J. R. Statist. Soc. A (2002)
165, *Part* 1, *pp.* 97–119

Measurement error evaluation of self-reported drug use: a latent class analysis of the US National Household Survey on Drug Abuse

Paul P. Biemer and Christopher Wiesen

Research Triangle Institute, Research Triangle Park, USA

[Received May 2000. Revised May 2001]

Summary. Latent class analysis (LCA) is a statistical tool for evaluating the error in categorical data when two or more repeated measurements of the same survey variable are available. This paper illustrates an application of LCA for evaluating the error in self-reports of drug use using data from the 1994, 1995 and 1996 implementations of the US National Household Survey on Drug Abuse. In our application, the LCA approach is used for estimating classification errors which in turn leads to identifying problems with the questionnaire and adjusting estimates of prevalence of drug use for classification error bias. Some problems in using LCA when the indicators of the use of a particular drug are embedded in a single survey questionnaire, as in the National Household Survey on Drug Abuse, are also discussed.

Keywords: Classification error; Marijuana use; Non-sampling error; Survey error; Test–retest

1. Introduction

Survey data are subject to measurement errors from numerous sources including poor design of questionnaires, difficult and misunderstood concepts, inadequately trained interviewers, deliberate errors and the interview mode or setting. Measurement errors can have severe biasing effects on the estimation and data analysis and will often reduce the precision of the estimates and the power of statistical tests (Biemer and Trewin, 1997). The evaluation of measurement errors in US surveys has become an important goal of national statistical agencies (Martin and Straf, 1992) as well as non-Government survey research organizations.

The evaluation of measurement error is critical because

(a) data users need to understand the limitations of the data to interpret the survey results appropriately,
(b) survey designers use information on the quality of data to guide improvements of the survey process for future surveys,
(c) data analysts incorporate data on the non-sampling error distributions to reduce bias and to increase the power of their analyses (see, for example, Fuller (1991)) and
(d) survey practitioners rely on survey error evaluations for the quality assurance of field operations.

Traditional methods for assessing survey bias share an important shortcoming in that they rely on the accuracy of so-called gold standard measurements. Gold standard measurements

Address for correspondence: Paul P. Biemer, Research Triangle Institute, PO Box 12194, Research Triangle Park, NC 27606-2194, USA.
E-mail: ppb@rti.org

© 2002 Royal Statistical Society 0964–1998/02/165097

include reconciled reinterview responses, administrative record values, biological test results, diary entries and other measurements which are assumed to contain little or no non-sampling error. However, there is a growing literature that suggests that, in most situations, gold standard measurements are quite fallible. For example, in reconciled reinterviews, more experienced interviewers reinterview a sample of original interview respondents to identify discrepancies between the original and reinterview reports. These discrepancies are then reconciled to obtain revised, and presumably more correct, reports. However, recent work suggests that reconciled reinterview data can be as erroneous as the original measurements that they are intended to evaluate (see, for example, Biemer and Forsman (1992) and Biemer *et al.* (2001)). Administrative records data are often inaccurate and difficult to use (Jay *et al.*, 1994; Marquis, 1978) as a result of differing time reference periods, definitional differences and errors in the records themselves. Further, the collection of gold standard data is usually quite costly and complex (Biemer, 1988).

An alternative to a gold standard analysis involves the application of statistical models to replicate measurements of the same survey variable. For categorical data items for which two replicate measurements are available, the statistical model proposed by Hui and Walter (1980) can be applied to estimate the classification error probabilities in the measurements. This method requires that the population be partitioned into two groups (e.g. smokers and non-smokers) having different prevalence rates for the true characteristic but measurement error distributions that do not vary by group (this is referred to as the homogeneous error assumption). Under these assumptions, all parameters of the Hui–Walter model are identifiable; however, no degrees of freedom are available for testing the lack of fit. Biemer and Witt (1996) applied this model to estimate the classification error in reports of lifetime use of marijuana, cocaine and alcohol in the US National Household Survey on Drug Abuse (NHSDA).

A drawback of the Hui–Walter model is that standard Pearson or likelihood ratio χ^2 goodness-of-fit tests cannot be applied with this approach since the model is saturated. Further, identifying grouping variables for which the homogeneous assumption is plausible is difficult in many situations. However, when three replicate measurements (or indicators) are available, the additional degrees of freedom allow the homogeneous error assumptions to be relaxed and classification error parameters that vary by group are identifiable. In addition, with three indicators the residual degrees of freedom are often adequate for testing model fit for a wide range of models.

Of course, one difficulty is that three replicate measurements of the same survey variable are difficult to obtain in practice. A common approach for obtaining two replicate measurements is to use test–retest or reinterview methods that require recontacts of respondents. However, contacting respondents to obtain three replicates is often not practical because of the burden on respondents and their resistance, the risk of response conditioning effects by prior contacts and the costs that are associated with repeated contacts with the respondent.

The present paper considers the analysis of replicate measurements which are embedded in a single survey instrument and that can be obtained in a single interview. Since the replicate measurements are collected during the same interview, the risk that errors made for one measurement are correlated with the errors of the other two measurements must be considered in the analysis. Ideally, the wording of the replicate items should differ somewhat to conceal the redundancy from the respondent to avoid resistance by the respondents to the burden of answering the same questions repeatedly. Therefore, the indicators may not be parallel measures (i.e. measures having the same error distributions), which further complicates the analysis by introducing additional error parameters into the model.

However, in many situations, embedding replicate measurements is feasible for obtaining repeated observations for a limited number of items in the questionnaire. Our analysis will demonstrate how correlated errors and departures from parallel measures can be modelled and accounted for in the latent class analysis (LCA) estimates of measurement error. In the present study, we analyse data from the NHSDA to demonstrate the analytical technique for estimating classification errors in surveys having embedded replicate measurements. Our analysis shows that LCA can be an important tool for identifying the sources of measurement errors in such surveys and for adjusting survey estimates for measurement error bias by using embedded indicators.

