Conclusions about changes in categorical characteristics based on observed panel data can be incorrect when (even a small amount of) measurement error is present. Random measurement errors, referred to as independent classification errors, usually lead to overestimation of the total amount of gross change, whereas systematic, correlated errors usually cause underestimation of the transitions. Furthermore, the patterns of true change may be seriously distorted by independent or systematic classification errors. Latent class models and directed log-linear analysis are excellent tools to correct for both independent and correlated measurement errors. An extensive example on labor market states taken from the Survey of Income and Program Participation panel is presented.

Estimating True Changes When Categorical Panel Data Are Affected by Uncorrelated and Correlated Classification Errors

An Application to Unemployment Data

FRANCESCA BASSI
University of Padova
JACQUES A. HAGENAARS
MARCEL A. CROON
JEROEN K. VERMUNT
Tilburg University

It has long been recognized that turnover tables showing the transitions between discrete states provide important, basic tools for understanding processes of social change (Lazarsfeld and Rosenberg 1955, sec. 3; Plewis 1985). At the same time, it is well known that even small amounts of measurement error may result in distorted turnover tables and misleading conclusions about the changes that are taking place (Maccoby 1956; Hagenaars 1990, 1994). The main purpose of this article is to show how to find the true changes and analyze transition data when the data are affected by random, but especially by cor-

AUTHORS' NOTE: *Thanks are due to the anonymous reviewers for important suggestions and helpful comments on previous drafts of this article.*

SOCIOLOGICAL METHODS & RESEARCH, Vol. 29 No. 2, November 2000 230-268

related, systematic measurement errors. Only categorical variables will be dealt with here, and, accordingly, measurement errors will be denoted as classification errors or misclassifications. By way of example, data concerning labor flows will be used in which the transitions between labor market states, characterized as employed (E), unemployed (U), and not in the labor force (N), are observed at successive points in time.

Data to estimate gross flows (gross changes) can in principle be obtained in two ways: (1) longitudinal (panel) surveys, in which the same respondents are interviewed on successive occasions during each of which information about the respondent's current labor status is obtained, and (2) retrospective surveys, in which at one moment in time information is gathered about the respondent's labor status in the past. In practice, all kinds of mixtures of these two basic approaches occur. With longitudinal (panel) data, the classification errors for the successive occasions are usually only moderately or not at all correlated with one another given that the moments of observation are not too close to each other in time. These (nearly) independent measurement errors usually attenuate the associations between the variables and lead to spurious observed transitions and overestimation of the amounts of gross changes in the labor market. In retrospective surveys, on the other hand, classification errors are usually of a systematic nature and often lead to underestimation of the turnover, since respondents tend to be consistent in their answers and to forget about past changes in their labor market status (Kalton and Citro 1994; O'Muircheartaigh 1996).

In most classical methods for correcting for measurement errors and estimating the true gross flows, it is assumed that the measurement errors are independent of one another; that is, the independent classification error (ICE) assumption is made (Biemer and Trewin 1997; Kuha and Skinner 1997). More specifically, it is assumed that errors referring to two different occasions are independent of each other conditional on the true (labor market) states, and that errors depend only on the present true state, not on what has happened in the past, nor on what has been observed before. One group of classical methods compares the survey data with a gold standard, that is, with data that are considered to be (almost) perfectly valid. Such standards may be obtained from administrative sources or from specifically

arranged reinterviews. If validation data are not available, which is the rule rather than the exception, the role of the gold standard is taken over by an ICE model in which explicit assumptions are made about the error structure and the nature of the true transition processes (Abowd and Zellner 1985; Poterba and Summers 1986; Chua and Fuller 1987; Sutcliffe 1965a, 1965b). To be useful in empirical research, at least a portion of the model assumptions should be empirically testable. A powerful model in this respect is Lazarsfeld's latent class model, either in its standard form (Lazarsfeld and Henry 1968; Goodman 1974a, 1974b) or in the form of a latent Markov model (Van de Pol and Langeheine 1990; Collins and Wugalter 1992).

Unfortunately, there is much empirical evidence showing that the ICE assumption very often does not hold, especially in retrospective surveys, and that estimation procedures based on the ICE assumption often yield quite misleading and unrealistic descriptions of labor market dynamics (Lemaitre 1988; Skinner and Torelli 1993). In this article, strategies are proposed to correct gross flows estimates when the data are subjected to correlated classification errors. These strategies are based on a reformulation of the latent class model as a log-linear model with latent variables (Haberman 1979, chap. 10), more specifically as a causal directed log-linear model with latent variables (Hagenaars 1988, 1990, 1993, 1998; Vermunt 1996, 1997b; see also Goodman 1973; Whittaker 1990; Lauritzen 1996).[1]

The labor market data that serve as our example are from the 1986 panel of the Survey of Income and Program Participation (SIPP), one of the major longitudinal labor surveys in the United States. To understand the analyses to come, it is necessary to have some basic knowledge of the design of this study.

SIPP: A BRIEF DESCRIPTION

The SIPP is a multipanel survey of the U.S. noninstitutional population age 15 and older. Its main aim is to collect information on income, program participation, labor force activity, and household composition (U.S. Department of Commerce 1991; Citro and Kalton 1993). It was started in 1984 by the U.S. Bureau of the Census, and each year a

new panel of about 20,000 individuals is included in the survey. The respondents are contacted on a personal basis; telephone and proxy interviews are avoided as much as possible. The SIPP panel of respondents is divided into four rotation groups. A particular rotation group is interviewed at the beginning of every fifth month; every month, one and just one of the four groups is being interviewed. During the interview, the respondents are asked to provide information on the topics mentioned above for the reference period, that is, for the four calendar months preceding the interview. This basic design is repeated eight times. As a consequence of this particular rotation design, the SIPP has characteristics of both a retrospective study and a longitudinal study.

In the Labor Force and Recipiency section of the SIPP questionnaire, each respondent is asked to report on a weekly basis about his or her own labor market status during the past four calendar months. The respondents are asked first whether they had a job or a business at any point in time during the preceding four months. If they give a negative answer, respondents are asked whether they spent any time looking for work or were in layoff and, if so, in which weeks. If their first answer is positive, they are asked whether they worked for the entire reference period (all 18 weeks). If they report they worked for a shorter period, a long series of questions starts: Respondents have to indicate exactly in which weeks they had a job and in which weeks they were in layoff or looking for a job. Moreover, they have to tell for any of the weeks they had a job or were in a business, whether they were absent without pay from work and, if so, why they were absent. The weekly information is usually recoded to obtain a monthly classification with the three categories mentioned above: employed (E), unemployed (U), and not in the labor force (N). If a respondent belongs to different states in a single month, we follow Martini (1989), who classified the respondents in each month according to their modal position in the labor market, taking into consideration all four or five weeks of the month.

Because of its detailed questionnaire, the SIPP provides one of the best data sources on U.S. labor market dynamics. At the same time the task for the respondents is quite formidable, and there are theoretical reasons and empirical indications for assuming that SIPP data are affected by correlated response errors. To discuss the nature of these classification errors, some terminology must be introduced. The moment

a particular rotation group is being interviewed will be called a wave. As remarked above, during that wave, the rotation group concerned provides information about its labor market behavior during the four previous months (the reference period). After four months, a new wave takes place for this rotation group. So, observed labor market transitions between any two months will be based on information either from the same wave or from two different waves. Month-to-month transitions observed retrospectively within the same wave will be called within-wave transitions, whereas transitions observed on the basis of information gathered at two different waves of the same rotation group will be referred to as between-wave transitions. With regard to the within-wave transitions, it is likely that, going backward in time during one wave, the errors for each weekly report will be systematic and correlated due to all kinds of conditioning effects (Martini 1988; Kasprzyk et al. 1989, pt. 6). Failing memory will cause respondents to forget about spells of employment or unemployment, or misplace them in time. "Laziness" combined with the SIPP questionnaire structure may cause respondents to report a stable situation across all four months, or misplace changes of state toward the beginning or the end of the reference period. The status reported for the most recent week may be mechanically repeated for the entire reference period, or the misclassifications for one particular week may carry over to the next week. The true state at the beginning of the reference period, that is, the moment the interview takes place, may influence all answers for the whole reference period.

There are several empirical indications in SIPP data for the actual presence of these kinds of correlated misclassifications and conditioning effects. Although a true gold standard is missing, based on what is known from other sources the within-wave SIPP data seem to underestimate the gross changes in labor status. Furthermore, for a particular rotation group, as a rule, the within-wave transitions show more stability than the between-wave transitions. There is also the tendency that the within-wave stability between the two consecutive months at the beginning of the reference period, which are the farthest away from the moment of interviewing, is larger than the stability between the two months at the end of the reference period, which are the closest to the moment of interviewing. Finally, there is the notorious and well-documented phenomenon called the seam effect (Young 1989;

Burkhead and Coder 1985). If a turnover table is constructed for any two particular successive months for all four rotation groups, then, because of the typical structure of the SIPP, for three of the rotation groups the information is based on within-wave transitions and for one group on between-wave transitions. Now, the seam effect is called the phenomenon that the amount of gross changes for two particular successive months is far less when estimated on the basis of within-wave transitions than when based on between-wave transitions.

The data in Table 1 illustrate these tendencies. The data refer to the SIPP 1986 panel, which was started in February 1986 (asking questions to the first rotation group about the period from October 1985 to January 1986) and ended in August 1988. We considered the period from January 1986 to December 1987, in which we have the information for all four rotation groups. Row ALL contains the average observed transition rates for all two consecutive months during that period based on the complete sample. The last row in the table, row 41, contains the average between-wave transition rates for all four groups during the same period, that is, the average transition rates at the seam between the last month of a particular reference period and the first month of the next reference period. Rows 12, 23, and 34 contain the average within-wave transition rates during the same period and for all four groups. The 12 transitions are the average transitions occurring between the first two months of the reference periods, whereas 34 transitions refer to the more recent average transitions between the third and fourth month of the reference period (with obvious meaning for the remaining row 23). The tendencies that are visible in Table 1 and have been described above indicate that the data are not error free, and that the classification errors are systematic and correlated with each other.

In the following sections, we will show how one may deal with these kinds of longitudinal data when they can be assumed to be error free or only affected by independent classification errors. The (Markov) models described will serve as baseline models for analyses that take the possible systematic nature of the misclassifications into account. From application to SIPP data, many practical difficulties for carrying out these kinds of analyses become visible. Because these practical difficulties are in no way unique for SIPP data, they will be given explicit attention.

TABLE 1: Survey of Income and Program Participation Observed Gross Flows (average monthly transitions, percentages) (January 1986 to December 1987)

	EE	*EU*	*EN*	*UE*	*UU*	*UN*	*NE*	*NU*	*NN*
All	97.04	1.31	1.62	20.27	67.65	12.11	2.41	2.13	95.46
12	98.27	1.04	0.69	15.46	79.63	4.91	1.15	1.42	97.43
23	97.41	1.13	0.96	17.37	75.96	6.70	1.38	1.71	96.91
34	97.85	1.20	0.95	19.23	73.25	7.52	1.28	1.69	97.07
41	94.04	2.10	3.87	26.81	42.20	30.99	5.65	3.77	90.58

NOTE: E = employed, U = unemployed, N = not in the labor force.

