/black answers to the RR question are denoted by F^*. The question is whether the respondents earned money by doing some odd jobs without informing the office that provides their social benefit. This is a sensitive question since not informing the office is against regulations. Let the binary variable F denote the not-observed yes/no answers that we will call the true answers.

Applying the chi-square test to the observed values in Table 1a yields $X^2 = 3.377$ with 1 degree of freedom and a p value of .066.

TABLE 1: (a) Classification by Gender (G) and Randomized Response (RR) Answer (F^*) and (b) Estimated Classification by Gender (G) and True Answer (F)

(a)

G	F* *Red*	*Black*	*Total*
Male	218	500	718
Female	152	438	590
Total	370	938	1,308

(b)

G	F *Yes*	*No*	*Total*
Male	124.00	594.00	718
Female	56.67	533.33	590
Total	180.67	1127.33	1,308

When we choose a significance level of $\alpha = .05$, the data do not give a reason to reject the null hypothesis.

We now show that ignoring the results of Korn (1981) and taking the misclassification into account leads to the same value of X^2. Let $\boldsymbol{n}^* = (n^*_{11}, n^*_{12}, n^*_{21}, n^*_{22})^t$ denote the observed frequencies in Table 1a. To use the moment estimator (6), we first define the transition matrix $\boldsymbol{P}_{GF}$ of the compounded variable. Since the RR design to F is applied with matrix (2) and gender (G) is not perturbed, we obtain

$$\boldsymbol{P}_{GF} = \begin{pmatrix} 8/10 & 2/10 & 0 & 0 \\ 2/10 & 8/10 & 0 & 0 \\ 0 & 0 & 8/10 & 2/10 \\ 0 & 0 & 2/10 & 8/10 \end{pmatrix}. \tag{7}$$

This matrix is used to estimate frequencies $\widehat{\boldsymbol{n}} = (\widehat{n}_{11}, \widehat{n}_{12}, \widehat{n}_{21}, \widehat{n}_{22})^t$ in the classification by G and F by

$$\widehat{\boldsymbol{n}} = \boldsymbol{P}^{-1}_{GF}\boldsymbol{n}^*$$

(see Table 1b). Next, we estimate the expected frequencies in this table, denoted by $\widehat{\boldsymbol{m}} = (\widehat{m}_{11}, \widehat{m}_{12}, \widehat{m}_{21}, \widehat{m}_{22})^t$, under the model of

independence by $\widehat{m}_{ij} = \widehat{n}_{i+}\widehat{n}_{+j}/N$. Since we want to fit the model of independence, we compute the fitted frequencies under this model, denoted by $\widehat{\boldsymbol{m}}^*$, by

$$\widehat{\boldsymbol{m}}^* = \boldsymbol{P}_{GF}\widehat{\boldsymbol{m}}$$

and compare them with the observed $\boldsymbol{n}^*$. Again we get $X^2 = 3.377$.

When measuring the association between G and F by estimating the odds ratio, the misclassification should be taken into account explicitly. To show this, we will first ignore the misclassification and, second, give the adjusted estimate of the odds ratio.

Using only Table 1a to compute an estimate of the odds ratio η in the standard way yields $\widehat{\eta}^* = (218 \times 438)/(500 \times 152) = 1.26$. The large-sample standard error of $\log \widehat{\eta}^*$ is $(1/218 + 1/500 + 1/152 + 1/438)^{1/2} = 0.12$ (see Agresti 1990), so that $\widehat{\eta}^*$ has the 95 percent confidence interval (0.99, 1.61). This interval includes 1 and therefore does not justify rejecting the null hypothesis of independence. However, this estimate of η is biased toward 1 (see Magder and Hughes 1997) and is therefore not trustworthy.

An adjusted estimate can be deduced from Table 1b: $\widehat{\eta} = (124.00 \times 533.33)/(594.00 \times 56.67) = 1.96$. The logarithm of this estimate has a large-sample standard error of .40, so that $\widehat{\eta}$ has the 95 percent confidence interval (0.90, 4.29) (see Greenland 1988; van den Hout and van der Heijden 2002). As expected, the interval is larger than the interval of $\widehat{\eta}^*$ since the extra variance due to the RR design is taken into account. We see that, in accordance with the chi-square test to the observed values in Table 1a, the adjusted estimation of the odds ratio does not justify rejecting H_0.

THE LOG-LINEAR MODEL

This section discusses log-linear analysis when one or more categorical variables are observed using a RR design such as (3). This section can be seen as an extension to Section 3 since testing the log-linear model of independence for two variables is equal to the chi-square test of independence. First, we use the log-linear parameterization of the LCM to show that the RR design can be described by an LCM. Second, we give an example of log-linear analysis when one of the variables is an RR variable. The link between RR and LCMs is

useful since it turns out that widely available latent class software can be used to fit log-linear models that contain RR variables.

When one or more variables in the standard log-linear model concern observed values in a RR design, log-linear analysis using only the observed table may lead to wrong inference about the parameters. Consider, for instance, the variables G and F^* that are cross-classified in Table 1a. The standard saturated model (GF^*) to describe this table is given by

$$\log m_{gf^*} = \lambda_0 + \lambda_g^G + \lambda_{f^*}^{F^*} + \lambda_{gf^*}^{GF^*},$$

where m_{gf^*} denotes the expected frequency in the (g, f^*) cell, and $g, f^* \in \{1, 2\}$. The λ terms are restricted by

$$\sum_{g=1}^{2} \lambda_g^G = \sum_{f^*=1}^{2} \lambda_{f^*}^{F^*} = \sum_{g=1}^{2} \lambda_{gf^*}^{GF^*} = \sum_{f^*=1}^{2} \lambda_{gf^*}^{GF^*} = 0.$$

The estimate $\exp(4\widehat{\lambda}_{11}^{GF^*})$ is equal to the estimate of the odds ratio $\widehat{\eta}^* = 1.26$, as given in Section 3. It was already noted that this estimate is biased.