2. Statistical framework

We begin by developing the essential analytic tools for estimating classification errors in the case of three embedded replicate measurements. Extensions to more than three indicators are straightforward and will not be discussed here. Initially, we assume simple random sampling from the target population and later extend the methodology to more complex survey designs. In Section 3, this methodology will be applied to the NHSDA and the results will be evaluated and discussed.

2.1. Model notation and assumptions

Let S denote a simple random sample of size n from a large population and let i denote a particular unit in S. We are interested in evaluating the classification error associated with a categorical response variable measured for the members of S. Let X_i, which is assumed to be an unobservable or latent variable, denote the true value of the characteristic for unit i. Let A_i , B_i and C_i denote three repeated observations (or indicators) of the characteristic for unit i. For example, A_i may denote the response to question A in the survey, B_i the response to question B and C_i the response to question C, where all three questions are designed to measure the variable X_i but perhaps by different methods.

In the following, we shall assume that X_i is a dichotomous variable and that A_i, B_i and C_i are dichotomous indicators of X_i. Thus, X_i, A_i, B_i and C_i take the values 1 or 2 for all units in S. Extensions to $k > 2$ categories for X_i, A_i, B_i and C_i do not afford any difficulties but will not be considered here. For ease of notation, the subscript i is dropped in the following.

Let π_z, for arbitrary random variable Z, denote $\Pr(Z = z)$. Thus, π_x denotes $\Pr(X = x)$, π_{xa} denotes $\Pr(X = x, A = a)$, $\pi_{a|x}$ denotes $\Pr(A = a|X = x)$ and so on. Note that the classification error probabilities for the indicator A are $\pi_{a|x}$ for $a \neq x$. For example, $\pi_{a=1|x=2}$ is the false positive probability for A and $\pi_{a=2|x=1}$ is the false negative probability. Likewise, $\pi_{b=1|x=2}$ and $\pi_{b=2|x=1}$ are the false positive and false negative probabilities respectively for B, and $\pi_{c=1|x=2}$ and $\pi_{c=2|x=1}$ are the false positive and false negative probabilities respectively for C. Hence the probabilities $\pi_{a|x}$, $\pi_{b|x}$ and $\pi_{c|x}$ are often referred to as error probabilities. It is tacitly assumed that the error probabilities are the same for all members of the population. Later we shall introduce covariates and grouping (or stratification) variables into the model to account for heterogeneous error probabilities in the population.

Let $XABC$ denote the cross-classification table for the variables X, A, B and C and let (x, a, b, c) denote the cell associated with $X = x$, $A = a$, $B = b$ and $C = c$. Thus, π_{xabc} is the expected proportion in cell (x, a, b, c) of the table and can be rewritten as

$$\begin{aligned}\pi_{xabc} &= P(X = x)\,P(A = a|X = x)\,P(B = b|A = a, X = x)\,P(C = c|A = a,\ B = b, X = x)\\ &= \pi_x\,\pi_{a|x}\,\pi_{b|ax}\,\pi_{c|abx} \qquad (1)\end{aligned}$$

and, hence, the probability that a unit is classified in cell (a, b, c) is

$$\pi_{abc} = \sum_{x} \pi_x \, \pi_{a|x} \, \pi_{b|ax} \, \pi_{c|abx}. \tag{2}$$

Let n_{abc} denote the number of observations in cell (a, b, c), where

$$n = \sum_{a,b,c} n_{abc},$$

and assume that the n_{abc} are distributed as a set of multinomial random variables. Then the kernel of the likelihood of observing the table ABC is

$$L(ABC) = \prod_{a} \prod_{b} \prod_{c} \pi_{abc}^{n_{abc}}. \tag{3}$$

Note that there are eight (i.e. 2^3) cells in the observed table whereas the likelihood contains 15 model parameters plus an additional parameter for the overall mean. A necessary condition for the model parameters to be identifiable (or estimable) is that the degrees of freedom for the model (the number of cells minus the number of parameters including the overall mean) are not negative. By this condition, equation (1) is not identifiable. To overcome this problem, restrictions on the probabilities can be introduced to reduce the number of parameters associated with the model. The traditional latent class model for the table $XABC$ introduces the restrictions $\pi_{b|ax} = \pi_{b|x}$ and $\pi_{c|abx} = \pi_{c|x}$ which eliminate two and six parameters respectively. This assumption, called *local independence* in the LCA literature (see, for example, McCutcheon (1987)), is also quite common for estimating test–retest reliability (see, for example Bohrnstedt (1983)). It specifies that errors in the indicators A, B and C are mutually independent. With these restrictions, the number of model parameters is reduced to 7 and the model is identifiable. However, as we discuss later, having non-negative degrees of freedom is generally not sufficient for model identifiability with latent class models.

The local independence assumption is more plausible when the three indicators are obtained in three separate interviews rather than within the same interview. However, even in an interview–reinterview–reinterview situation, independent classification error is not guaranteed (see Biemer *et al.* (2001) for an example of this). For example, respondents may have repeated their erroneous responses across interview occasions either from memory or from replicating the response process that led to the original classification error.

When measurements are embedded in the same questionnaire, the risk of correlated error is greater owing to within-interview factors and increased memory effects. One means of counteracting these effects is to vary the method for measuring the underlying construct, for example, by altering the wording of the question while maintaining its original meaning and intent (see, for example, Saris and Andrews (1991)) and by separating the questions in time within the interview to the extent possible. Despite these precautions, the potential for correlated error must still be considered in the analysis of the repeated measures data.