MANIFEST AND LATENT MARKOV MODELS

Markov chains have been widely used in the analyses of (labor) turnover tables. Figure 1 represents the basic causal model that is relevant here, where A, B, C, and D denote respondents' labor market position in four consecutive months. Figure 1 represents a directed graph in which the arrows indicate direct effects from one variable to another controlling for the appropriate antecedent and intervening variables (Lauritzen 1996; Cox and Wermuth 1996). Given the causal order of the variables in Figure 1 and following the principles of Goodman's (1973) modified path approach, the probability π^{ABCD}_{abcd}, denoting the joint probability of belonging to labor market states a, b, c, and d on variables A, B, C, and D, with the subscripts a, b, c, and d varying over the set of labor market states {E, U, N}, may be decomposed as follows:

$$\pi^{ABCD}_{abcd} = \pi^{A}_{a}\pi^{B|A}_{ba}\pi^{C|AB}_{cab}\pi^{D|ABC}_{dabc}, \tag{1}$$

where π^{A}_{a} indicates the probability of being observed in category a of A, $\pi^{B|A}_{ba}$ is the conditional probability of scoring b on B provided that one belongs to category a of A, and the other symbols have similar and obvious meanings. Furthermore, all probabilities are subjected to the usual restrictions: Their lower bounds are zero, their upper bounds are one, and they sum to one where appropriate. In equation (1), the score on each successive variable depends in principle on all variables that are causally prior to the variable concerned. To evaluate the complete modified path model, the appropriate logit or (the equivalent) log-linear models are defined in agreement with the investigator's hypo-

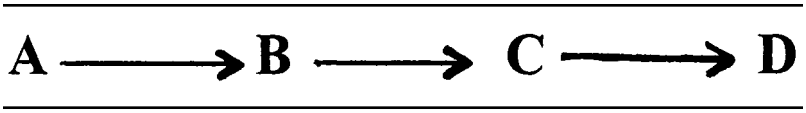


Figure 1: First-Order Markov Chain for Four Measurements Over Time

theses for each of the (marginal) tables corresponding to the conditional probabilities on the right-hand side of equation (1). The resulting estimates of the right-hand side parameters are used to obtain the estimate for the joint probability π_{abcd}^{ABCD} on the left-hand side of equation (1).[2]

Although leading to exactly the same results, it may be insightful to first simplify equation (1) before defining the appropriate log-linear (sub)models (Lauritzen 1996; Vermunt 1997b; Hagenaars 1998). If the model in Figure 1 is true in the population, each variable is influenced only by its causally immediately preceding variable, which leads to certain conditional independence relations between the variables: Variable C is conditionally independent of A given B, variable D is conditionally independent of A given B or C, and variable D is conditionally independent of B given C. Therefore, if the model in Figure 1 is true, equation (1) can be simplified as follows:

$$\pi_{abcd}^{ABCD} = \pi_{a}^{A} \pi_{ba}^{B|A} \pi_{cb}^{C|B} \pi_{dc}^{D|C} . \tag{2}$$

Figure 1 and equation (2) represent a first-order, nonstationary Markov chain in which there are only effects from time point t to time $t + 1$. Second-order chains may be defined by adding direct effects from t to $t + 2$, and so on (and replacing in equation [2] $\pi_{cb}^{C|B}$ by $\pi_{cab}^{C|AB}$ and $\pi_{dc}^{D|C}$ by $\pi_{dbc}^{D|BC}$). For first- or higher order, nonstationary Markov chains, standard methods provide the maximum likelihood estimates $\hat{\pi}_{abcd}^{ABCD}$ that can be compared with the observed proportions p_{abcd}^{ABCD} in the usual way by means of chi-square statistics to test the model assumptions (Goodman 1973; Lauritzen 1996).

In terms of Goodman's modified path approach or directed log-linear modeling, the first-order, nonstationary Markov chain amounts to estimating a saturated log-linear (logit) model for each of the elements on the right-hand side of equation (2). Employing the usual shorthand notation for hierarchical log-linear models, model $\{A\}$ is applied to

observed marginal table A to estimate π_a^A, model $\{AB\}$ is applied to observed marginal table AB to obtain estimates for $\pi_{ba}^{B|A}$, and so on. It is of course also possible to define nonsaturated, restricted log-linear models for one or more of these (marginal) tables, for example, symmetry or quasi-independence models. The Markov chain can also be made stationary by equating the transition tables A-B, B-C, and C-D, that is, $\pi_{ji}^{B|A} = \pi_{ji}^{C|B} = \pi_{ji}^{D|C}$ (with or without the restriction of dynamic equilibrium $\pi_i^A = \pi_i^B = \pi_i^C = \pi_i^D$). Introducing the stationarity assumption yields a model that is strictly speaking no longer a log-linear model, since it requires defining restrictions for sums of frequencies, in this case the equality of particular marginal one- or two-variable distributions. Such restrictions can be handled by the marginal modeling approach (Bishop, Fienberg, and Holland 1975, chap. 7; Vermunt 1997b:43-45; Lang and Agresti 1994; Becker and Yang 1998; Bergsma 1997; Diggle, Liang, and Zeger 1996).

Because in our example data from several (rotation) groups are available, an extra grouping variable G must be introduced, with subscript g running from one through four (since there are four rotation groups):

$$\pi_{gabcd}^{GABCD} = \pi_g^G \pi_{ag}^{A|G} \pi_{bga}^{B|GA} \pi_{cgb}^{C|GB} \pi_{dgc}^{D|GC}. \tag{3}$$

By defining the appropriate log-linear models for the elements on the right-hand side of equation (3), completely homogeneous models might be defined in which the Markov chain model is completely identical for all groups, or completely heterogeneous models might be defined in which all (Markov) parameters are different for all groups. And, of course, many "in-between" models exist.

If the most restricted, that is, homogeneous, first-order, stationary Markov model does not fit the data for the four rotation groups, a well-fitting model can always be found by relaxing one or more of the restrictions. In the end, the heterogeneous, highest order, nonstationary Markov model will always fit the data perfectly (with zero degrees of freedom). Another important strategy for obtaining well-fitting (parsimonious) models advocated by Van de Pol and Langeheine (1990) is to assume the population is heterogeneous and consists of two or more (unknown, latent) groups that behave differently but, hopefully, according to a simple Markov chain, albeit with

different sets of parameters. Their mixed Markov model deals with this unobserved heterogeneity.

So far, it is assumed that all measurements are perfect and that there are no classification errors, a presumption that is known to be false for SIPP data. Therefore, a (Markov) model that takes measurement error into account is needed. Such a model is the latent Markov model, originally developed by Wiggins and Poulsen (Wiggins 1955, 1973; Lazarsfeld and Henry 1968, chap. 9; Poulsen 1982) and put into a much more general framework by Van de Pol and Langeheine (1990) (see also the literature on hidden Markov chains [Juang and Rabiner 1991; Hughes, Guttorp, and Charles 1999]). The latent Markov model can also be viewed as a modified path model with latent variables, referred to as a modified LISREL model (Hagenaars 1990) or, in terms of the above, a directed log-linear model with latent variables.

The model in Figure 2 represents the standard basic latent Markov model. *A*, *B*, *C*, and *D* denote the labor market positions in four consecutive months as reported by the respondents (see Figure 1). Variables *W*, *X*, *Y*, and *Z* are their latent, not directly observed counterparts. *W* through *Z* are trichotomous latent variables, representing the true labor market states E, U, and N in the four successive months. The observed variables *A* through *D* are probabilistically, not perfectly, related to the latent variables: There is a chance that the respondents give the wrong answers and report states that are different from their true (latent) states.

Essential to latent variable models in general, and to the latent Markov model in particular, is that the joint probability that now involves both observed and latent variables be decomposed into a structural (causal) portion and a measurement portion (Hagenaars 1998):

$$\pi_{wxyzabcd}^{WXYZABCD} = (\pi_{wxyz}^{WXYZ})(\pi_{abcdwxyz}^{ABCD|WXYZ}) . \tag{4}$$

The first element on the right-hand side of equation (4) (π_{wxyz}^{WXYZ}) refers to the joint probability distribution for variables *W* through *Z*; it forms the structural portion, from which it may be inferred how variables *W* through *Z* are (causally) related to each other. All variables in the structural portion happen to be latent. The second right-hand side

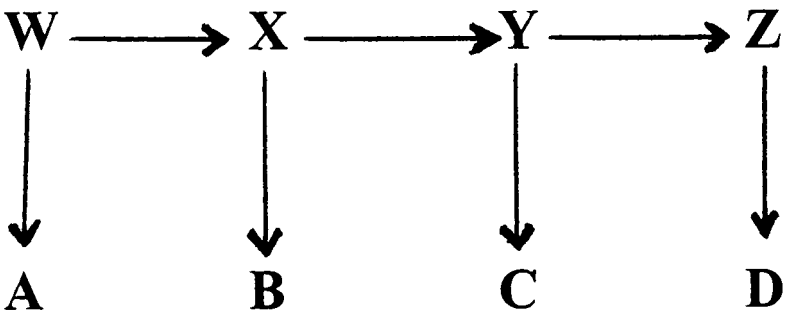


Figure 2: Standard First-Order Independent Classification Error Latent Markov Model

element ($\pi^{ABCD|WXYZ}_{abcdwxyz}$) is the measurement portion and indicates how the scores on indicators *A* through *D* depend on the structural variables *W* through *Z*. In the standard latent Markov model in Figure 2, it is assumed that the latent state transitions follow a first-order Markov chain over time. Furthermore, the standard latent class assumption of local independence is made, implying that the states observed at different occasions are independent of each other given the true state. In other words, the classification errors in the observed variables are conditionally independent of each other given the latent variables. Furthermore, each observed variable depends directly on only one latent variable. Obviously, the standard ICE assumption is made here. Given the validity of the first-order Markov and the ICE assumptions, equation (4) can be rewritten as follows:

$$\pi^{WXYZABCD}_{wxyzabcd} = (\pi^{W}_{w}\,\pi^{X|W}_{xw}\,\pi^{Y|X}_{yx}\,\pi^{Z|Y}_{zy})(\pi^{A|W}_{aw}\,\pi^{B|X}_{bx}\,\pi^{C|Y}_{cy}\,\pi^{D|Z}_{dz}). \quad (5)$$

The elements of the (first) structural portion on the right-hand side of equation (5) have an important interpretation: Given the validity of the model, they represent the true labor market transitions from time point *t* to *t* + 1, corrected for the misclassifications in the observed variables. The elements of the measurement portion are directly related to the reliabilities at the different points in time. For example, $\pi^{A|W}_{aw}$ is the conditional probability of observing state *a* in the first month of the reference period when the true state in the first month is *w*; *a* and *w*

may refer to the same state of the set {E, U, N}, in which case $\pi_{aw}^{A|W}$ indicates the conditional probability of the observed classification being correct, that is, in agreement with the true state, or *a* and *w* may refer to a different state, resulting in the conditional probability of a misclassification.[3]

If the data from the four rotation groups are simultaneously dealt with, an extension of equation (5) similar to the extension from equation (2) to equation (3) is required. Because of the way the SIPP is designed, the four rotation groups are in principle equivalent samples from the same population. Consequently, the (joint) distribution of the latent variables *W* through *Z* that are considered to reflect the true labor market positions in the population is not influenced by grouping variable *G*. However, the conditional response probabilities (the reliabilities) may vary across groups, since the temporal distance between the date of the interview and a particular month for which the information is given is different for the four groups. These considerations lead to equation (6):

$$\pi_{gwxyzabcd}^{GWXYZABCD} = (\pi_g^G \pi_w^W \pi_{xw}^{X|W} \pi_{yx}^{Y|X} \pi_{zy}^{Z|Y})(\pi_{agw}^{A|GW} \pi_{bgx}^{B|GX} \pi_{cgy}^{C|GY} \pi_{dgz}^{D|GZ}). \quad (6)$$

As before, saturated or nonsaturated log-linear models can be defined for the elements on the right-hand side of equation (6). Given appropriate (Poisson or [product]multinomial) sampling schemes, maximum likelihood estimates for the elements on the right-hand side of equation (5) or (6) can be found by standard methods, employing (combinations) of EM, Newton/Raphson, or scoring algorithms. Implementations of these algorithms with the kinds of models in hand have been described by Hagenaars (1990, 1993), Van de Pol and Langeheine (1990), Collins, Fidler, and Wugalter (1996), and, especially, Vermunt (1996, 1997b). The analyses for this article were carried out by means of Vermunt's (1997a) program *i*EM.