To apply log-linear models correctly, we should take into account the misclassification due to the RR design. In the standard application of LCMs, there are two kinds of variables: directly observed manifest variables and indirectly observed latent variables. The general idea is that the latent variables explain relationships among the manifest variables. Say we have one latent variable, X, and three manifest variables, S, T, and U. An important assumption in LCMs is local independence: Given the latent variable, manifest variables are independent. The log-linear parameterization of the LCM is therefore

$$\log m_{stux} = \lambda_0 + \lambda_s^S + \lambda_t^T + \lambda_u^U + \lambda_x^X + \lambda_{sx}^{SX} + \lambda_{tx}^{TX} + \lambda_{ux}^{UX}, \quad (8)$$

where the possible values x of the latent variable X and the number of categories of X are not known beforehand. An example of latent class analysis is the situation when the manifest variables concern attitudes toward political issues and the latent variable is binary and indicates political orientation (e.g., left wing vs. right wing). The idea here is that the latent variable explains dependencies between the attitudes. More about this example and the general LCM can be found in Hagenaars (1993).

The RR situation is rather different from the standard latent class situation. Say we have an observed red/black variable A^* that is the misclassified version of the yes/no variable A. The relation between the variables is one to one: Manifest variable A^* corresponds to latent variable A, and the assumption of local independence does not apply since there are no other manifest variables besides A^*. Furthermore, we do not have to investigate how many categories A has since the number is equal to the number of categories of A^*. The log-linear parameterization of this LCM is

$$\log m_{a^*a} = \lambda_0 + \lambda_{a^*}^{A^*} + \lambda_a^A + \lambda_{a^*a}^{A^*A}, \tag{9}$$

where $a^*, a \in \{1, 2\}$. An important property of (9) is that $\lambda_{a^*}^{A^*}$ and $\lambda_{a^*a}^{A^*A}$ are fixed since the conditional probabilities $I\!P(A^* = a^* | A = a)$ are fixed by the RR design. The relations between these terms and conditional probabilities are given in, for example, Heinen (1996, chap. 2); see also Section 5.

Once we have a log-linear parameterization of the RR design, we can add manifest variables that are not RR variables and investigate different log-linear models. We elaborate the social benefit example by considering, besides variables F and G, the categorical variable P, which denotes the population size of the place of residence and has five levels. Consider Table 2, which cross-classifies F^* with G and P, and the log-linear model (FGP, FF^*), given by

$$\begin{aligned}\log m_{f^*gpf} = {} & \lambda_0 + \lambda_{f^*}^{F^*} + \lambda_g^G + \lambda_p^P + \lambda_f^F \\ & + \lambda_{fg}^{FG} + \lambda_{fp}^{FP} + \lambda_{gp}^{GP} + \lambda_{f^*f}^{F^*F} + \lambda_{fgp}^{FGP},\end{aligned} \tag{10}$$

where $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$ are fixed by the RR design, and $f^*, g, f \in \{1, 2\}$, $p \in \{1, \ldots, 5\}$.

We call (10) the saturated model for the latent table FGP. In what follows, we will assess different log-linear models for the latent table FGP. All the models include the fixed terms $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$ since the RR design should always be taken into account.

The preceding discussion shows that we can use latent class software when this software allows for restrictions on conditional probabilities. The program ℓEM (Vermunt 1997) is an example of this kind of software. Since the LCMs that correspond to the RR design are very restricted, estimation of the models becomes less complex

TABLE 2: Classification by Randomized Response (RR) Answer (F^*), Gender (G), and Population Size of the Place of Residence (P)

		P ($\times 1,000$)				
F*	G	≥ 400	100–400	50–100	20–50	≤ 20
Red	Male	12	34	51	79	42
	Female	19	30	33	47	23
Black	Male	32	89	79	198	102
	Female	35	101	105	150	47

TABLE 3: Test Statistics

Model	df	X^2	p	L^2	p
1. (FG, FP, GP, FF^*)	4	6.78	0.15	6.70	0.15
2. (FG, GP, FF^*)	8	11.54	0.17	11.10	0.20
3. (FP, GP, FF^*)	5	10.34	0.07	10.39	0.06
4. (GP, FF^*)	9	14.85	0.10	14.49	0.11

when they describe RR data. Using ℓEM to estimate log-linear models for RR data is easy and fast. The code that we used for the example with variables F^*, G, and P is given in the appendix. Apart from the fixed interaction terms, the models for the RR data also differ from standard log-linear models because they concern an incomplete contingency table (i.e., in table $FGPF^*$, variable F is not observed). Because of this incompleteness, the EM algorithm (Dempster, Laird, and Rubin 1977) is applicable. The estimation of the models and the formulation of the EM algorithm are discussed in Section 5.

Since we are not interested in the relation between G and P, we only consider models that contain G and P jointly. In Table 3, the values of the test statistics of several models are given. Estimating the frequencies under the saturated log-linear model for the latent table FGP yields the estimates in Table 4. This table can also be estimated by using the unbiased moment estimator (6).