Since local independence models with three measurements are saturated models, models which introduce additional terms for correlated error (so-called local dependence models) are not identifiable unless further restrictions are place on the model (Hagenaars, 1988). For example, by imposing the restrictions $\pi_{a|x} = \pi_{b|x} = \pi_{c|x}$—i.e. the classification error distributions for A, B and C are identical—2 degrees of freedom are saved which can be used to estimate the two additional parameters introduced by relaxing the independence assumption for two of the three indicators, e.g. $\pi_{b|ax}$. However, this equal error probability restriction is not plausible and is likely to be violated if the method (i.e. question wordings) for obtaining A, B and C varies within the questionnaire.

Another technique for increasing the model degrees of freedom is to introduce a grouping variable G having L levels. Now, the number of cells of the $GABC$-table is L times the number of cells in the ABC-table. Equating some parameters of the model across the L groups to free enough degrees of freedom for estimating the correlated error parameters often results in more plausible assumptions for the model than are possible without the grouping variable.

For the case of two groups, say $G = 1$ for males and $G = 2$ for females, the $GABC$-table has a total of 2×2^3 or 16 cells. Denoting the conditional classification probabilities for group g by $\pi_{a|gx}$, $\pi_{b|gx}$ and $\pi_{c|gx}$, we assume that

$$\pi_{\alpha|g=1,x} = \pi_{\alpha|g=2,x} = \pi_{\alpha|x} \tag{4}$$

for $\alpha = a, b, c$, i.e., for each indicator, the classification error probabilities for males and females are equal (referred to as group homogeneity). We further assume that the three measurements are correlated (Hagenaars, 1988) and introduce terms $\pi_{b|ax}$ and $\pi_{c|bx}$ to model the correlation. Thus, we assume that the joint error distributions, like the marginal distributions, are homogeneous across groups and that the three-way interactions between the indicators given the true value of the characteristic are 0, i.e. $\pi_{c|abx} = \pi_{c|bx.}$

For three indicators embedded in a single questionnaire, it seems plausible to assume a causal ordering of errors in the indicators that reflects the temporal ordering of the indicators in the interview, i.e. we assume local dependence between chronologically adjacent indicators in the NHSDA questionnaire so that the error in B depends on A and the error in C depends on B, but the error in C conditional on B does not depend on A. Thus, the probability of an observation in cell (g, a, b, c) of the $GABC$-table is

$$\pi_{gabc} = \pi_g \sum_x \pi_{x|g}\, \pi_{a|x}\, \pi_{b|ax}\, \pi_{c|bx}\,, \tag{5}$$

which yields a likelihood for $GABC$ with 13 model parameters (plus two parameters for the overall means for the two groups), leaving 1 degree of freedom to assess the fit of the model.

These ideas can be extended in various ways, some of which will be explored in the next section. Additional grouping variables can be added to the model which may be desirable, not only to provide additional degrees of freedom for estimation of the error distributions, but also to capture the heterogeneity of response errors across various population subgroups. As the number of grouping variables in the analysis increases, a greater range of model assumptions can be explored that reflect the interrelationships between the latent variable, the indicators and the subgroups.

In the next section, we draw on the work of Haberman, Goodman and others who have provided a convenient structure for exploring alternative probability models in the context of a log-linear analysis with latent variables.

2.2. Relationship of latent probability models and log-linear models

Haberman (1979) showed that latent probability models like equation (5) can be reformulated as log-linear models with latent variables. Thus, much of the statistical theory that has been developed for log-linear analysis can be directly applied to latent structure analysis. Goodman (1973) showed that the model for the cell probabilities in equation (5), including the latent variable classification, can be written as

$$\begin{aligned}\log(n_{xgabc}) = u + u_g^G + u_x^X + u_a^A + u_b^B + u_c^C + u_{gx}^{GX} + u_{xa}^{XA} + u_{xb}^{XB} \\ + u_{xc}^{XC} + u_{ab}^{AB} + u_{bc}^{BC} + u_{xab}^{XAB} + u_{xbc}^{XBC}\end{aligned} \tag{6}$$

where $n_{xgab} = n\pi_{xgab}$ for n the number of observations. This is a hierarchical linear model owing to the inclusion of all lower order interactions involving variables for the highest order interactions. Correlated errors are modelled by the interactions terms involving two or more repeated measures. Model (6) can be represented in shorthand notation as $\{XG, XAB, XBC\}$. In this notation, only the highest order terms involving each variable in the model are shown within the braces. The lower order terms are all implicitly included by the hierarchical model structure. Methods for testing the fit of model (6) to the $XGABC$ contingency table have been developed that parallel those used for ordinary log-linear models. The primary difference is that, since X is not observed, the fit of the latent log-linear model is assessed through the χ^2 goodness-of-fit criterion applied to the observed table only, i.e. $GABC$. (See Hagenaars (1990, 1993) for further details.)

In the next section, we shall illustrate the analysis of classification errors using embedded repeated measurements and log-linear models with latent variables for data from the NHSDA.

3. Illustration using the National Household Survey on Drug Abuse

3.1. Description of the data

The NHSDA is a multistage household survey designed to measure the population's current and previous drug use activities. The 1996 survey was the 16th study conducted in a series initiated in 1971. Since 1990, the survey has been conducted annually, with independent samples of households and people selected each time. For this study, data from the 1994, 1995 and 1996 surveys were used in the analysis: a total of 53 825 interviews. The subsequent description of the NHSDA will be restricted to design and implementation issues related to these surveys.

3.2. Survey design and data collection

The NHSDA design is a stratified, multistage cluster sample of dwelling units selected in approximately 127 primary sampling units in 1994 and 115 primary sampling units in 1995 and 1996. The primary sampling units represent geographic areas in the USA, generally defined as counties, groups of counties or metropolitan statistical areas. The target population includes people who are 12 years old or older who live in households, certain group quarters (e.g. college dormitories and homeless shelters) and civilians living on military installations. Active military personnel and most transient populations, such as homeless people not residing in shelters, are not included. The annual sample sizes for the 1994, 1995 and 1996 surveys are provided in Table 1. Hispanics, blacks, younger people and the residents of certain metropolitan statistical areas are oversampled to ensure that the sample sizes are adequate to produce the subpopulation estimates that are of interest.