Once the maximum likelihood estimates for the right-hand side elements of equation (5) or (6) have been obtained, the maximum likelihood estimate of the joint probability $\pi_{wxyzabcd}^{WXYZABCD}$ (equation [5]) or $\pi_{gwxyzabcd}^{GWXYZABCD}$ (equation [6]) can be computed. Summation of $\hat{\pi}_{wxyzabcd}^{WXYZABCD}$ (or $\hat{\pi}_{gwxyzabcd}^{GWXYZABCD}$) over the latent variables yields the maximum likelihood estimates $\hat{\pi}_{abcd}^{ABCD}$ (or $\hat{\pi}_{gabcd}^{GABCD}$), which after multiplication by sample

size N can be compared in the usual ways with the observed frequencies f_{abcd}^{ABCD} (or f_{gabcd}^{GABCD}) to test the model.

When applying the latent Markov models in equations (5) and (6) to SIPP data, several difficulties were encountered. Most problems follow from the sparse and unbalanced nature of the observed frequency table. As can be inferred from Table 1, the monthly changes in labor market states are very small, and consequently variables A through D are highly multicollinear, that is, highly correlated with one another. Therefore, even though each rotation group has about 5,000 respondents and observed table $ABCD$ contains only $3^4 = 81$ cells, the observed table is a very sparse table with many zero entries. For most models, this will result in many extremely small estimated cell frequencies. Consequently, the approximation of the distributions of the standard chi-square test statistics toward the theoretical chi-square distribution will be problematic, as will be the approximation of the distributions of the maximum likelihood estimates of the parameters toward the normal distribution. So, one should proceed very cautiously when employing these conventional test statistics. The main problem is that the asymptotic theory underlying maximum likelihood estimation and testing (or for that matter estimation and testing procedures based on principles other than maximum likelihood) may not be appropriate for such sparse tables. This issue will be taken up again in the discussion. Below, a few practical consequences of sparse tables will be dealt with that are closely related to the main problem but deserve to be mentioned separately.

Sparse tables easily lead to boundary estimates, that is, to estimated (conditional) probabilities equal to zero or one. Often, such solutions with boundary estimates are local solutions of the maximum likelihood equations, and better, "more likely" solutions exist, as may be discovered by using a large number of different sets of initial parameter estimates. But even if the boundary solution is the best in the sense that no solution with a higher likelihood exists within the allowable parameter space, the estimates found must be regarded as either "terminal estimates" (Goodman 1974b) or conditional maximum likelihood estimates given that the boundary values are true for the population. However, this latter position is hard to defend if the boundary estimates result from sampling zeroes in the sparse table and not from a priori intended structural zeroes. It is therefore unclear whether to

take the quasi-structural zeroes into account when determining the number of degrees of freedom.

Besides making it hard to determine the nature of the estimates, boundary estimates can make it difficult to investigate the identifiability of the model. Identifiability can be a serious problem in latent variable models (Goodman 1974a; De Leeuw, Van der Heijden, and Verboon 1990; Clogg 1981) and must be expected to be a major problem in the kinds of models and data discussed here. A sufficient condition for local identifiability is that the information matrix (or, equivalently, the variance-covariance matrix of the parameter estimates) has full rank. In practice, one has to work with the estimated information matrix (for which *i*EM provides all eigenvalues). For a model to be identifiable, all eigenvalues should be strictly positive. Boundary estimates lead to nonpositive eigenvalues. If these boundary estimates have not been fixed a priori, it may be unclear whether the model is identifiable without these boundary estimates. For SIPP data, the models in equations (5) and (6) yielded many (estimated) nonpositive eigenvalues in combination with many boundary estimates. To determine the identifiability of the models as such, simulated data were used to obtain a solution without boundary estimates, and the eigenvalues were inspected for this solution.

Application of this procedure made it clear that neither in equation (5) nor in equation (6) were all parameters identified. This is well known from the literature on latent Markov chains and has been demonstrated by Poulsen (1982) and Van de Pol and Langeheine (1990), who indicated how this problem might be solved. Extra restrictions are needed to make the models identifiable, such as assuming equal reliabilities for all four indicators or stationarity of the first-order latent Markov chain, or, for the model in equation (6), equating particular reliabilities across groups. (For similar problems and solutions with continuous variables, see Heise [1969].) Several variants and combinations of these extra restrictions were used. The test outcomes and the parameter estimates for the several variants were similar. All models had to be rejected ($p < .00$) on the basis of the Pearson chi-square and likelihood ratio L^2. For example, the test results for the first-order, nonstationary Markov chain (equation [5]) applied to the first rotation group, with the additional restriction that the corresponding probabilities of a misclassification be the same for all four indica-

tors yielded $\chi^2 = 147.49$, $L^2 = 86.75$, $df = 54$, $p = .00$; adding to this model the stationarity assumption of equal latent transition tables resulted in the following test results: $\chi^2 = 189.33$, $L^2 = 104.61$, $df = 66$, $p = .00$. Given the problems caused by the sparse tables discussed above, it is uncertain whether these test results should be trusted. However, this is not too important, since there are other compelling reasons for rejecting these models. All models that are essentially based on equation (5) or (6) represent ICE models, and the resulting latent transition tables correspond only to the true changes if the ICE assumption is true. As discussed above, there are strong theoretical and empirical indications that the classification errors in the SIPP are not independent of each other. Not surprisingly, then, in all these ICE models the peculiarities found in the observed SIPP data (see Table 1) and that led to the conclusion of correlated errors did not disappear at the latent level. The latent turnover tables show even more stability than the observed tables, the latter already being considered too high. Also, the notorious seam effect was present at the latent level. Obviously, models are needed that reckon with dependent, correlated misclassifications.

In the next section, we will present models by which one can tackle the problem of correlated measurement errors. To keep the exposition of the approach simple and comprehensible, these models will be defined for the reference period of just one rotation group, using only one indicator for each latent variable. A disadvantage of the simplified approach is that the proposed models as such are not identified unless severe (and perhaps unrealistic) restrictions are imposed on the model parameters. Therefore, in the next section, no empirical results will be presented (although these restricted models did fit the data). Later, a more elaborate empirical example will be presented.

MODELS FOR CORRELATED CLASSIFICATION ERRORS

Because of the presence of the within-wave response consistencies, the associations between the observed variables are not completely explained by the direct effects of the latent variables on the indicators (as in Figure 2). There exist additional sources of association between the indicators over and above the portion that is explained by the indi-

cators' relations with their latent variables (Hagenaars 1988). A very general, but also from a certain point of view, uninformative way to model this additional source of association is to assume that there exists a hidden, unmeasured (latent) variable that causes consistency among the answers. The model in Figure 3a represents this approach, where V is the extra latent variable that influences all answers. V is assumed to be independent of the proper latent variables W through Z. This is the usual kind of assumption, made for reasons of identifiability or interpretability of the model, but it is not a necessary restriction. If Figure 3a is viewed as a directed graph, it corresponds to the following equation:

$$\pi^{VWXYZABCD}_{vwxyzabcd} = (\pi^{V}_{v}\pi^{W}_{w}\pi^{X|W}_{xw}\pi^{Y|X}_{yx}\pi^{Z|Y}_{zy})(\pi^{A|VW}_{avw}\pi^{B|VX}_{bvx}\pi^{C|VY}_{cvy}\pi^{D|VZ}_{dvz}). \quad (7)$$

Variable V is treated as a categorical latent variable. Therefore, the number of categories of V has to be determined. One might start with two categories and add more categories until the model is no longer identified or until a good fit with the data has been obtained. Another possibility is of course that one has theoretical reasons to start with a certain number of categories; here, for instance, three categories denoting the overall tendency of people to give the answer "employed," "unemployed," or "not in the labor force," respectively, regardless of their true position. This categorical hidden variable approach is closely related to (more standard, linear) models with correlated error terms (Bollen 1989, chap. 5), unobserved heterogeneity (Heckman and Singer 1982; DeSarbo and Wedel 1993; Vermunt 1996, 1997b), or random coefficients (Bryk and Raudenbush 1992; Qu, Tan, and Kutner 1996; Hadgu and Qu 1998).

Because an extra latent variable V has been introduced without additional observed indicators, the model of equation (7) has to be examined very carefully for (extra) identifiability problems. One possible but nonsufficient way of achieving identifiability is to define more restrictive log-linear models, for example, no-three-variable-interaction models for the three-way tables involving latent variable V in equation (7). Finding appropriate and meaningful restrictions may be problematic, since it is often difficult to provide a compelling substantive interpretation of latent variable V. Usually, this variable just accounts for the correlated error terms, but, from a substantive point of view, in an

unknown way and a large number of different interpretations can be attached to *V*. In general, preference should be given to models that incorporate the presumed nature of the systematic response errors.