We partitioned the likelihood ratio goodness-of-fit statistic to find the best model (see Table 5), and this leads to model (GP, FF^*). This means that there is no convincing evidence for the dependence between F, on one hand, and G and P taken jointly, on the other

TABLE 4: Estimated Classification by True Answer (F), Gender (G), and Population Size of the Place of Residence (P)

		P ($\times$1000)				
F	G	$\geq$ 400	100–400	50–100	20–50	$\leq$ 20
Yes	Male	5.3	15.7	41.7	39.3	22.0
	Female	13.7	6.3	9.0	12.7	15.0
No	Male	38.7	107.3	88.3	237.7	122.0
	Female	40.3	124.7	129.0	184.3	55.0

TABLE 5: Hypothesis Test for Various Pairs of Nested Models in Table 3

Comparison	Δdf	ΔL^2	p
2 versus 1	4	4.40	0.36
4 versus 2	1	3.39	0.07
3 versus 1	1	3.69	0.05
4 versus 3	4	4.10	0.39

TABLE 6: Estimates of λ Parameters in Model (GP, FF^*) and Their Standard Errors

Parameter	*Estimate*	*Standard Error*	*Parameter*	*Estimate*	*Standard Error*
λ_1^F	–0.92	0.09	λ_4^P	0.72	0.05
λ_1^G	0.07	0.03	λ_{11}^{GP}	–0.18	0.08
λ_1^P	–0.85	0.08	λ_{12}^{GP}	–0.10	0.06
λ_2^P	0.11	0.06	λ_{13}^{GP}	-0.10	0.06
λ_3^P	0.16	0.06	λ_{14}^{GP}	0.10	0.05

hand. Estimated λ terms and their estimated standard errors for model (GP, FF^*) are given in Table 6.

The data indicate that not informing the social benefit office about money earned is independent of gender and population size of the place of residence taken together.

In the remainder of this section, we make some general remarks with respect to the log-linear models for RR data. Certain log-linear models for RR data can be tested when standard log-linear analysis is applied to observed variables, even when some of the variables are misclassified. Korn (1981) showed that a hierarchical model is

preserved by misclassification if the misclassified variable appears only once in the specification of the model. *Preserved* means that the misclassification will not change the fact that the observed table satisfies the model. When testing the model to the observed table, the same significance level is achieved, but power is reduced. An example is applying the chi-square to a two-dimensional table (see Section 2). Another example is the model (AB, BC) that is preserved under misclassification in A and in C but not under misclassification in B.

In practice, it often will be the case that we want to investigate log-linear models that do not meet the criterion formulated by Korn (1981), and we still need the adjustments described by the methods above. Also, even when a model is preserved, the estimation of the λ terms in the model should take the misclassification into account. In the example, only model (FG, FP, GP, FF^*) does not satisfy the assumptions of Korn.

Another important point is whether local maxima of the likelihood at hand are possible. In the standard hierarchical log-linear model, the likelihood function has a unique maximum when the solution is in the interior of the parameter space. Regarding the general LCM, it is known that it is possible that the likelihood function has local maxima (see Haberman 1979). However, the restricted LCM in this article that describes the RR variables seems to have different properties than the general LCM. In the example above and in other not reported log-linear analyses, we did not encounter local maxima of the likelihood functions. As an illustration, Figure 1 depicts the likelihood given by (12) of the independence model for a latent table FG denoted by (G, FF^*) and applied to Table 1a. As can be seen in Figure 1, the parameter space of the likelihood seems to have one maximum.

We obtain $(\widehat{\lambda}_1^G, \widehat{\lambda}_1^F) = (0.098, -0.92)$ as the point where the maximum is attained. Note the flatness of the likelihood in the direction of λ_1^F compared with the direction of λ_1^G. This flatness shows the extra variance due to the fact that F is a RR variable.

We conjecture that the log-linear models for RR data have a unique maximum if there is a solution in the interior of the parameter space. With respect to the saturated model for the latent table, this is true since the saturated model is just a reparametrization of the multinomial distribution, and van den Hout and van der Heijden (2002) prove that in that case, there is a unique maximum. The conjecture for more parsimonious models might be investigated using research concerning

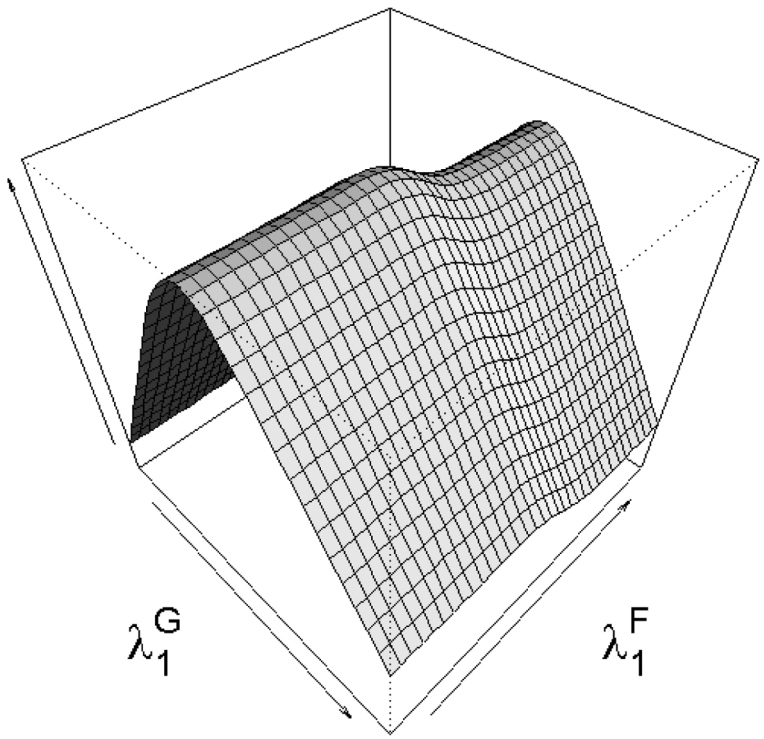

Figure 1: Likelihood of Model (G, FF^*)

product models (Haberman 1977) or research into marginal models (see, e.g., Bergsma and Rudas 2002). Both fields seem to address related problems. We hope to provide a decisive answer in future research.