The non-response rates for the NHSDA were 19%, 20% and 21% for 1994, 1995 and 1996 respectively. To reduce non-response bias in the estimates of drug use, the NHSDA incorporates weight adjustments based on age, race and sex cells. The analysis to follow is based on cell counts that have been weighted for unequal probabilities of selection and rescaled to the original sample size. Since the inferential population for our study is the population of NHSDA respondents, the NHSDA non-response adjustments were not used in our weighting.

The NHSDA interview process takes about an hour to complete and collects drug and demographic data from each respondent by using a combination of interviewer-administered and self-administered instruments. The interview begins with a set of inter-

Table 1. Sample sizes (number of households) for the 1994, 1995 and 1996 NHSDA by analysis domain

Domain	*Numbers of households for the following years:*		
	1994	*1995*	*1996*
Sex			
Males†	7950	7652	7774
Females	9859	10095	10495
Race			
Hispanic	4706	4599	4841
Black	4010	4208	4372
White or other	9093	8940	9056
Age (years)			
12–17	4698	4595	4538
18–25	3706	3963	4366
26–34	5223	5213	5262
35 and older	4182	3976	4103
Total sample size	17809	17747	18269

†The NHSDA sample contains slightly fewer males owing to the differential non-response rate by gender.

viewer-administered questions to collect data on basic demographic characteristics. The remainder of the questionnaire is divided into sections corresponding to each substance of interest: tobacco, alcohol, marijuana, cocaine, 'crack', hallucinogens, inhalants, analgesics, tranquilizers, stimulants and sedatives. For each section, the interviewers present the answer sheets to the respondents and ask them to record their responses on it. Depending on the complexity of an answer sheet, the interviewer will either read the questions to the respondent or, if preferred, respondents can read the questions themselves. On the completion of an answer sheet, the respondent is requested to place the answer sheet in an envelope without allowing the interviewer to see the responses. The motivation for conducting the interview in this manner is to ensure that the respondent understands the questions and does not erroneously skip over major parts of the questionnaire as well as to guarantee privacy for the respondent.

Most of the answer sheets are designed so that even respondents who have never used a particular drug will answer each question on the answer sheet. Since both users and non-users of a drug are asked to respond to essentially the same number of questions, the interviewer is less likely to guess that the respondent is a user or non-user on the basis of the time that the respondent takes to complete an answer sheet. This is another feature of the survey that is designed to protect the privacy of the respondent. In addition, some respondents who indicate that they never used the drug under direct questioning will later answer an indirect question about the drug in a way that implies use of the drug (as shown in Table 2). This redundancy in the questionnaire provides the basis for constructing three repeated measures for estimating accuracy of reporting on drug use using the LCA discussed in the previous section.

3.3. Repeated measures of drug use

The definitions of the three indicators in terms of specific NHSDA questions appear in the exhibits in Appendix A. For each of the three years, indicator A is the response to the

Table 2. Observed inconsistencies in the three indicators of use of marijuana

Indicator	*Results for the following years:*					
	1994		*1995*		*1996*	
	n	*%*	*n*	*%*	*n*	*%*
A versus B	241	1.35	263	1.48	293	1.61
A versus C	854	4.80	380	2.14	452	2.48
B versus C	883	4.96	409	2.31	491	2.69
A versus B versus C	989	5.55	526	2.96	618	3.39

so-called recency of use (or recency) question which asks about the length of time since marijuana or hashish was last used. Also, for all three years, indicator B is the response to the so-called frequency of use (or frequency) question which asks how frequently, if ever, the respondent has used marijuana or hashish in the past year. Indicator C is a composite of several questions (seven in 1994 and eight in 1995 and 1996) on the so-called drug answer sheet. An affirmative response to any one of these is coded as 'yes' for C and otherwise C is coded as 'no'. Note from Appendix A that in 1994 this composite question included question 7 which required that the respondent responds 'yes' or 'no' to two questions:

(a) whether you wanted to cut down on the use of marijuana or hashish and
(b) whether you were able to cut down.

As we shall see later, this question proved to be quite difficult for respondents. In 1995 and 1996, question 7 was replaced by two very different questions (see Appendix A). In addition, questions 5 and 6 were modified from the 1994 version.

Clearly, indicators A, B and C appear to satisfy the goal of embedding repeated measurements in a single instrument in that the methods used to measure the underlying construct are varied. For this research, interest lies primarily in estimating the false positive and false negative probabilities separately for A, B and C and in comparing the LCA estimates of π_x, the true prevalence of use of marijuana in the past year, with the corresponding estimates from the NHSDA.

The classification variable used for the NHSDA official estimates of use of marijuana in the past year, which we shall denote by T, may be defined as $T = 1$ if either A or B is 1, and $T = 0$ otherwise, i.e. an individual is classified as a user in the past year for NHSDA estimation if either A or B indicates use in the past year. However, disagreement between A and B may also be the result of an error in the report of use in the past year (i.e. a false positive error). Since the NHSDA does not take into account the potential for false positive error (as the LCA estimates do), we expect that the LCA-derived estimates of π_x will be smaller than the corresponding NHSDA estimates.

For example, for indicators A and B in 1995, the estimated false positive error rates are 0.01% and 0.78% respectively, and the false negative error rate estimates are 8.96% and 0.90% respectively. The adjusted rate of use in the population from the LCA is 7.68%. If a positive response to either or both questions is used to indicate use, the estimated rate would be about 8.40% or about three-quarters of a percentage point larger than the LCA estimate.