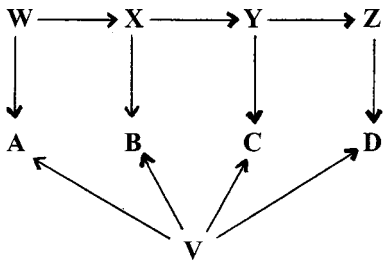
One possible substantive explanation for the correlated misclassifications within a particular reference period is that the respondent has the tendency to adapt the information about the past to the present, true position. In other words, the true position at the time of the interview influences all answers. Because the SIPP interview takes place at the beginning of the fifth month, latent variable *Z* (i.e., the true labor market position for the fourth month) comes closest to the true position at the time of the interview. Therefore, it is assumed, as depicted graphically in Figure 3b, that *Z* directly influences not only its own indicator *D* but also *A*, *B*, and *C*:

$$\pi^{WXYZABCD}_{wxyzabcd} = (\pi^{W}_{w}\,\pi^{X|W}_{xw}\,\pi^{Y|X}_{yx}\,\pi^{Z|Y}_{zy})(\pi^{A|WZ}_{awz}\,\pi^{B|XZ}_{bxz}\,\pi^{C|YZ}_{cyz}\,\pi^{D|Z}_{dz}). \quad (8)$$

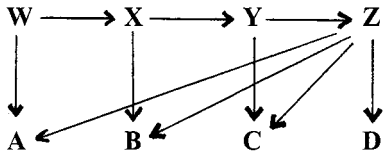
Restricted (log-linear or logit) models may be defined (and are necessary to achieve identifiability) for the probabilities on the right-hand side of equation (8). The model in Figure 3b still fulfills the local independence assumption (i.e., the indicators are independent of each other within the categories of the latent variables that influence the indicators) but not the ICE assumption (i.e., the answers at a particular point in time do not depend only on the true position at that particular point in time).

A possible response mechanism to account for the correlated misclassifications that violates the local independence assumption occurs when the respondent adapts an answer to previously given answers. During the SIPP interview, the respondent is shown a calendar and asked to recall his or her labor history during the reference period on a weekly basis. It is assumed here that while performing this task, the respondent most probably thinks first about his or her present status and then recalls his or her labor market states going back in time. This assumption implies that *D* may have an influence on *A*, *B*, and *C*; *C* on *A* and *B*; and *B* on *A*. Of course, we have no proof of the validity of these back-in-time dependencies, but they do lead to better fitting models than forward-in-time dependencies (which assume that *A* influences *B*, *C*, and *D*, etc.). The graphical models in Figures 3c and 3d depict two possible variants. In Figure 3c, it is assumed that only the first given answer (*D*) has a direct influence on all later answers:

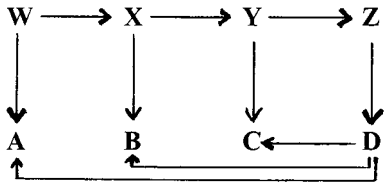
a) Unmeasured 'consistency-trait'



b) Consistency with true position at time of interview



c) Consistency with first given answer



d) Consistency with previously given answer

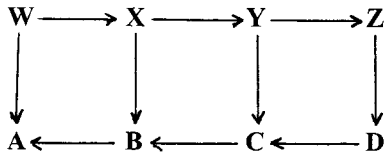


Figure 3: First-Order Latent Markov Model with Correlated Classification Errors

$$\pi^{WXYZABCD}_{wxyzabcd} = (\pi^{W}_{w}\pi^{X|W}_{xw}\pi^{Y|X}_{yx}\pi^{Z|Y}_{zy})(\pi^{A|WD}_{awd}\pi^{B|XD}_{bxd}\pi^{C|YD}_{cyd}\pi^{D|Z}_{dz}). \quad (9)$$

In Figure 3d, the assumption is made that only each immediately previous answer directly influences the next one:

$$\pi^{WXYZABCD}_{wxyzabcd} = (\pi^{W}_{w}\pi^{X|W}_{xw}\pi^{Y|X}_{yx}\pi^{Z|Y}_{zy})(\pi^{A|WB}_{awb}\pi^{B|XC}_{bxc}\pi^{C|YD}_{cyd}\pi^{D|Z}_{dz}). \quad (10)$$

None of the above models is identifiable without further restrictions. Which particular model and which particular set of additional restrictions to choose depend on the design of a particular study and the way the data have been collected, and, of course, on theoretical considerations and empirical results: whichever theoretically meaningful model explains the observed data best.[4] An extensive application to SIPP data of one particular model will be presented in the next section.

MODELING SYSTEMATIC MISCLASSIFICATIONS IN THE SIPP

Although the approach sketched above on how to handle systematic misclassifications is very general and flexible, in this concrete (SIPP) case these models are not identified as such and, even with identifying restrictions, are often still just "poorly identified" (Davidson and MacKinnon 1993, sec. 6.3), certainly in combination with the very sparseness of the table. Using more data in the sense of using the data from all four rotation groups of the SIPP does not really solve the identification problem; an important source of the identification problem here is that each latent variable has only one indicator. Fortunately, for each time period of the SIPP study, a second (dichotomous) indicator for respondents' labor market status is available using a question from another section of the SIPP questionnaire, that is, the Earnings and Employment section, in which the respondents were interviewed again about their labor market position but in a cruder manner. The question posed to respondents was whether they did or did not have a job during the whole reference period. If they stated that they did not have a job, they received a score of NJ (no job) on all four

monthly second indicators. If they answered that they had a job, they were asked to indicate precisely in which calendar period(s) they were employed. On the basis of these answers, respondents were assigned either to category NJ or to category J (had a job) of the second indicator (applying the previously discussed monthly modal assignment rule). The basic ICE latent Markov model with two indicators is depicted in Figure 4.

Variables *A* through *D* in Figure 4 indicate the four trichotomous labor market indicators that were used above; more precisely, they refer to the labor market status in the months January, February, March, and April (1986). The four trichotomous latent variables *W* through *Z* are (again) their unobserved counterparts. The four new dichotomous indicators are denoted by *E*, *F*, *H*, and *I*, with categories indicated as *e*, *f*, *h*, *i*, respectively, each varying over the set {J, NJ}. For a particular month, if no classification errors were made, all people in observed category J would belong to latent category E and all respondents in observed category NJ would belong to either latent category U or latent category N. It was assumed that given the more global character of this question and the less complicated task the respondent had to perform, the errors for these four dichotomous indicators would have the ICE property. However, this was not the case, as was clear from an inspection of the observed relationships between the four dichotomous indicators. The same irregularities as for indicators *A* through *D*, such as the seam effect, showed up, although to a smaller extent.

The basic starting equation corresponding to the graphical ICE model in Figure 4, but now extended to include the four rotation groups used in the analyses, looks as follows (denoting the group variable again by *G* and assuming, as above, that the distribution of the latent variables is the same for all groups; see equation [6]):

$$\pi^{GWXYZABCDEFHI}_{gwxyzabcdefhi} = (\pi^{G}_{g}\,\pi^{W}_{w}\,\pi^{X|W}_{xw}\,\pi^{Y|X}_{yx}\,\pi^{Z|Y}_{zy})$$
$$(\pi^{A|GW}_{agw}\,\pi^{B|GX}_{bgx}\,\pi^{C|GY}_{cgy}\,\pi^{D|GZ}_{dgz}\,\pi^{E|GW}_{egw}\,\pi^{F|GX}_{fgx}\,\pi^{H|GY}_{hgy}\,\pi^{I|GZ}_{igz}). \qquad (11)$$

The maximum likelihood estimates of $\pi^{GWXYZABCDEFHI}_{gwxyzabcdefhi}$ in equation (11) can be found by defining saturated log-linear models for the (marginal) tables corresponding to the (conditional) probabilities on the

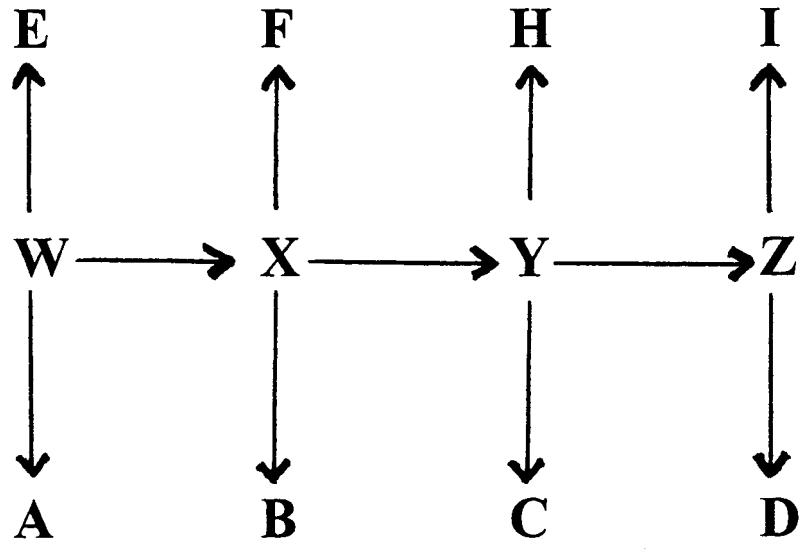


Figure 4: First-Order Independent Classification Error Latent Markov Model With Two Indicators

right-hand side of equation (11), that is, model {*G*} for marginal table *G*; model {*W*} for marginal table *W*; models {*WX*}, {*XY*}, and {*YZ*} for marginal tables *WX*, *XY*, and *YZ*, respectively; model {*WGA*} for marginal table *WGA*; and so on (employing the standard shorthand notation for hierarchical log-linear models).

In the directed log-linear model of equation (11), it is assumed that within each rotation group, the ICE assumption holds, an assumption that is known to be false for SIPP data. The ICE model in equation (11) serves as a baseline model, albeit in a simplified form. In the analyses reported below, it will be assumed that the reliabilities of indicators *A* through *D* are equal to each other and are the same for all groups, as are the reliabilities of indicators *E*, *F*, *H*, and *I*. Possible distortions of the equalities of the reliabilities are supposed to be caused by possible extra effects among the indicators.

Reliability is essentially determined by the nature and strength of the relation between a latent variable and its indicator(s). With categorical data, two approaches prevail. One is defining reliability in

terms of conditional response probabilities, that is, in terms of the probability of giving the correct answer in agreement with one's true (latent) position (see note 3). In the other approach, reliability is defined in terms of the odds ratio describing the relationship between the latent variable and the indicator. Although the two approaches have much in common, they are not completely equivalent. In the conditional response probabilities approach, the equal reliabilities assumption implies for equation (11), imposing the appropriate equality restrictions within the set of conditional probabilities $\{\pi^{A|GW}_{agw}, \pi^{B|GX}_{bgx}, \pi^{C|GY}_{cgy}, \pi^{D|GZ}_{dgz}\}$ and within the set $\{\pi^{E|GW}_{egw}, \pi^{F|GX}_{fgx}, \pi^{H|GY}_{hgy}, \pi^{I|GZ}_{igz}\}$ (Mooijaart and Van der Heijden 1992; Vermunt 1996, 1997b [to be implemented by means of, for example, *i*EM]). In terms of the log-linear parameterization, such equality restrictions on the conditional probabilities imply particular restrictions on both one- and two-variable log-linear effects (Haberman 1979:551; Hagenaars 1990:31; Heinen 1996:66-71). In the odds ratio approach to reliability (chosen here), only restrictions on particular odds ratios (or the corresponding two-variable log-linear effects, such as $\lambda^{WA}_{wa}, \lambda^{XB}_{xb}$) are imposed, and no restrictions are imposed on one-variable log-linear effects (such as λ^{A}_{a}, λ^{B}_{b}).[5]