ESTIMATING THE LOG-LINEAR MODEL

This section presents techniques for estimating the log-linear models for RR data discussed in Section 4. First, we specify the likelihood. Second, we discuss the EM algorithm that can be used to maximize the likelihood. For log-linear models with latent

variables, the algorithm was formulated by Haberman (1979). Both Chen (1989) and Kuha and Skinner (1997) use the algorithm in the situation of misclassification when misclassification probabilities are known, although the formulations of the algorithm differ. Chen explicitly discusses RR data. We will review the two formulations since the difference is important when a boundary solution is encountered. By a *boundary solution*, we mean an estimated expected cell frequency in the latent table that equals zero. Section 6 will give examples of RR data where boundary solutions occur.

To give the general formula of the likelihood, let the latent frequencies $\boldsymbol{n} = (n_1, \ldots, n_D)^t$ be multinomially distributed with parameters N and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)^t$. We specify log-linear models by $\eta_d = \log \theta_d$, $d \in \{1, \ldots, D\}$, and $\boldsymbol{\eta} = \boldsymbol{M}\boldsymbol{\lambda}$, where $\boldsymbol{M}$ is the $D \times r$ design matrix that defines the log-linear model, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^t$ is the parameter vector of the model. Ignoring constants, the likelihood is given by

$$L(\boldsymbol{\lambda}|\boldsymbol{n}^*) = \prod_{i=1}^{D} (\theta_i^*)^{n_i^*} = \exp\left\{ \sum_{i=1}^{D} n_i^* \log(p_{i1}e^{\eta_1} + \cdots + p_{iD}e^{\eta_D}) \right\}, \tag{11}$$

where $\boldsymbol{n}^*$ is the vector with observed frequencies, and p_{ij}, $i, j \in \{1, \ldots, D\}$, are the entries of the transition matrix that describe the misclassification with respect to $\boldsymbol{\theta}$. Since $\boldsymbol{n}$ is assumed to be multinomially distributed, $\boldsymbol{n}^*$ is also multinomially distributed due to the specific form of the transition matrix (see Van den Hout and Van der Heijden 2002, Section 5).

As an example, consider the likelihood of the independence model (G, F) applied to Table 1a, given by

$$L(\boldsymbol{\lambda}|\boldsymbol{n}^*) = \prod_{i=1}^{4} (\theta_i^*)^{n_i^*} = \exp\left\{ \sum_{i=1}^{4} n_i^* \log(p_{i1}e^{\eta_1} + \cdots + p_{i4}e^{\eta_4}) \right\}, \tag{12}$$

where $\boldsymbol{n}^*$ are the frequencies in Table 1a; p_{ij}, $i, j \in \{1, \ldots, 4\}$, are the entries in $\boldsymbol{P}_{GF}$ given by (7); and the nonredundant parameters

are λ_1^G and λ_1^F. Note that parameter λ_0 is not a free parameter in the log-linear model—in this case,

$$\lambda_0 = -\log \sum_{gf} \exp(\lambda_g^G + \lambda_f^F),$$

where $g, f \in \{1, 2\}$.

We can maximize (11) directly using the Newton-Raphson method, but we can also maximize it using an EM algorithm. The program ℓEM (Vermunt 1997) uses both procedures. The program starts with an EM algorithm and uses Newton-Raphson when close to the maximum. The applicability of the EM algorithm to RR data becomes clear when RR data are viewed as incomplete data. For each respondent, we can associate with the observed value of F^* the not-observed, nonperturbed value of F. Together, these pairs form an incomplete data matrix. (In the framework of Rubin [1976], the missing data are missing at random since they are missing by design.)

Next we review the formulations of the EM algorithm given by Chen (1989) and Kuha and Skinner (1997). We use the example in the preceding section and start with the formulation by Chen (1989). Say we want to fit model (FGP, FF^*). With $v = 0, 1, 2...$ denoting the cycles, the algorithm is given by

Initial estimate: $m_{fgp}^{(0)}$

E-step: $$n_{f^*gpf}^{(v)} = n_{f^*gp}\left(m_{fgp}^{(v)}\pi_{f^*|f}\right)\Big/\left(\sum_{t=1}^{2} m_{tgp}^{(v)}\pi_{f^*|t}\right)$$

M-step: Fit (FGP, FF^*) to $n_{f^*gpf}^{(v)}$ and use estimated expected frequencies to compute $m_{fgp}^{(v+1)}$,

where $m_{fgp}^{(0)}$ is the initial estimate of the frequencies in latent table FGP, n_{f^*gp} are the observed frequencies in the F^*GP table, and in each step $f^*, g, f \in \{1, 2\}$, $p \in \{1, \ldots, 5\}$. The conditional probabilities $\pi_{f^*|f}$ are fixed and provided for by the transition matrix for F given by (2).

To test the model after convergence, compare $m_{f^*gp+}^{(\infty)}$ with n_{f^*gp} using, for instance, the chi-square test or the likelihood ratio test. The degrees of freedom of the chi-square distributions of these test

statistics are the number of cells that are compared minus the number of parameters fitted. For model (FGP, FF^*), we have $20 - 20 = 0$ degrees of freedom since the $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$ are fixed due to the RR design.