3.4. Analysis of classification error for use of marijuana in the past year

3.4.1. Evidence of inconsistent reporting

It is well known that the various indicators of drug use in the NHSDA questionnaire are inconsistent (see, for example, Cox *et al.* (1992)). In Table 2 we show the rate of disagreement among all combinations of the three indicators. This rate is computed as the unweighted proportion of observations in the off-diagonal cells of the cross-classification table for the variables in each combination. Thus, Table 2 shows the extent of the inconsistencies by year among the three indicators. The disagreement rate for *A versus B* varies around 1.50% whereas the rate for *C versus A* or *B* is considerably higher, particularly in 1994 where the disagreement rate among all three measurements is 5.55%.

Since indicator *C* is a composite of many questions, one plausible hypothesis for a high rate of inconsistency with *C* is that *C* is the more accurate indicator (i.e. a gold standard) and disagreement with *C* is an indication of classification errors in the other two measurements. In support of this argument, estimates of use of marijuana in the past year are highest for *C*, which suggests greater accuracy since the use of marijuana tends to be underreported in the NHSDA (see, for example, Turner *et al.* (1992) and Mieczkowski (1991)). As we shall see, this hypothesis can be explored by using an LCA of the three measurements *A*, *B* and *C*.

3.4.2. Modelling the error in use of marijuana in the past year

In our analysis, rather than including a term in the model for the survey year and fitting a single model to the data pooled across years, separate LCA models were fitted to each year of the NHSDA. This was done for various reasons. First, we can use the comparison of the best fitting models for each year as an indicator of validity of the model. Since there were no major design changes for the NHSDA from 1994 to 1996, the error structure in the data should be the same for all three years. Thus, the failure of a well fitting model from one year to fit the data from another year would be evidence of an invalid model. Second, a separate analysis facilitated the interpretation of comparisons of the LCA estimates with the official NHSDA estimates which are also computed separately for each year. Third, the indicator *C* was different in 1994 from that in the other two years so this complexity would have to be handled somehow in the pooled data models. Finally, we have encountered some convergence problems with LCA estimation software when the number of parameters in the models is quite large. For the pooled data set, the models would have contained up to 400 parameters. Although some loss of precision in the estimation will result by using a separate analysis for each year, this is not a problem given the large data sets that we are analysing for each year.

Using the three indicators of use of marijuana in the past year, a wide range of latent class models was explored to identify the 'best' model for producing estimates of classification error. Following Lin and Dayton (1997), three criteria were applied to select the best model as follows.

(a) The model should be identifiable.
(b) The likelihood ratio χ^2 p-value for the model should be greater than 0.01, indicating that the model fits the data reasonably well.
(c) The Bayesian information criterion BIC defined as $L^2 - \log(N)$ degrees of freedom should be the smallest among all competing models.

Identifiability of the models was verified by using the sufficient condition suggested by Dayton and Macready (1980) that the variance–covariance matrix for the parameters should be of full rank. Note that the p-value criterion in equation (2) is more liberal than the

minimum 0.05 *p*-value that is typically used for model selection. This is because, with over 18000 observations for each year, the power of the χ^2-test using a *p*-value of 0.05 is quite high, which could result in overparameterization of the model. The less stringent 0.01 *p*-value allows a consideration of models with expected cell probabilities that may only differ trivially from the observed data while satisfying the other selection criteria. Criterion (c) provides for the most parsimonious model that fits the data. Lin and Dayton (1997) found that this criterion is most appropriate for comparing models when large sample sizes are involved as in the present case.

The types of model considered in our analysis were limited to simple extensions of the basic latent class model given by equation (4) with grouping variables defined by age (*G*), race (*R*) and sex (*S*). The fit statistics for equation (4) and its four extensions for all three years are provided in Table 3. Also provided in Table 3 is the fit of the same models to a revised version of the 1994 data that will be explained subsequently. The following is a description of the models that were used in the analysis.

Table 3. Model diagnostics for alternative classification error models by year

Model	*Degrees of freedom*	*Number of parameters*	L^2	*p-value*	*BIC*
1994 data					
Model 0: {*GRSX*, *AX*, *BX*, *CX*}	138	54	199.24	0.0005	−1151
Model 1: same as model 0 + {*AG*, *BG*, *CG*, *AR*, *BR*, *CR*, *AS*, *BS*, *CS*}	120	72	82.96	0.9960	−1092
Model 2: same as model 1 + {*AB*, *BC*}	118	74	72.11	0.9997	−1083
Model 3: {*GRSX*, *GAX*, *GBX*, *GCX*, *RAX*, *RBX*, *RCX*, *SAX*, *SBX*, *SCX*}	102	90	63.61	0.9990	−935
Model 4: same as model 3 + {*AB*, *BC*}	100	92	57.25	0.9998	−921
1995 data					
Model 0: {*GRSX*, *AX*, *BX*, *CX*}	138	54	252.20	0.0000	−1098
Model 1: same as model 0 + {*AG*, *BG*, *CG*, *AR*, *BR*, *CR*, *AS*, *BS*, *CS*}	120	72	90.38	0.9799	−1084
Model 2: same as model 1 + {*AB*, *BC*}	118	74	89.77	0.9753	−1065
Model 3: {*GRSX*, *GAX*, *GBX*, *GCX*, *RAX*, *RBX*, *RCX*, *SAX*, *SBX*, *SCX*}	102	90	51.42	1.0000	−947
Model 4: same as model 3 + {*AB*, *BC*}	100	92	51.22	1.0000	−927
1996 data					
Model 0: {*GRSX*, *AX*, *BX*, *CX*}	138	54	244.46	0.0000	−1110
Model 1: same as model 0 + {*AG*, *BG*, *CG*, *AR*, *BR*, *CR*, *AS*, *BS*, *CS*}	120	72	155.46	0.0163	−1022
Model 2: same as model 1 + {*AB*, *BC*}	118	74	139.83	0.0831	−1018
Model 3: {*GRSX*, *GAX*, *GBX*, *GCX*, *RAX*, *RBX*, *RCX*, *SAX*, *SBX*, *SCX*}	102	90	111.95	0.2354	−889
Model 4: same as model 3 + {*AB*, *BC*}	100	92	107.19	0.2933	−874
1994 revised data					
Model 0: {*GRSX*, *AX*, *BX*, *CX*}	138	54	212.08	0.0000	−1139
Model 1: same as model 0 + {*AG*, *BG*, *CG*, *AR*, *BR*, *CR*, *AS*, *BS*, *CS*}	120	72	105.00	0.8336	−1069
Model 2: same as model 1 + {*AB*, *BC*}	118	74	96.08	0.9308	−1059
Model 3: {*GRSX*, *GAX*, *GBX*, *GCX*, *RAX*, *RBX*, *RCX*, *SAX*, *SBX*, *SCX*}	102	90	75.27	0.9782	−923
Model 4: same as model 3 + {*AB*, *BC*}	100	92	64.95	0.9974	−914