For marginal tables *WGA*, *XGB*, *YGC*, and *ZGD* corresponding with conditional probabilities $\pi^{A|GW}_{agw}, \pi^{B|GX}_{bgx}, \pi^{C|GY}_{cgy}, \pi^{D|GZ}_{dgz}$ in equation (11), no-three-variable-interaction (logit) models are defined ({*WG*, *WA*, *GA*}, {*XG*, *XB*, *GB*}, {*YG*, *YC*, *GC*}, and {*ZG*, *ZD*, *GD*}, respectively), and analogous no-three-variable-interaction models are defined for marginal tables *WGE*, *XGF*, *YGH*, and *ZGI*. Furthermore, the equal reliabilities assumption is made by restricting the relations between the latent variables and their indicators in the following way:

$$\lambda^{WA}_{ij} = \lambda^{XB}_{ij} = \lambda^{YC}_{ij} = \lambda^{ZD}_{ij} \qquad \textit{for all } ij. \qquad (12)$$

$$\lambda^{WE}_{ij} = \lambda^{XF}_{ij} = \lambda^{YH}_{ij} = \lambda^{ZI}_{ij}$$

Because of the absence of all three-variable interactions (involving *G*), the relations between a particular latent variable and its two indicators are the same for all four rotation groups. Because of the restrictions in equation (12), the reliabilities of indicators *A* through *D* are identical, as are the reliabilities for indicators *E*, *F*, *H*, and *I*. The test

outcomes for the thus restricted (and identifiable) model of equation (11) are presented in model 1 of Table 2. Given the extreme sparseness of the observed table, the test statistics cannot be expected to follow the theoretical chi-square distribution, as is immediately clear from the extremely large difference between Pearson and log-likelihood ratio chi-squares. The *p* value found has no meaning here (and is therefore not reported). But the test outcomes for this model are useful as a standard of comparison for the models to come that are not based on the ICE assumption.

From some preliminary analyses and substantive considerations, it was concluded that the best way to introduce extra effects among the indicators would be to assume that within a particular reference period, each answer is directly influenced by the immediately preceding answer (see Figure 3c). Because the answers obtained from two different, successive interviews are four months apart, the misclassifications for the indicators belonging to two different reference periods are assumed to be independent of each other. Given the SIPP interviewing scheme, all information for rotation group 1 about the labor market states in the months January through April has been obtained within the same interview, and therefore direct test-retest effects should be expected between each pair of two successive months. This is depicted in Figure 5 by the three arrows for group 1. For rotation group 2, the information on the states in February through April belongs to one reference period but the scores in January to another. Therefore, in Figure 5 there are no arrows (direct effects) between the answers in January and February for group 2. For similar reasons, the arrow between February and March is missing for group 3 and the arrow between March and April is missing for group 4. All this leads to the following restricted correlated error model (model 2 in Table 2). First, to introduce the direct test-retest effects of the previous answer, equation (11) is replaced by the following equation:

$$\pi^{GWXYZABCDEFHI}_{gwxyzabcdefhi} = (\pi^{G}_{g}\,\pi^{W}_{w}\,\pi^{X|W}_{xw}\,\pi^{Y|X}_{yx}\,\pi^{Z|Y}_{zy})$$
$$(\pi^{A|GWB}_{agwb}\,\pi^{B|GXC}_{bgxc}\,\pi^{C|GYD}_{cgyd}\,\pi^{D|GZ}_{dgz}\,\pi^{E|GWF}_{egwf}\,\pi^{F|GXH}_{fgxh}\,\pi^{H|GYI}_{hgyi}\,\pi^{I|GZ}_{igz}). \quad (13)$$

To obtain the intended maximum likelihood estimates for the elements on the right-hand side of equation (13), saturated (logit) models

TABLE 2: Models for Survey of Income and Program Participation Data (rotation groups 1 through 4) (January 1986 to April 1986)

Model	*Pearson* ^{2}	L^2	df	*BIC*
1. Equation (11): Independent classification error, but no three-variable interactions and equal reliabilities (equation [12])	8,354,349.78	5,547.19	5,106	–45,021.04
2. Equation (13): Not independent classification error, but equal reliabilities and with test-retest effects for within-wave observation	110,860.59	2,472.98	5,061	–47,649.58
3. Model 2, but restricted test-retest effects "a," "b," "c" in Figure 5	35,291.97	2,686.89	5,091	–47,732.78

$\{G\}$, $\{W\}$, $\{WX\}$, $\{XY\}$, and $\{YZ\}$ are defined for subtables G, W, XW, XY, and YZ, respectively. To obtain the estimate of $\pi_{agwb}^{A|GWB}$, a logit model has to be defined in which the effects of G, W, and B on A are defined for subtable $GWAB$ in agreement with the hypotheses. The appropriate log-linear model is a restricted version of hierarchical model $\{WGB, WA, GAB\}$:

$$\pi_{wgab}^{WGAB} = (\lambda + \lambda_w^W + \lambda_g^G + \lambda_b^B + \lambda_{wg}^{WG} + \lambda_{wb}^{WB} + \lambda_{gb}^{GB} + \lambda_{wgb}^{WGB}) + (\lambda_a^A + \lambda_{wa}^{WA}) + (\lambda_{ga}^{GA} + \lambda_{gb}^{GB} + \lambda_{ab}^{AB} + \lambda_{gab}^{GAB}). \quad (14)$$

The effects represented by the term $\{WGB\}$ reflect the fact that in the relevant logit equation for the effects of W, G, and B on A, one conditions on the joint distribution of the independent variables. The term $\{WA\}$ refers to the direct effects of W on A, which do not interact with G or B; therefore, the reliabilities are equal across groups. Moreover, λ_{wa}^{WA} in equation (14) has to be restricted according to equation (12) because it is assumed that the reliabilities are constant over time. The effects represented by the term $\{GAB\}$ imply that there is a direct effect of B on A and that this effect may be different for the four rotation groups. This three-variable interaction effect is necessary because A and B have different positions within the interview scheme of the four

rotation groups. In particular, it is assumed that *B* has no effect on *A* within rotation group 2 because the information on *A* and *B* has been obtained during two different interviews (see Figure 5). Such a restriction about the conditional relationship between *A* and *B* for group 2 being zero (i.e., $\lambda_{ab2}^{AB|G}$) can be imposed by reparameterizating equation (14) and replacing the terms $(\lambda_{ab}^{AB} + \lambda_{gab}^{GAB})$ by $\lambda_{abg}^{AB|G}$, resulting in the following equation:

$$\pi_{wgab}^{WGAB} = (\lambda + \lambda_{w}^{W} + \lambda_{g}^{G} + \lambda_{b}^{B} + \lambda_{wg}^{WG} + \lambda_{wb}^{WB} + \lambda_{gb}^{GB} + \lambda_{wgb}^{WGB}) + (\lambda_{a}^{A} + \lambda_{wa}^{WA}) + (\lambda_{ga}^{GA} + \lambda_{gb}^{GB} + \lambda_{abg}^{AB|G}). \quad (15)$$

For each rotation group *g*, there are four independent conditional effects $\lambda_{abg}^{AB|G}$ to estimate, imposing the usual identifying restrictions. These 16 independent parameters replace completely the (16 independent) effects $(\lambda_{ab}^{AB} + \lambda_{gab}^{GAB})$ and, without further restrictions, equations (14) and (15) yield completely identical estimates $\hat{\pi}_{wgab}^{WGAB}$. By setting up the appropriate design matrix for conditional effects, $\lambda_{abg}^{AB|G}$ can be defined, including the postulated restriction $\lambda_{ab2}^{AB|G} = 0$ (Evers and Namboodiri 1978).

In a completely analogous way, the appropriate restricted log-linear models are defined for subtables *XGBC* and *YGCD* to get the estimates for $\pi_{bgxc}^{B|GXC}$ and $\pi_{cgyd}^{C|GYD}$ in equation (13). Because measures "earlier" than *D* are ignored in this analysis, to obtain the estimates for $\pi_{dgz}^{D|GZ}$, model {*ZG*, *ZD*, *GD*} is defined for subtable *DGZ*. Finally, the whole procedure is analogously repeated for the dichotomous indicators *E* through *I*, including the equal reliabilities restriction in equation (12). All this results in an identified model, denoted in Table 2 as model 2.

The introduction of the test-retest effects between the successive indicators in the described manner costs 45 degrees of freedom compared to model 1 but yields an enormous improvement in terms of L^2 ($L_1^2 - L_2^2 = 5{,}547.19 - 2{,}472.98 = 3{,}074.21$). When using L^2 or BIC as a descriptive measure of fit, there is no doubt that model 2 is the preferred model compared to model 1. Although with extremely sparse tables one has to be very careful in drawing definitive conclusions on the basis of these fit indexes, there is evidence that outcomes for the comparisons of parsimonious models (here, 5,184 cells and about

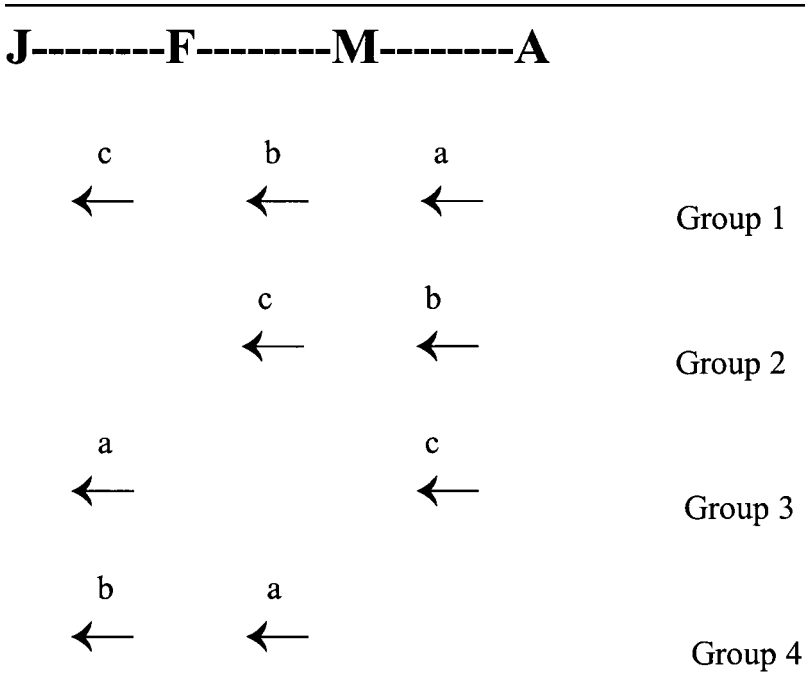


Figure 5: Survey of Income and Program Participation Interviewing Scheme for the Four Rotation Groups in the Period From January Through April 1986

20,000 respondents compared to 5,061 degrees of freedom for model 2) of conditional L^2 tests and comparisons of BIC indexes are generally more trustworthy (as test statistics) than noncomparative tests (Haberman 1978:341; Raftery 1993, 1995; but see Weakliem 1999). It seems safe to conclude that, as expected, not-ICE model 2 with the described direct effects of the successive indicators on each other has to be preferred above ICE model 1.[6]

Model 2 might be further restricted in agreement with the SIPP interviewing scheme by assuming that the effects of the first answer on the second, going backward in time during the reference period, are the same for all four groups (the arrows indicated by "a" in Figure 5), as are the effects of the second answer on the third ("b") and of the third answer on the fourth one ("c"). The test results of introducing these extra restrictions are reported in model 3 of Table 2. Compared

to model 2, model 3 has 30 more degrees of freedom, while $L^2_{3/2} = 213.91$. In traditional terms, model 3 has to be rejected in favor of model 2. But does L^2 (even conditional L^2) approximate the theoretical chi-square distribution adequately, or is it "too big" and "too progressive" given the extremely sparse table? According to BIC, model 3 should be chosen. But BIC might be too conservative, favoring the more parsimonious model 3 over the less parsimonious model 2. Parametric bootstrapping to determine the sampling distribution of L^2 might provide the answer but was not feasible with these models and data; it simply takes too much (days of) computing time. Therefore, the parameter estimates of models 2 and 3 were inspected and compared. Although it might be due to sampling error, the parameter estimates of model 2 did not at all suggest the "a," "b," "c," restricted pattern in Figure 5. So it was decided to accept model 2 as the final model. For some reasons, perhaps related to the particulars of the fieldwork of the SIPP, the direct influence of the first answer on the second, and so on, is not the same for all four rotation groups.