Chen (1989) is not explicit with respect to the fixed λ terms in models such as (FGP, FF^*). We think that it is important to stress that when fitting a log-linear model in the M-step, one should check whether the restrictions due to the RR design are maintained. The relations between λ terms and conditional probabilities in our example are

$$\pi_{f^*|f} = \frac{\exp(\lambda_{f^*}^{F^*} + \lambda_{f^*f}^{F^*F})}{\sum_{f^*} \exp(\lambda_{f^*}^{F^*} + \lambda_{f^*f}^{F^*F})}, \tag{13}$$

where f^*, $f \in \{0, 1\}$, and the summation is over values $f^* \in \{0, 1\}$ (see, e.g., Heinen 1996, chap. 2). The relation (13) is the same for more parsimonious models for the latent table FGP. Using (13), we obtain

$$\begin{aligned}
\lambda_1^{F^*} &= -1/4\Big(\log \pi_{2|1} - \log \pi_{1|1} - \log \pi_{1|2} + \log \pi_{2|2}\Big) = 0, \\
\lambda_{11}^{F^*F} &= -1/4\Big(\log \pi_{2|1} - \log \pi_{1|1} + \log \pi_{1|2} - \log \pi_{2|2}\Big) \approx 0.693.
\end{aligned} \tag{14}$$

When fitting (FGP, FF^*) in the M-step in a standard way without maintaining the restrictions, $\widehat{\lambda}_1^{F^*}$ and $\widehat{\lambda}_{11}^{F^*F}$ converge to the fixed values of $\lambda_1^{F^*}$ and $\lambda_{11}^{F^*F}$. However, when there is a boundary solution, this may not be the case. In Section 6, we give examples of RR data with boundary solutions. When the restrictions are not maintained, estimated expected frequencies $m_{f^*gp+}^{(\infty)}$ are wrong; consequently, the value of the test statistic is wrong. This means that the formulation of the EM algorithm in Chen (1989) does not always yield the EM algorithm that we want.

To apply the EM algorithm correctly—that is, in such a way that it also yields the right estimates in the case of boundary solutions—there are two possible adjustments. We will explain these two adjustments and show that they are one and the same due to the collapsibility theorem. First, we can fit (FGP, FF^*) in the M-step using the fixed

values of $\lambda_{f^*}^{F^*}$ and $\lambda_{f^*f}^{F^*F}$, given by (14). The disadvantage is that this is not completely standard log-linear analysis since we should take care of these restrictions in estimating expected frequencies.

Second, we can use the E-step and M-step, as given in Kuha and Skinner (1997), who refer to Chen (1989) but nevertheless give a different formulation of the algorithm—namely,

E-step:
$$n_{f^*gpf}^{(v)} = n_{f^*gp}\left(m_{fgp}^{(v)}\pi_{f^*|f}\right)\Big/\left(\sum_{t=1}^{2} m_{tgp}^{(v)}\pi_{f^*|t}\right)$$
$$n_{fgp}^{(v)} = n_{+gpf}^{(v)}$$

M-step: Fit (FGP) to $n_{fgp}^{(v)}$ to obtain estimated expected frequencies $m_{fgp}^{(v+1)}$.

The advantage is that the M-step is standard. With respect to the testing of the model after convergence, Kuha and Skinner (1997) do not explicitly give a procedure, but we suggest doing the following. Structure the estimated frequencies $\widehat{m}_{fgp}$ in the three-dimensional latent table FGP as a one-dimensional table of a compounded variable—say, $\widehat{\boldsymbol{m}}$—and compute the fitted frequencies under this model, denoted by $\widehat{\boldsymbol{m}}^*$, by

$$\widehat{\boldsymbol{m}}^* = \boldsymbol{P}_{FGP}\widehat{\boldsymbol{m}}, \tag{15}$$

where $\boldsymbol{P}_{FGP}$ is the transition matrix of the compounded variable. Next, $\widehat{\boldsymbol{m}}^*$ can be compared with the observed $\boldsymbol{n}^*$.

The above adjustments yield one and the same EM algorithm. This follows from the applicability of the collapsibility theorem (see Bishop et al. 1975). The theorem states that the interaction between F, G, and P in model (FGP, FF^*) can be measured from the table of sums obtained by collapsing table F^*GPF over F^* (see Figure 2).

This is why, in the EM algorithm, we can collapse the estimated complete table in the E-step *before* we apply log-linear analysis in the M-step.

To test the fitted model after convergence of the EM algorithm formulated in Kuha and Skinner (1997), we can combine the estimated λ terms (i.e., the main effects and the interactions) of the latent table FGP with the fixed λ terms given by (14), compute λ_0, and estimate

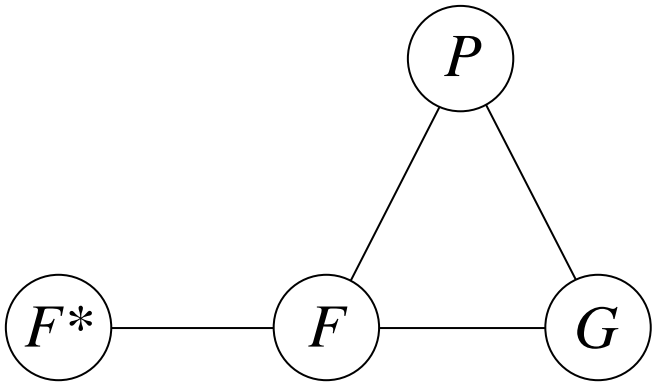

Figure 2: Saturated Model for the Latent Table *FGP*

the complete table F^*GPF to which the model (FGP, F^*F) exactly fits. However, since the information of the fixed λ terms is completely given by the transition matrix of F, we can also proceed after the EM algorithm, as described by (15). In this way, we can stay away from the estimation of λ terms and work with cell frequencies instead.

When the reader wants to implement the EM algorithm, we advocate using the EM algorithm in Kuha and Skinner (1997) and testing the models using (15). Note, however, that the EM algorithm does not yield estimated standard errors for the estimated λ terms. For estimated standard errors, one could use a method such as Newton-Raphson, as is done in ℓEM.