(a) *Model 0*: $\{GRSX, AX, BX, CX\}$ is essentially equation (4) with the $GRSX$-effect representing the variation in the true prevalence rate by age, race and sex, and AX, BX and CX representing the classification error in the indicators which is assumed not to vary by age, race and sex.
(b) *Model 1*: the first-level extension of model 0 is the addition terms that reflect variation in the indicator error rates by age, race and sex. The simplest way to introduce this error variation is by the addition of nine two-factor interactions: AG, BG, CG, AR, BR, CR, AS, BS and CS.
(c) *Model 2*: at the next level of complexity (in terms of the number of parameters) terms that reflect possible correlations between the indicators (i.e. local dependence) are added. This can be accomplished by adding the interactions AB and BC to model 1. Note that the AC-interaction is missing, which suggests that the correlations follow a Markov process, i.e. the third indicator, C, does not depend on the first, A, when conditioned on the second, B.
(d) *Model 3*: $\{GRSX, GAX, GBX, GCX, RAX, RBX, RCX, SAX, SBX, SCX\}$ is a more complex representation of the variation in the indicator error rates by age, race and sex, extending the two-factor interactions of model 1 to three factors. However, the assumption of local independence is preserved.
(e) *Model 4*: replaces the local independence assumption of model 3 by a Markov correlated error assumption through the addition of terms AB and BC.

All the models contain the term $GRSX$ which is referred to as the 'structural' part of the model. It postulates that the use of marijuana in the past year, X, varies across the cells of the GRS-table. Although more parsimonious representations of the variation in X in the population may fit the data, our focus is on the terms of the model involving A, B and C, referred to as the 'measurement' part of the model. For modelling the measurement error, there is not much to be gained in trying to reduce the number of parameters used to describe the structure of X.

Model 0 represents the simplest error structure for the indicators A, B and C. It postulates that the error rates for A, B and C do not depend on age, race or sex. Skipping models 2 and 3 for the moment, model 4 postulates that error rates depend on the grouping variables. The terms GAX, GBX and GCX reflect the variation in the error terms AX, BX and CX by the levels of G. Likewise, RAX, RBX and RCX and SAX, SBX and SCX model the variation in the error terms by race and sex respectively.

Model 1 reflects a type of dependence of the error terms on the grouping variables but not the full dependence represented in model 4. To see this, consider the simple case where we wish to model $\pi_{a|xf}$, where the grouping variable F has two levels. It is shown in Appendix B that for any number of groups the model $\{AX, AF\}$ implies that the product of the odds of making a false positive error times the odds of making a false negative error are constant across the groups, i.e. $\phi_f\theta_f = \gamma_0$ where ϕ_f is the odds of a false positive result in group f, θ_f is the odds of a false negative result in group f and γ_0 is a constant. For this to occur requires that a group responding to an item with a higher false positive error probability also responds to the item with a lower probability of a false negative error and vice versa to maintain a constant $\phi_f\theta_f$ across the levels of the grouping variable F.

An indicator that always classifies an individual as a drug user would have $\theta_f = 0$ and $\phi_f = 1$. Likewise, an indicator that always classifies an individual as a non-user would have $\theta_f = 1$ and $\phi_f = 0$. Therefore, it is not uncommon for indicators to tend to classify respondents erroneously more in one direction than in the other.

Examples of an indicator where the pair (ϕ_f, θ_f) exhibits a negative correlation across groups are also encountered quite often in practice (see for example, Biemer *et al.* (2001)). For example, consider the question 'Have you used marijuana in the last 12 months?'. Interpretations of this question may vary in the population depending on factors such as education, age and cultural differences. As an example, younger respondents may tend to interpret the term 'use' incorrectly as 'regular use' whereas those in older age groups may feel (correctly) that any use, casual or otherwise, constitutes use of marijuana. In this case, younger age groups could exhibit a lower false positive error and higher false negative odds whereas the older age group could exhibit the opposite effect. Thus, the product of two error probability odds, $\theta_f \, \phi_f$, could still be constant across the two groups even though the error rates may vary across the groups.

Now, turning to models 2 and 4, the interpretation of the terms AB and BC has been discussed in detail in Hagenaars (1988). For example, consider the differences between two logistic models for the probability $\pi_{b|xa}^{B|XA}$: $\{BX\}$ and $\{BX, BA\}$. The former model postulates that the errors in A and B are locally independent. However, the residuals (i.e. the differences between the observed and predicted frequencies) from this model may indicate a tendency for positive residuals for the (1, 1) and (2, 2) cells of the AB-table and negative residuals for the (1, 2) and (2, 1) cells. This situation suggests a greater propensity for B to be positive or negative given that A is positive or negative respectively. This additional propensity is absorbed by the BA-term in the latter model.

The differences between the models considered in our study can be summarized as follows. Model 0 postulates that the false positive and false negative error probabilities do not vary by groups. Model 3 postulates essentially unconstrained variation in the error probabilities across groups and model 1 postulates variation in the error probabilities across groups while the product, $\phi_f \theta_f$, is constrained to be constant. Model 2 and model 4 add the local dependence assumption to models 1 and 3.