Of course, one can never claim that model 2 is the correct model, that is, the population model that generated the data. Other well-fitting models may be found. For example, Bassi et al. (1995), analyzing the same data set, decided on another "final" latent directed log-linear model. We will briefly discuss this directed log-linear model to demonstrate its flexibility. Bassi et al. followed a suggestion by Hubble and Judkins (1987). With regard to the within-wave classification errors, Bassi et al. assumed that during a particular interview (wave), once the respondents give a wrong answer for a certain month, they continue to report that incorrect answer for the following months, going backward in the wave, with absolute certainty. With regard to the between-wave classification errors, Bassi et al. (1995) assumed (as above) the ICE mechanism. Hubble and Judkins's suggestions lead to a complex model with four variable interaction effects, in which the answers for a particular month depend not only on the current true state but also on the discrepancies between the past true and past reported states. However, given the (partly) deterministic nature of the response mechanism, many of the conditional response probabilities are a priori fixed to zero or one, yielding a rather parsimonious model. (The test outcomes were $\chi^2 = 985{,}630.55$, $L^2 = 3{,}094.50$, $df = 5{,}097$, BIC = –47,384.65.)

In model 2, the estimates of the reliabilities as measured by the direct effects of the latent variables on their indicators are very high. Not surprisingly , the trichotomous indicators are more reliable than the dichotomous ones, for example, for category E (employed), $\hat{\lambda}_{11}^{WA}$ $(=\hat{\lambda}_{11}^{XB}=\hat{\lambda}_{11}^{YC}=\hat{\lambda}_{11}^{ZD})=5.411(\hat{\sigma}_{\lambda}=0.293)$, while $\hat{\lambda}_{11}^{WE}$ $(=\hat{\lambda}_{11}^{XF}=\hat{\lambda}_{11}^{YH}=\hat{\lambda}_{11}^{ZI})=3.4733(\hat{\sigma}_{\lambda}=0.074)$. The reliabilities for the not-in-the-labor-force category (N) have more or less the same sizes and pattern: $\hat{\lambda}_{33}^{WA}=4.489$ and $\hat{\lambda}_{32}^{WE}=2.667$. The most unreliable category is category U (unemployed): $\hat{\lambda}_{22}^{WA}=0.828(\hat{\sigma}_{\lambda}=0.155)$ and $\hat{\lambda}_{22}^{WE}=0.804$ $(\hat{\sigma}_{\lambda}=0.066)$. If the distributions of *W* and *A* are fixed at their marginal distributions, the values found for $\hat{\lambda}_{wa}^{WA}$ imply that the conditional probabilities of giving the correct answer for *A* given the true state of *W* are almost perfect for categories E and N ($\pi_{11}^{A|W}=0.998$ and $\pi_{33}^{A|W}=0.992$, respectively) but very bad for category U ($\pi_{22}^{A|W}=0.534$). The reason for the low reliability of category U might be that "being unemployed" does not have the same clear meaning for the respondent as "being employed" and "not being in the labor force"; additionally, to call oneself unemployed is difficult because it is a labor market state that is considered to be socially undesirable. In this respect, one might consider defining models (as a reviewer suggested) in which one only assigns errors to reporting of unemployment status.

As expected, the distorting (test-retest) effects of the previous answer are consistently larger for the trichotomous than for the dichotomous indicators. For example, the direct effects of *A* on *B* for category E and rotation group 1 are $\hat{\lambda}_{111}^{AB|G}=3.895(\hat{\sigma}_{\lambda}=0.492)$, while the effect of *E* on *F* is $\hat{\lambda}_{111}^{EF|G}=1.860(\hat{\sigma}_{\lambda}=0.193)$. The effects for category N of the trichotomous indicators *A* and *B* are about the same size as for the category E, but the effects for category U are much smaller: $\hat{\lambda}_{221}^{AB|G}=1.219(\hat{\sigma}_{\lambda}=0.243)$. It appears that mentioning being employed or not being in the labor force on *B* does influence the answer on *A* strongly in the same direction, but that mentioning being unemployed does not have such a large impact on *A*. This is in agreement with the supposition that being unemployed is socially undesirable.

All these effects are very large. However, they have to be interpreted with care. Given the many extremely small expected cell frequencies, the absolute sizes of the parameters depend very much on the accuracy of the convergence criterion for the iterative estimating

procedure. Even very small changes in the fifth or seventh decimal of the estimated probabilities may change the log-linear parameter estimates, for example, from 5.0 to 8.0. Therefore, we carefully checked whether the tendencies described above were found for the other indicators and rotation groups and in tables where the estimated probabilities were rounded off to four decimals and where .0001 was added to all probabilities. The tendencies reported above were consistently found. Other things were less clear; for example, what did people who are latently (truly) unemployed answer if they gave the wrong answer, were employed, or were not in the labor force? The relevant parameter estimates (e.g., $\hat{\lambda}^{WA}_{21}$, $\hat{\lambda}^{WA}_{23}$) varied too much and were numerically unstable, and therefore were omitted from the above discussion.

The consequences of the unreliabilities of the indicators and their direct effects on each other for the estimates of the changes in the labor market are shown in Table 3. The observed transition probabilities in the first row of Table 3 for the first rotation group for the months January through February ($B|A$) are calculated by means of data that were obtained within one and the same interview. For the second rotation group (row 2), the data come from two different successive reference periods. It is clearly seen that the stabilities (i.e., the conditional probabilities of not changing one's state) are larger according to rotation group 1 than to group 2. The true stabilities $X|W$ estimated on the basis of model 2 are in between, at least for categories E (column EE) and N (column NN).

What happens to category U is more complicated. Not surprisingly, the observed and latent stabilities (column UU) of category U are much smaller than the corresponding stabilities of categories E and N: By definition, unemployed people are looking for a job and are eager to lose their unemployment status. Also not unexpected is the fact that the observed stability of category U is much larger in group 1 than in group 2 and that the estimated latent stability of category U is much less than the latent stabilities of categories E and N. But what may seem surprising is that the estimated latent stability is not in between the observed stabilities of groups 1 and 2. How does one explain this? It is well-known from the literature that random measurement error in combination with a skewed distribution will make the smallest, less

TABLE 3: Survey of Income and Program Participation Gross Flows (percentages) (January 1986 [indicator *A*, latent variable *W*] to February 1986 [indicator *B*, latent variable *X*])

	EE	*EU*	*EN*	*UE*	*UU*	*UN*	*NE*	*NU*	*NN*
Observed group 1 $p_{b1a}^{B\|GA}$	98.29	1.16	0.55	17.44	77.95	4.62	0.63	1.45	97.83
Observed group 2 $p_{b2a}^{B\|GA}$	94.52	2.01	3.47	22.27	44.55	33.18	5.03	3.45	91.52
Latent *W*-*X* $\hat{\pi}_{xw}^{X\|W}$	97.19	2.17	0.64	6.29	86.48	7.23	2.99	1.37	95.64

NOTE: E = employed, U = unemployed, N = not in the labor force.

frequently occurring categories look very volatile and prone to change, much more than the bigger categories. This is true even when all categories in reality (truly, at the latent level) have the same amount of stability and are affected by the same (small) amount of random measurement error (Maccoby 1956; Hagenaars 1993:52-55, 1994). Now, according to model 2, the true (latent) distribution of employment status is $\hat{\pi}_1^W = 0.5930$, $\hat{\pi}_2^W = 0.0787$, and $\hat{\pi}_3^W = 0.3283$, with U being by far the smallest category. Consequently, it must be expected that because of measurement error alone, the observed changes for category U will be large and much larger than for categories E and N. This is confirmed by the observed transition probabilities in column UU for group 1 and, especially, group 2 (and would have occurred for category U—being the smallest—even if the stabilities and unreliabilities had been the same as for categories E and N). Furthermore, it must be expected that when one corrects for random measurement error, as is (also) done in latent variable models, the latent stability of the smallest category will be much larger than the observed stability, a result that is found in the present study. So, the latent stability of category U does not have a value in between the observed stabilities of groups 1 and 2. And although the latent stability of category U is smaller than the stabilities of categories E and N, the true (latent) difference between U on one hand and E and N on the other is much less than suggested by the observed data.

DISCUSSION

As this application illustrates, directed log-linear modeling with latent variables provides an excellent and flexible tool for analyzing data that are affected by random and systematic classification errors. At the same time, it has become clear that there may be serious problems to overcome. Most of the problems discussed below have to do with the nature of these models and with the potential sparseness of the frequency tables. As was indicated above, sparse tables may cause serious estimation and testing difficulties. Finally, attention will be paid to the sometimes problematic nature and meaning of latent variables acting as gold standards.

The introduction of systematic classification errors may cause extra identification problems that have to be solved. Especially when each latent variable is measured by just one indicator, non- or hardly identified models may be almost unavoidable. The remedy is clear: more than one indicator for each latent variable, but these data may not be available. Defining parsimonious models is then the next best solution, but only when such very restricted models are not completely unrealistic. The only choice remaining, but still usually much better than completely ignoring systematic measurement errors, is to set up models in which certain effects representing random and/or systematic classification errors are given particular a priori values and to investigate the consequences for the latent and manifest changes under study.

In the standard application of maximum likelihood estimation and testing procedures for log-linear modeling, it is assumed (as here) that the data follow a (product)multinomial sampling distribution, or, in practical terms, that the data arise from simple random sampling or stratified sampling with simple random sampling within the strata. Other sampling schemes, such as cluster sampling, will not result in (product)multinomial distributions, and the most likely sampling design effect is an inflation of the test statistics. Such possible design effects have not been taken into account for the SIPP data in the present study.