BOUNDARY SOLUTIONS

This section discusses boundary solutions that we encountered in the RR data concerning violations of regulations for social benefit (Van Gils et al. 2001). On the basis of these examples, a more general discussion is given with respect to boundary solutions in RR data and in PRAM data. This section generalizes the discussion of boundary solutions in van den Hout and van der Heijden (2002) to the situation with more than one variable.

A boundary solution is encountered when an estimated cell frequency in the latent table equals zero. This situation might occur

TABLE 7: (a) Classification by Randomized Response (RR) Answers F_1^* and F_2^* and (b) Estimated Classification by True Answers F_1 and F_2

(a)

F_1^*	F_2^* *Red*	*Black*	*Total*
Red	133	237	370
Black	147	791	938
Total	280	1,028	1,308

(b)

F_1	F_2 *Yes*	*No*	*Total*
Yes	107.21	66.22	173.43
No	0.00	1134.57	1134.57
Total	107.21	1200.79	1,308

when we combine several RR variables. From the research concerning violations of regulations for social benefit, we consider three binary RR variables—F_1^*, F_2^*, and F_3^*—with latent counterparts that are denoted by F_1, F_2, and F_3. Variable F_1^* is the same as F^* in Sections 3 and 4. Variable F_2^* denotes observed answers concerning the question of whether the respondents had a (temporary) legal job without informing the office that provides their social benefit. Variable F_3^* concerns the question of whether the respondent had an illegal job without informing the office. One transition matrix is used for each of the three variables and is given by (2).

As an example, consider Table 7a, which contains observed frequencies of RR variables F_1^* and F_2^*, and the estimated latent Table 7b, which contains estimated expected frequencies under model $(F_1F_2, F_1F_1^*, F_2F_2^*)$.

Testing the saturated model for the latent table F_1F_2 yields $X^2 = 18.67$ and $L^2 = 20.12$. If there had been no estimated zeroes in Table 7b, X^2 and L^2 would have been zero.

A second example is given by Table 8a, which contains observed frequencies, and Table 8b, which contains estimated expected frequencies under model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$.

TABLE 8: (a) Classification by Randomized Response (RR) Answers F_1^*, F_2^*, and F_3^* and (b) Estimated Classification by True Answers F_1, F_2, and F_3

(a)

F_1^*	F_2^*	F_3^* *Red*	*Black*
Red	Red	66	67
	Black	68	169
Black	Red	52	95
	Black	123	668

(b)

F_1	F_2	F_3 *Yes*	*No*
Yes	Yes	101.92	11.07
	No	18.56	45.38
No	Yes	0.00	0.00
	No	0.00	1131.06

The sample size is again 1308. Testing the saturated model for the latent table $F_1F_2F_3$ yields $X^2 = 38.53$ and $L^2 = 41.61$. It is clear from Table 8b that the questions are strongly related.

Boundary solutions might occur due to random error. When some of the latent frequencies are close to zero, an estimate of these frequencies after RR has been executed might result in a boundary solution. As an example, consider the binary latent variable H with latent frequencies $(98, 2)^t$ and assume that the transition matrix is given by (2). Possible frequencies of H^* due to the misclassification by the RR design are $(85, 15)^t$, and on the basis of these frequencies, the latent frequencies are estimated as $(100, 0)^t$—a boundary solution. When fitted frequencies are estimated by left-multiplying the transition matrix with $(100, 0)^t$, we get fitted frequencies $(80, 20)^t$ and $X^2 > 0$.

So, the fact that $X^2 > 0$ for the saturated model for the latent table is in itself not an indication that something is amiss. However, if the difference between zero and the X^2 of model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$ is large, it might be an indication that the perturbation of the latent frequencies is not due to

misclassification alone. This can be shown by two reasonings. First, a parametric bootstrap in which data are sampled from the estimated expected frequencies under the model can show that the large value of X^2 is unlikely when only misclassification is taken into account. We carried out such a bootstrap to investigate $X^2 = 38.53$ for the model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$ given the RR design. To describe the bootstrap, we switch to binary RR variables H_1^*, H_2^*, and H_3^*. From the estimated expected frequencies in the latent table $F_1F_2F_3$ under model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$, we sampled 100 tables $H_1H_2H_3$ and next simulated 100 tables $H_1^*H_2^*H_3^*$ using the RR design. For each of these tables $H_1^*H_2^*H_3^*$, the test statistic X^2 is computed. The mean value of the 100 simulated X^2 is 2.00, and the maximum is 9.25. This maximum is not even close to the $X^2 = 38.53$ of model $(F_1F_2F_3, F_1F_1^*, F_2F_2^*, F_3F_3^*)$. (Results for the likelihood ratio test L^2 are very similar.)

A second way to investigate whether the RR data can be described by misclassification alone is the following. Continuing with the example: Assume that none of the respondents committed fraud or, in other words, that the latent score is 2 for each variable F_1, F_2, and F_3. Under this assumption, the expected number of respondents with observed score "black" for each variable F_1^*, F_2^*, and F_3^* is $(8/10)^3 \times 1308 = 669.70$. The observed number in the survey is 668. This might suggest that the assumption is correct and that there are no frauds at all in the survey. However, this is contradicted by the 66 respondents who have the score "red" for each variable F_1^*, F_2^*, and F_3^*—a frequency that is much higher than the expected $(2/10)^3 \times 1308 = 10.46$ under the assumption of no fraud at all.