These five models were applied separately for each of the three years for the full NHSDA data set. The maximum likelihood solutions were estimated by using the lEM software (Vermunt, 1997) which employs an EM algorithm similar to that suggested by Goodman (1974). The model fitting process does not take into account the complex sampling design of the NHSDA with regard to clustering and unequal probability sampling effects. As a result, models are more easily rejected since the reported p-values are likely to be smaller than the actual p-values for the tests.

As seen from Table 3, model 0 does not describe the data for any year. Model 1 fits the data well, satisfies criterion (c) and is identifiable for all three years. Models 2 and 4, which add the correlated error terms, also fit the data very well; however, the models do not satisfy criterion (c). Moreover, the difference in L^2 (the likelihood ratio χ^2-statistic) between models 1 and 2 is significant for 1994 and 1996, whereas it is not significant for 1995. This is implausible since the surveys were essentially the same for all three years. Another important consideration is that interpreting the differences in the estimates among the years would be made easier by using the same model for all three years. Therefore, model 1 was accepted for all three years and this model was selected for the subsequent analysis.

Thus we see that, for the data for use of marijuana in the past year, the local dependence models did not fit the data as well as those postulating locally independent errors. This somewhat surprising result suggests that responses to the three indicators are not affected much by social desirability or other external factors that could influence the respondent to deny their drug use throughout the interview consciously and consistently. Nor is there strong evidence that the response processes which respondents used are sufficiently similar across the indicators

to produce errors that are significantly correlated. Rather, response errors for the indicators appear to behave independently and are more likely to arise from such factors as confusion generated by the wording of the questions or differences in the interpretation of the questions.

This latter explanation also may explain the presence of the indicator-by-grouping variable interaction terms in the best model. Problems of comprehension and interpretation of the questions may be related to the demographics of the respondents in such a way that false negative error rates vary inversely with the false positive error rates across the groups. For example, groups that adopt a broader interpretation of the question may tend to respond positively, and thus have a higher false positive error rate, whereas groups adopting a narrower interpretation may tend to respond negatively, resulting in a higher false negative error rate.

These patterns of response error may be quite different for questions about more stigmatized and strictly outlawed drugs such as cocaine and heroin or for questions regarding the current use of illegal drugs rather than use in the past year. For these the tendency deliberately to deny use may be greater, which will induce more local dependence among the indicators. Thus, model 2 or 4 may emerge as the best model for describing the error in more sensitive questions on drug use.

In the next two sections, we present the model 1 estimates of the classification error rates for each indicator, first for the total population and then by the age, race and sex. We compared these estimates with estimates derived from models 2–4 and found that they differ somewhat from the estimates which we report, particularly at the group level. However, the major findings that are reported here do not depend on the choice of model.

3.4.3. Total population level estimates of classification error

Table 4 shows the estimated classification error rates (expressed as percentages) for the total population, for all three years including a revised 1994 data set, denoted by 1994′. This data set is identical with the 1994 data set for indicators A and B but differs importantly for indicator C as described below. Standard errors, which are provided in parentheses, assume simple random sampling and do not take into account the unequal probability cluster design of the NHSDA. Consequently, they may be understated.

Several key points can be made from these results.

(a) The false positive rates for all three indicators are very small across all three years except for indicator C in 1994 where it is 4.07%: more than four times that of the other two measurements.

Table 4. Comparison of estimated percentage classification error by indicator†

True classification	*Indicator of use in past year*	*Classification errors (%) for the following data sets:*			
		1994	*1994′*	*1995*	*1996*
Yes ($X = 1$)	Recency ≡ no ($A = 2$)	7.29 (0.75)	6.93 (0.72)	8.96 (0.80)	8.60 (0.79)
	Direct ≡ no ($B = 2$)	1.17 (0.31)	1.18 (0.31)	0.90 (0.28)	1.39 (0.34)
	Composite ≡ no ($C = 2$)	6.60 (0.70)	7.18 (0.72)	5.99 (0.67)	7.59 (0.74)
No ($X = 2$)	Recency ≡ yes ($A = 1$)	0.03 (0.02)	0.03 (0.02)	0.01 (0.01)	0.08 (0.02)
	Direct ≡ yes ($B = 1$)	0.73 (0.07)	0.76 (0.07)	0.78 (0.07)	0.84 (0.07)
	Composite ≡ yes ($C = 1$)	4.07 (0.15)	1.23 (0.09)	1.17 (0.08)	1.36 (0.09)

†Standard errors are given in parentheses.

(b) The false negative error rates vary from 6.93% (1994′) to 8.96% (1995) for the recency indicator A, from 5.99% (1995) to 7.59% (1996) for the composite indicator C and only from 0.90% (1995) to 1.39% (1996) for the frequency indicator B.
(c) Across all four data sets, the same general results hold, i.e. substantial false negative rates for A and C, low false negative rates for B and low false positive rates for A, B and C except as noted in (b).

The large false positive rate for C in 1994 suggests that the high inconsistency rate between C and the other two indicators is due to classification error in C and not classification error in the other two indicators as hypothesized earlier. To investigate further, the questions comprising C for 1994 were compared with those in 1995 and 1996 to identify changes in the questionnaire that might explain this finding. The primary difference between 1994 and 1995 or 1996 is the change in question 7 after 1994 (see Appendix A). For 1994, the instructions for question 7 were quite complicated and the response task (mark a response in column B for each yes in column A) could have been confusing to many respondents. After 1994, this question was replaced by other questions that do not suffer from this problem. Therefore, it is plausible that question 7 is the primary cause of the high false positive rate for C in 1994.

To test this hypothesis, a new indicator was created, denoted by C', by deleting question 7 from indicator C. This new indicator replaced C in the revised data set denoted by 1994′ and the five models described above were refitted to these data. The last six rows of Table 3 show the fit statistics and, again, model 1 was selected by the selection criteria. The error parameter estimates from this model are in Table 4 under the 1994′ column.