Even with extremely sparse data, the estimation results are often surprisingly robust, at least for rather parsimonious models. Nevertheless, with sparse data one must check the robustness of the estimates (as was done here on a modest scale) and not take it for granted. The use of cross-validation (reserving part of the sample for cross-validation)

is essential. It is also easy to consider the final parameter estimates as population values and draw a few samples from them, not necessarily to carry out a complete parametric bootstrapping procedure but, rather, to determine which parameters obtain (very) different values. For the same purposes, one might add (very) small constants to the cell frequencies in an arbitrary manner (e.g., adding .001 to every cell frequency) or in a (pseudo) Bayesian manner to stay away from the boundary of the parameter space (Schafer 1997) and investigate the influence on the parameter estimates. It is also very useful to add or delete certain rather small effects and study the consequences for the (other) parameter estimates. In this way, one discovers which results (for this particular model) are robust and which should be treated with care.

The biggest problem caused by sparse data is the testing and selection of models. Parametric bootstrapping procedures (Van der Heijden, Hart, and Dessens 1997; Collins et al. 1993; Langeheine, Pannekoek, and Van de Pol 1996) or fully Bayesian analyses (Rubin and Stern 1994; Gelman et al. 1995) may provide a solution to these problems. However, these procedures may still be impractical for some problems even with modern computing equipment; perhaps even more important, we still do not know how these procedures behave in the case of extremely sparse tables, such as those discussed in this article. In the end, it is often necessary to base the final model selection largely on descriptive measures of fit and on theoretical side information. The need to test these models on new data is obvious.

A final discussion point worth mentioning is the status and meaning of the latent variables, especially in longitudinal studies. What does it mean when we say that in reality, the estimated proportion of the employed in January is $\hat{\pi}_1^W = 0.5930$, or when we interpret the last row in Table 3 in terms of estimated true changes from January to February? Of course, the model used to obtain these estimates has to be valid. If that is true, the estimates of the latent parameters are the true ones in the sense that they are the ones that would have been obtained if the observed data had been free of misclassifications. The misclassifications introduced here are partly systematic, involving the direct effects of the answers (the indicators) on each other and partly random, that is, ICE. It is the random portion that is actually the most

difficult to interpret, and so the systematic portion will be further ignored in this discussion.

In the early days of the development of latent class models, it was recognized that there are three causes of observed changes (Lazarsfeld 1972; Kendall 1954; Wiggins 1955, 1973; Hagenaars 1990:181-83). First, the observed changes may reflect real changes, that is, the changes in the true score. Second, the observed scores mirror the accidental changes people experience, that is, changes because of mood, chance factors, accidental and temporary loss of job, and so on. Finally, there are the proper measurement errors, accidental response errors made by the respondents or the interviewers, processing errors, and so on. (Lazarsfeld [1972] assigned a minor role to these 'psychometric' errors.) Now, it should be recognized that latent variable models correct not only for the proper measurement errors but also for the observed true but accidental changes, and that the true, latent changes are observed changes purified of both sources of random change. Latent class analysis as such cannot make the difference between somebody who is usually employed but at one point in time happens to be erroneously recorded as unemployed and a person who is usually employed but at one point in time happens to be without a job. Even if the true labor market states had been observed without any response error in the strict psychometric sense (but, as is often the case, a portion of the true observed movements had a random character), applying the latent variable model would have led to latent turnover tables that show more stability than the corresponding true observed tables. Whether this is problematic depends on the purposes of the investigation.

In this respect, the intended definition of true score becomes important. A distinction can be made between the platonic true score and the operational true score (Sutcliffe 1965a, 1965b; Lord and Novick 1968, sec. 2.9; Hagenaars 1990, sec. 4.4; Sobel 1994). In the platonic true score model, it is assumed that there exists a real and actual true score. Somebody is married or not, works or not, has a certain weight, and so on. In the operational true score model, the true score is defined as the score a person obtains on average in a series of independently conducted experiments. Whether this distinction matters depends on the purposes of the analysis. Here, it is argued that sometimes it does matter. If one wants to measure the true underlying attitude of a person

or, for example, a person's weight, in the sense of the value that is not affected by daily or even hourly random fluctuations in the true, platonic value, one is interested in the true score as the operational true score (whether or not it is meaningful to assume that the platonic score really exists). Because latent variable models correct for both random errors and random behavior, they are appropriate to employ when the interest lies in the operational true score. However, if one is interested in the real number of people that is actually unemployed at a certain moment in time (the platonic true score) regardless whether they are "always" unemployed or typically employed but accidentally at this moment unemployed, latent variable models will generally not be appropriate because in principle the existing real number of unemployed people will not be the same as the latent number of unemployed people, and the amount of real changes that are occurring in labor market will be underestimated by the latent changes. For an estimation of the platonic score, a model including a gold standard is needed; for the operational score, latent variable models are most appropriate and, as seen in the present study, provide the researcher with a flexible tool to correct for all kinds of classification errors.

NOTES

1. The term *causal* is used somewhat loosely to denote asymmetrical relationships between variables. For more exact definitions of causality and some opposing views, see Rubin (1974), Sobel (1995), Glymour et al. (1987), Pearl (1995), and Raftery (1998).

2. In principle, modified path models (or directed log-linear models) must be estimated in the stepwise manner described here. Sometimes, however, it is possible to obtain the estimates for the joint probabilities in the full table directly by specifying one (log-linear) model rather than a sequence of submodels. This possibility has to do with the "collapsibility" of the log-linear model(s) (Agresti 1990, sec. 5.4.2.) and whether the causal model satisfies the "Wermuth condition" and is a "moral graph" (Whittaker 1990, sec. 3.5; Pearl 1995).

3. Interpreting the conditional response probabilities in terms of probabilities of misclassifications is most appropriate when the indicator directly depends on just one particular latent variable and when there is a one-to-one correspondence between the categories of the indicator and the latent variable (Sutcliffe 1965a, 1965b; Hagenaars 1990; Kuha and Skinner 1997). See also the discussion below and note 5.

4. Depending on the (identifying) restrictions imposed, several of these models may be empirically indistinguishable from each other, since they yield the same estimated expected frequencies.

5. In the standard (linear) approach, reliability is defined as the proportion of the variance of the indicator(s) that is explained by the latent variable (the true scores). When categorical variables are seen as realizations of underlying continuous variables, the same basic (standard) ap-

proach essentially still applies, as ingeniously shown by Bartholomew and Schuessler (1991). For truly categorical data, in addition to the two approaches mentioned in the main text, the explained variance definition of reliability might be used, but now with measures of qualitative variance such as entropy or concentration. (For an overview of such measures, see Vermunt [1997a:76].) To our knowledge, the properties of this latter approach have not been investigated.

6. An additional difficulty when comparing models 1 and 2 in Table 2 arises from the fact that zero estimates occurred in the latent turnover tables (model 1 $\hat{\pi}_{32}^{Y|X} = \hat{\pi}_{31}^{Z|Y} = \hat{\pi}_{32}^{Z|Y} = 0$, model 2 $\hat{\pi}_{32}^{Y|X} = \hat{\pi}_{32}^{Z|Y} = 0$ [and $\hat{\pi}_{31}^{Z|Y} = .0053$]). The models are identified, also without the zero estimates. Furthermore, several different initial estimates and different algorithms were used, always with the same results, including the zeroes. The zero estimates were not treated as restrictions in the computation of the number of degrees of freedom but, rather, as estimated parameters.

REFERENCES

Abowd, John M. and Arnold Zellner. 1985. "Estimating Gross Labor Force Flows." *Journal of Business and Economic Statistics* 3:254-83.

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

Bartholomew, David J. and Karl F. Schuessler. 1991. "Reliability of Attitude Scores Based on a Latent Trait Model." Pp. 97-123 in *Sociological Methodology 1991*, edited by Peter V. Marsden. Oxford: Blackwell.

Bassi, Francesca, Marcel Croon, Jacques A. Hagenaars, and Jeroen K. Vermunt. 1995. "Estimating Latent Turnover Tables When Data Are Affected by Correlated and Uncorrelated Classification Errors." WORC Paper No. 95.12, Tilburg University.

Becker, Mark P. and Ilsoon Yang. 1998. "Latent Class Marginal Models for Cross-Classifications of Counts." Pp. 293-326 in *Sociological Methodology 1998*, edited by Adrian E. Raftery. Washington, DC: American Sociological Association.

Bergsma, Wicher P. 1997. *Marginal Models for Categorical Data*. Tilburg: Tilburg University Press.

Biemer, Paul P. and Dennis Trewin. 1997. "A Review of Measurement Error Effects on the Analysis of Survey Data." Pp. 603-32 in *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: Wiley.

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

Bollen, Kenneth A. 1989. *Structural Equations With Latent Variables*. New York: Wiley.

Bryk, Anthony S. and Stephen W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Newbury Park, CA: Sage.

Burkhead, Dan and John Coder. 1985. "Gross Changes in Income Recipiency From the Survey of Income and Program Participation." Pp. 351-56 in *Proceedings of the Social Statistics Section*. Washington, DC: American Statistical Association.

Chua, Tin C. and Wayne A. Fuller. 1987. "A Model for Multinomial Response Error Applied to Labor Flows." *Journal of the American Statistical Association* 82:46-51.

Citro, C. F. and Graham Kalton. 1993. *The Future of the SIPP*. Washington, DC: National Academy Press.

Clogg, Clifford C. 1981. "New Developments in Latent Structure Analysis." Pp. 215-46 in *Factor Analysis and Measurement in Sociological Research*, edited by David J. Jackson and Edgar F. Borgatta. Beverly Hills, CA: Sage.

Collins, Linda M., Penny L. Fidler, and Stuart E. Wugalter. 1996. "Some Practical Issues Related to the Estimation of Latent Class and Latent Transition Parameters." Pp. 133-46 in *Categorical Variables in Developmental Research: Methods of Analysis*, edited by Alexander von Eye and Clifford C. Clogg. San Diego, CA: Academic Press.

Collins, Linda M., Penny L. Fidler, Stuart E. Wugalter, and Jeffrey D. Long. 1993. "Goodness-of-Fit Testing for Latent Class Models." *Multivariate Behavioral Research* 28:375-89.

Collins, Linda M. and Stuart E. Wugalter. 1992. "Latent Class Models for Stage-Sequential Dynamic Latent Variables." *Multivariate Behavioral Research* 27:131-57.

Cox, D. R. and Nanny Wermuth. 1996. *Multivariate Dependencies: Models, Analysis and Interpretations*. London: Chapman & Hall.

Davidson, Russell and James G. MacKinnon. 1993. *Estimation and Inference in Econometrics*. New York: Oxford University Press.

De Leeuw, Jan, Peter G. Van der Heijden, and Peter Verboon. 1990. "A Latent Time-Budget Model." *Statistica Neerlandica* 44:1-22.

DeSarbo, Wayne S. and Michel Wedel. 1993. *A Review of Recent Developments in Latent Class Regression Models*. Research Memorandum No. 521. Groningen: University of Groningen.

Diggle, Peter J., Kung-Yee Liang, and Scott L. Zeger. 1996. *Analysis of Longitudinal Data*. Oxford: Clarendon.

Evers, Mark and N. Krishnan Namboodiri. 1978. "On the Design Matrix Strategy in the Analysis of Categorical Data." Pp. 86-111 in *Sociological Methodology 1979*, edited by Karl F. Schuessler. San Francisco: Jossey-Bass.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 1995. *Bayesian Data Analysis*. London: Chapman & Hall.