The two reasonings show that, given the RR design in the research and the estimated expected frequencies in Table 8b, it is rather unlikely that Table 8a is the observed table. However, since the observed data are the starting point of statistical inference, we should state the conclusion the other way around: Given Table 8a and the RR design, Table 8b is probably not a good estimate of the latent frequencies. The cause for this estimation problem is probably that some respondents do not always follow the RR design and answer "black" too often, irrespective of the question asked. A reason for this might be that some respondents do not trust the privacy protection offered by the RR design and answer "black" since "black" is associated with

no. These respondents bring about a second perturbation of the latent frequencies besides the misclassification due to the RR design. In the conclusion, we will return to this problem.

To make the discussion more general, note that when the statistical disclosure method PRAM is applied (see Section 1), $X^2 > 0$ for the saturated model for the latent table can only occur due to random error since the misclassification is executed by the computer. Also, in the case when RR data are not unlikely in the sense as discussed above, a method is needed to deal with the fact that X^2 might be unequal to zero. In the related field of incomplete data, it also may occur that $X^2 > 0$ (see Schafer 1997). Therefore, we suggest following Schafer (1997), who proposes taking the deviation from the null as a baseline for assessing nonsaturated models and defines an adjusted test statistic

$$X^2_{adj} = X^2 - X^2_0,$$

where X^2_0 is the X^2 of the saturated model. The likelihood ratio test is adjusted in the same way. The behavior of this adjusted test might be studied using Gelman, Meng, and Stern (1996), who use Bayesian analysis to assess model fit in situations when $X^2 > 0$ but when X^2 is expected to be zero if the model is true.

CONCLUSION

This article discusses log-linear analysis of randomized response data. It is shown that this kind of analysis can be executed using existing latent class software.

The RR data from the example in this article are difficult data. The problem is not the theoretic misclassification due to the RR design since this article shows that we can handle this misclassification. The problem is that some respondents do not follow the RR design and—of course—that we cannot identify these "cheaters." To some extent, we can use X^2_0 (i.e., the test statistic for the saturated model for the latent table) as a measure of the bias of the RR data (see Section 6), but it does not provide a decisive answer.

Asking sensitive questions will always produce incomplete or biased data. So one must make do with what one has got, and our idea

is that RR performs relatively well (see van der Heijden et al. 2000). Analysis of RR data in the future might profit from research into more methodological aspects of RR designs (see Boeije and Lensvelt-Mulders 2002, who discuss cheating in RR designs). A possible form of cheating is when a respondent answers "black," irrespective of the question asked since he or she does not trust the privacy protection. When respondents understand the privacy protection offered by the RR design better, data might be less biased. Another approach might be to add extra parameters to the model to describe cheating behavior. This is not straightforward since it is difficult to model cheating behavior, and we also might run into identifiability problems (see, e.g., Goodman 1974).

The question of how the bias in the data influences the analysis is difficult to answer. It is obvious that results should be interpreted with care and that cross-classifying several RR questions might increase the bias of the results. We suggest the following: When the saturated model fits perfectly (i.e., $X_0^2 = 0$), we advocate log-linear modeling, as described in Section 4. When $X_0^2 > 0$, one should be more careful, and the reasonings used in Section 6 can be used to assess bias in the RR data due to the "cheaters." When the parametric bootstrap in Section 6 makes the value of X_0^2 unlikely, it is unclear how to interpret the results of the log-linear analysis. When X_0^2 is not too large (i.e., the deviation from the null can be explained by random error), an adjusted test statistic can be used to test the fitting of log-linear models.

APPENDIX

We present two input files that can be used in ℓEM to fit the models in Section 4. The program ℓEM and the manual can be downloaded for free from www.kub.nl/faculteiten/fsw/ organisatie/ departementen/mto/software2.html. The "*" symbol denotes comment. By removing and adding this symbol in the text, different models can be fitted. The first input fits the saturated model for the latent table *FGP* and follows the log-linear parameterization of the LCM.

```
lat 1          * 1 latent variable
man 3          * 3 manifest variables
dim 2 2 2 5    * dimensions of variables
```

```
lab F R G P       * Labels:
                  * F = fraud
                  * R = observed RR answer
                  * G = gender
                  * P = pop. size of place of residence

  mod FGP {FGP,wei(FR)}              * Sat. model for table FGP with weighted interaction FR
  *mod FGP {FG,FP,PG,wei(FR)}        * no 3-way interaction model
  *mod FGP {FG,PG,wei(FR)}           * conditional independence
  *mod FGP {FP,GP,wei(FR)}           * conditional independence
  *mod FGP {F, GP,wei(FR)}           * model of joint independence
  sta wei(FR) [.8 .2 .2 .8]          * misclassification probabilities determine weights
  dat [12 34 51 79 42 19 30 33 47 23 * observed data
  32 89 79 198 102 35 101 105 150 47]
```

The second input also fits the saturated model for the latent table *FGP* but follows the log-linear modified path model parameterization (Goodman 1973; Hagenaars 1993:15). We only give the input for the saturated model for the latent table *FGP*, but restrictive models can easily be formulated.

```
lat 1             * 1 latent variable
man 3             * 3 manifest variables
dim 2 2 2 5       * dimensions of variables
lab F R G P       * Labels

  mod FGP {FGP}                      * Saturated model
  R|F {wei(RF)}                      * specifying weights
  sta wei(RF) [.8 .2 .2 .8]          * using misclassification probabilities
  dat [12 34 51 79 42 19 30 33 47 23 * observed data
  32 89 79 198 102 35 101 105 150 47]
```

REFERENCES

Agresti, Alan. 1990. *Categorical Data Analysis*. New York: John Wiley.

Assakul, Kwanchai and Charles H. Proctor. 1967. "Testing Independence in Two-Way Contingency Tables With Data Subject to Misclassification." *Psychometrika* 32:67-76.