Glymour, Clark, Richard Scheines, Peter Spirtes, and Kevin Kelly. 1987. *Discovering Causal Structure*. Orlando, FL: Academic Press.

Goodman, Leo A. 1973. "The Analysis of a Multidimensional Contingency Table When Some Variables Are Posterior to the Others." *Biometrika* 60:179-92.

———. 1974a. "Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models." *Biometrika* 61:215-31.

———. 1974b. "The Analysis of Systems of Qualitative Variables When Some of the Variables Are Unobservable: Part I. A Modified Latent Structure Approach." *American Journal of Sociology* 79:1179-259.

Haberman, Shelby J. 1978. *Analysis of Qualitative Data: Vol. 1. Introductory Topics*. New York: Academic Press.

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

Hadgu, Alula and Yinsheng Qu. 1998. "A Biomedical Application of Latent Class Models With Random Effects." *Applied Statistics* 47:603-16.

Hagenaars, Jacques A. 1988. "Latent Structure Models With Direct Effects Between the Indicators, Local Dependence Models." *Sociological Methods & Research* 16:379-405.

———. 1990. *Categorical Longitudinal Data Log-Linear Panel, Trend and Cohort Analysis*. Newbury Park, CA: Sage.

———. 1993. *Loglinear Models With Latent Variables*. Newbury Park, CA: Sage.

———. 1994. "Latent Variables in Log-Linear Models of Repeated Observations." Pp. 329-52 in *Latent Variable Analysis: Applications for Developmental Research*, edited by Alexander von Eye and Clifford C. Clogg. Thousand Oaks, CA: Sage.

———. 1998. "Categorical Causal Modeling: Latent Class Analysis and Directed Log-Linear Models With Latent Variables." *Sociological Methods & Research* 26:436-87.

Heckman, James J. and Burton Singer. 1982. "Population Heterogeneity in Demographic Models." Pp. 567-99 in *Multidimensional Mathematical Demography*, edited by Kenneth Land and Andrei Rogers. New York: Academic Press.

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

Heise, David R. 1969. "Separating Reliability and Stability in Test-Retest Correlation." *American Sociological Review* 34:93-101.

Hubble, Dan L. and David R. Judkins. 1987. "Measuring the Bias in Gross Flows in the Presence of Autocorrelated Measurement Error." Survey of Income and Program Participation Working Paper No. 8712.

Hughes, James P., Peter Guttorp, and Stephen P. Charles. 1999. "A Non-Homogeneous Hidden Markov Model for Precipitation Occurrence." *Applied Statistics* 48:15-30.

Juang, B. H. and L. R. Rabiner. 1991. "Hidden Markov Models for Speech Recognition." *Technometrics* 33:251-72.

Kalton, Graham and C. F. Citro. 1994. "Panel Surveys: Adding the Fourth Dimension." *Survey Methodology* 19:205-15.

Kasprzyk, Daniel, Greg J. Duncan, Graham Kalton, and M. P. Singh. 1989. *Panel Surveys*. New York: Wiley.

Kendall, Patricia. 1954. *Conflict and Mood: Factors Affecting Stability and Response*. New York: Free Press.

Kuha, Jouni and Chris Skinner. 1997. "Categorical Data and Misclassification." Pp. 633-70 in *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith de Leeuw, Cathryn Dippo, Norbert Schwarz, and Dennis Trewin. New York: Wiley.

Lang, Joseph B. and Alan Agresti. 1994. "Simultaneous Modeling Joint and Marginal Distributions of Multivariate Categorical Responses." *Journal of the American Statistical Association* 89:625-32.

Langeheine, Rolf, Jeroen Pannekoek, and Frank Van de Pol. 1996. "Bootstrapping Goodness-of-Fit Measures in Categorical Data Analysis." *Sociological Methods & Research* 24:492-516.

Lauritzen, Steffen L. 1996. *Graphical Models*. Oxford: Clarendon.

Lazarsfeld, Paul F. 1972. "The Problem of Measuring Turnover." Pp. 115-25 in *Continuities in the Language of Social Research*, edited by Paul F. Lazarsfeld, Ann K. Pasanella, and Morris Rosenberg. New York: Free Press.

Lazarsfeld, Paul F. and Neil W. Henry. 1968. *Latent Structure Analysis*. Boston: Houghton Mifflin.

Lazarsfeld, Paul F. and Morris Rosenberg. 1955. *The Language of Social Research*. Glencoe, IL: Free Press.

Lemaitre, Georges E. 1988. "The Measurement and Analysis of Gross Flows." Working Paper Series No. SSMD-88-1-E, Statistics Canada.

Lord, Frederic M., and Melvin R. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Maccoby, Eleanor E. 1956. "Pitfalls in the Analysis of Panel Data: A Research Note on Some Technical Aspects of Voting." *American Journal of Sociology* 61:359-62.

Martini, Alberto. 1988. "Retrospective Versus Panel Data in Estimating Labor Force Gross Flows: Comparing SIPP and CPS." Paper presented at the annual meeting of the American Statistical Association, Social Science Section, August 22-25, New Orleans.

———. 1989. "Seam Effect, Recall Bias and the Estimation of Labor Force Transition Rates From SIPP." Pp. 387-92 in *Proceedings of the Section of Survey Research Methods*. Washington, DC: American Statistical Association.

Mooijaart, Ab and Peter G. M. Van der Heijden. 1992. "The EM Algorithm for Latent Class Analysis With Equality Constraints." *Psychometrika* 57:261-69.

O'Muircheartaigh, Colm. 1996. "Measurement Errors in Panel Surveys: Implications for Survey Design and for Survey Instruments." Pp. 207-18 in *Proceedings of the Scientific Reunion of the Italian Statistical Society.* Rimini: Maggioli.

Pearl, Judea. 1995. "Causal Diagrams for Empirical Research." *Biometrika* 82:669-710.

Plewis, Ian. 1985. *Analysing Change: Measurement and Explanation Using Longitudinal Data*. Chichester, UK: Wiley.

Poterba, James M. and Lawrence H. Summers. 1986. "Reporting Errors and Labor Market Dynamics." *Econometrica* 54:1319-38.

Poulsen, Chris S. 1982. "Latent Structure Analysis With Choice Modeling Applications." Ph.D. dissertation, Wharton School, University of Pennsylvania.

Qu, Y., M. Tan and M. H. Kutner. 1996. "Random Effects Models in Latent Class Analysis for Evaluating Accuracy of Diagnostic Tests." *Biometrics* 52:797-810.

Raftery, Adrian E. 1993. "Bayesian Model Selection in Structural Equation Models." Pp. 163-80 in *Testing Structural Equation Models*, edited by Kenneth A. Bollen and J. Scott Long. Newbury Park, CA: Sage.

———. 1995. "Bayesian Model Selection in Social Research." Pp. 111-64 in *Sociological Methodology 1995*, edited by Peter V. Marsden. Washington, DC: American Sociological Association.

———, ed. 1998. "Special Issue: Causality in the Social Sciences: In Honor of Herbert L. Costner." *Sociological Methods & Research* 27:139-348.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66:688-701.

Rubin, Donald B. and Hal S. Stern. 1994. "Testing in Latent Class Models Using a Posterior Predictive Check Distribution." Pp. 420-38 in *Latent Variables Analysis: Applications for Developmental Research*, edited by Alexander von Eye and Clifford C. Clogg. Thousand Oaks, CA: Sage.

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

Skinner, Chris and Nicola Torelli. 1993. "Measurement Error and the Estimation of Gross Flows From Longitudinal Economic Data." *Statistica* 3:391-405.

Sobel, Michael E. 1994. "Causal Inference in Latent Variable Models." Pp. 3-35 in *Latent Variables Analysis: Applications for Developmental Research*, edited by Alexander von Eye and Clifford C. Clogg. Thousand Oaks, CA: Sage.

———. 1995. "Causal Inferences in the Social and Behavioral Sciences." Pp. 1-38 in *Handbook of Statistical Modeling for the Social and Behavioral Sciences*, edited by Gerhard Arminger, Clifford C. Clogg, and Michael E. Sobel. New York: Plenum.

Sutcliffe, J. P. 1965a. "A Probability Model for Errors of Classification: I. General Considerations." *Psychometrika* 30:73-96.

———. 1965b. "A Probability Model for Errors of Classification: II. Particular Cases." *Psychometrika* 30:129-55.

U.S. Department of Commerce. 1991. *SIPP's User Guide*. Washington, DC: U.S. Department of Commerce.

Van de Pol, Frank and Rolf Langeheine. 1990. "Mixed Markov Latent Class Models." Pp. 213-47 in *Sociological Methodology*, edited by C. Clogg. New York: Blackwell.

Van der Heijden, Peter, Harm 't Hart, and Jos Dessens. 1997. "A Parametric Bootstrap Procedure to Perform Statistical Tests in a LCA of Anti-Social Behavior." Pp. 196-208 in *Applications of Latent Trait and Latent Class Models in the Social Sciences*, edited by Jürgen Rost and Rolf Langeheine. New York: Waxmann.

Vermunt, Jeroen K. 1996. *Loglinear Event History Analysis: A General Approach With Missing Data, Latent Variables, and Unobserved Heterogeneity.* Tilburg: Tilburg University Press.
———. 1997a. "EM: A General Program for the Analysis of Categorical Data." WORC paper, Tilburg University.
———. 1997b. *Loglinear Models for Event History Analysis*. Thousand Oaks, CA: Sage.
Weakliem, David L. 1999. "A Critique of the Bayesian Information Criterion for Model Selection [With Discussion]." *Sociological Methods & Research* 27:359-97.
Whittaker, Joe. 1990. *Graphical Models in Applied Multivariate Statistics*. New York: Wiley.
Wiggins, Lee M. 1955. *Mathematical Model for the Analysis of Multi-Wave Panels*. Ann Arbor, MI: University Microfilms.
———. 1973. *Panel Analysis: Latent Probability Models for Attitude and Behavior Change*. Amsterdam: Elsevier.
Young, Nathan. 1989. "Wave Seam Effect in SIPP." Pp. 393-98 in *Proceedings of the Survey Research Section*. Washington, DC: American Statistical Association.

Francesca Bassi is a researcher in the Statistics Department at the University of Padova, Italy. Her research interests are in measurement errors and gross flows estimation in longitudinal data; statistical models to describe economic behavior, in particular log-linear and latent class models; and the measurement and analysis of labor and unemployment.

Jacques A. Hagenaars is a professor in the Methodology Department at Tilburg University. His research interests are in social statistics, causal models, and research methodology. He is the author of Categorical Longitudinal Data *and* Log-Linear Models With Latent Variables.

Marcel A. Croon is an associate professor in the Methodology Department of the Faculty of Social Sciences at Tilburg University. His research interests are in applied statistics, measurement theory, and research methodology.

Jeroen K. Vermunt is an assistant professor in the Methodology Department of the Faculty of Social and Behavioral Sciences at Tilburg University, as well as a research associate at the Work and Organization Research Center. He specializes in the analysis of categorical data and latent class analysis.