Bergsma, Wicher P. and Tamas Rudas. 2002. "Marginal Models for Categorical Data." *The Annals of Statistics* 30:140-59.

Bishop, Yvonne M. M., Stephen E. Fienberg, and Paul W. Holland. 1975. *Discrete Multivariate Analysis: Theory and Practice*. Cambridge, MA: MIT Press.

Boeije, Hennie and Gerti Lensvelt-Mulders. 2002. "Honest by Chance: A Qualitative Interview Study to Clarify Respondents' (Non-)Compliance With Computer-Assisted Randomized Response." *Bulletin de Methodologie Sociologique* 75:24-39.

Chaudhuri, Arijit and Rahul Mukerjee. 1988. *Randomized Response: Theory and Techniques*. New York: Marcel Dekker.

Chen, T. Timothy. 1989. "A Review of Methods for Misclassified Categorical Data in Epidemiology." *Statistics in Medicine* 8:1095-1106.

Dempster, Arthur P., Nan M. Laird, and Donald B. Rubin. 1977. "Maximum Likelihood From Incomplete Data Via the EM Algorithm." *Journal of the Royal Statistical Society Series B* 39:1-38.

Fox, James A. and Paul E. Tracy. 1986. *Randomized Response: A Method for Sensitive Surveys*. Beverly Hills, CA: Sage.

Gelman, Andrew, Xiao-Li Meng, and Hal Stern. 1996. "Posterior Predictive Assessment of Model Fitness Via Realized Discrepancies." *Statistica Sinica* 6:733-807.

Goodman, Leo A. 1973. "The Analysis of Multidimensional Contingency Tables When Some Variables Are Posterior to Others: A Modified Path Analysis Approach." *Biometrika* 60:179-92.

———. 1974. "Explanatory Latent-Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215-231.

Greenland, Sander. 1988. "Variance Estimation for Epidemiologic Effect Estimates Under Misclassification." *Statistics in Medicine* 7:745-57.

Haberman, Shelby J. 1977. "Product Models for Frequency Tables Involving Indirect Observation." *The Annals of Statistics* 5:1124-47.

———. 1979. *Analysis of Qualitative Data: New Developments*. Vol. 2. New York: Academic Press.

Hagenaars, Jacques A. 1993. *Loglinear Models With Latent Variables*. Newbury Park, CA: Sage.

Heinen, Ton. 1996. *Latent Class and Discrete Latent Trait Models: Similarities and Differences*. Thousand Oaks, CA: Sage.

"IBM Scientists Rely on the Principle of Uncertainty to Develop Web-Privacy Answers." 2002. Retrieved from www.916.ibm.com/press

Kooiman, Peter, Leon C. R. J. Willenborg, and José M. Gouweleeuw. 1997. "PRAM: A Method for Disclosure Limitation of Microdata." Research paper no. 9705, Statistics Netherlands, Voorburg/Heerlen.

Korn, Edward L. 1981. "Hierarchical Log-Linear Models Not Preserved by Classification Error." *Journal of the American Statistical Association* 76:110-13.

Kuha, Jouni and Chris Skinner. 1997. "Categorical Data Analysis and Misclassification." In *Survey Measurement and Process Quality*, edited by Lyberg L., Biemer P., Collins M., de Leeuw E., Dippo C., Schwartz N., and Trewin D. New York: John Wiley.

Kuk, Anthony Y. C. 1990. "Asking Sensitive Questions Indirectly." *Biometrika* 77:436-38.

Magder, Laurence S. and James P. Hughes. 1997. "Logistic Regression When the Outcome Is Measured With Uncertainty." *American Journal of Epidemiology* 146:195-203.

Mote, V. L. and R. L. Anderson. 1965. "An Investigation of the Effect of Misclassification on the Properties of χ^2-Tests in the Analysis of Categorical Data." *Biometrika* 52:95-109.

Rosenberg, Martin J. 1979. "Multivariate Analysis by a Randomized Response Technique for Statistical Disclosure Control." Ph.D. dissertation, University of Michigan, Ann Arbor.

Rubin, Donald B. 1976. "Inference and Missing Data." *Biometrika* 63:581-92.

Schafer, Joseph L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.

van den Hout, Ardo and Peter G. M. van der Heijden. 2002. "Randomized Response, Statistical Disclosure Control and Misclassification: A Review." *International Statistical Review* 70:269-88.

van der Heijden, Peter G. M., Ger Van Gils, Jan Bouts, and Joop J. Hox. 2000. "A Comparison of Randomized Response, Computer-Assisted Self-Interview, and Face-to-Face Direct Questioning." *Sociological Methods & Research* 28:505-37.

Van Gils, Ger, Peter G. M. Van der Heijden, and Annemarie Rosebeek. 2001. *Onderzoek naar regelovertreding, Resultaten ABW, WAO en WW*. Amsterdam: NIPO (in Dutch).
Vermunt, Jeroen K. 1997. *LEM: A General Program for the Analysis of Categorical Data: User's Manual*. Tilburg: Tilburg University.
Warner, Stanley L. 1965. "Randomized Response: A Survey Technique for Eliminating Answer Bias." *Journal of the American Statistical Association* 60:63-69.

Ardo van den Hout is a Ph.D. student in the Department of Methodology and Statistics of the Faculty of Social Sciences, Utrecht University. He is interested in statistical models for the analysis of categorical data, particularly the analysis of randomized response data. He is also doing research into the field of statistical disclosure control.

Peter G. M. van der Heijden is a professor of statistics in the Department of Methodology and Statistics of the Faculty of Social Sciences, Utrecht University. He is interested in statistical models for the analysis of categorical data, particularly the analysis of randomized response data. He is also doing research into models for the estimation of the size of populations, such as capture-recapture models.