Note that the false positive rate for C using the revised 1994 data set dropped to 1.23% from 4.07%. Thus, our hypothesis that item 7 is the cause of the high false positive rate for C in 1994 is confirmed. To verify the LCA result, we checked the consistency of C' with A and B for 1994 and found it to be similar to that for C *versus* A and B in 1995 and 1996. We conclude that question 7 is the cause of the poor agreement of C with A or B in 1994 and, thus, in our subsequent analysis, we shall evaluate the error rate in C' using the 1994′ data set as well as in C. (After reviewing this report, an NHSDA staff member (Rachel Caspar) confirmed that item 7 in 1994 was dropped because many respondents were incorrectly completing the item and, on the basis of interviewer reports, experienced much difficulty in understanding the item.)

In addition to the large false positive rate finding in point (a), the small false negative rate for B noted in (b) was also quite unexpected. Why should the false negative rate for B be so much smaller than for A? We looked for an explanation in the statement of the survey questions for A and B. Indicator A is based on the question 'How long has it been since you last used marijuana or hashish?' whereas indicator B asks 'On how many days in the past 12 months did you use marijuana or hashish?'. For A, respondents who use the drug on only a few days must admit to 'using marijuana' which would classify them in a group ('marijuana users') to which they may think is inappropriate since they used the drug so infrequently. However, for B, respondents can report their frequency of use and, thus, some respondents who deny using marijuana in the past year for the recency question (A) may admit to using the drug on 1 or 2 days on the frequency question (B).

Thus, we hypothesized that respondents who responded falsely to the recency question but answered honestly to the frequency question are the infrequent users. To test this hypothesis, responses for the frequency question were cross-classified by the A-classification. Our hypothesis would be confirmed if a disproportionate number of respondents who were

classified as 'No use in the past 12 months' by indicator *A* and 'Yes, use in the past 12 months' by indicator *B* are light users who responded '1 to 2 days' to the frequency question.

The results of this analysis are reported in Table 5 and are consistent with this theory. Among respondents answering 'No past year use' for *A* and 'Past year use' for *B*, 58.62% (weighted) answered in the 1 to 2 days category for the frequency question. Compare this with only 15.66% in the 1 to 2 days category among people who were consistently classified as users by both indicators.

This analysis demonstrates the utility of LCA for identifying problems with a questionnaire. In addition, the causal analysis investigating findings (a) and (b) above provides evidence of the validity of the latent class estimates.

3.4.4. Estimates of classification error for demographic domains

Next, we examine the model 1 estimates of the classification error probabilities by sex, race or Hispanicity and age domains for each indicator. Estimates by sex are presented in Table 6, race or Hispancity in Table 7 and age in Table 8. The entries in Tables 6–8 are false positive (corresponding to the rows 'true' ≡ 'yes' and 'observed' ≡ 'no') and false negative (corresponding to the rows 'true' ≡ 'no' and 'observed' ≡ 'yes') probability estimates with their simple random sampling standard errors.

Indicator *A* is particularly important to the NHSDA since it is a key question in the newly redesigned instrument that was introduced in 1999. Focusing on the corresponding section of Tables 6–8, we see the following for indicator *A*:

(a) no significant differences in error rates between males and females for all three years;
(b) no significant differences in error rates between the Hispanics, blacks and whites or other demographic groups—however, across years (except for 1994′), the estimates for blacks and Hispanics are slightly larger than those for whites or other;
(c) except for 1995, no significant differences in error rates between the four age groups and no pattern of non-significant differences were observed—in 1995, the false negative error rate for the 35 years and older age group was significantly higher ($p < 0.05$) than for the other age groups.

Table 5. Distribution of reported days of use in the past year by recency of use in the past year: 1994–1996 NHSDAs†

Number of days used marijuana in past year from frequency question	*% reporting no use in past year on recency question*	*% reporting use in past year on recency question*
More than 300	5.84	10.00
201–300	0.96	5.54
101–200	0.93	9.06
51–100	1.45	10.01
25–50	2.96	10.50
12–24	4.76	11.95
6–11	6.06	11.76
3–5	18.41	15.51
1–2	58.62	15.66
Total	100.00	100.00

†Based on 53715 responses.

Table 6. Estimated classification probabilities and standard errors by sex for indicators *A*, *B* and *C*†

Classification		*Group*	*Estimated probabilities for the following indicators and data sets:*											
True	*Observed*‡		*Indicator A*				*Indicator B*				*Indicator C*			
			1994	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*
Yes	No	Males	7.28 (0.94)	6.69 (0.90)	8.35 (1.05)	7.96 (0.95)	1.03 (0.30)	1.01 (0.29)	0.68 (0.23)	1.17 (0.31)	6.04 (0.67)	6.54 (0.74)	4.98 (0.67)	6.39 (0.72)
($X = 1$)	($I = 2$)	Females	7.29 (1.29)	7.37 (1.28)	9.85 (1.39)	9.75 (1.44)	1.42 (0.44)	1.49 (0.46)	1.22 (0.42)	1.78 (0.48)	7.61 (0.90)	8.31 (1.05)	7.47 (1.02)	9.77 (1.14)
No	Yes	Males	0.04 (0.02)	0.04 (0.02)	0.02 (0.01)	0.09 (0.03)	0.87 (0.11)	0.94 (0.11)	1.03 (0.12)	1.04 (0.12)	4.71 (0.24)	1.46 (0.13)	1.50 (0.13)	1.72 (0.14)
($X = 2$)	($I = 1$)	Females	0.04 (0.02)	0.03 (0.02)	0.01 (0.01)	0.07 (0.02)	0.61 (0.08)	0.60 (0.08)	0.56 (0.08)	0.67 (0.19)	3.51 (0.19)	1.04 (0.10)	0.88 (0.09)	1.06 (0.10)

†Standard errors are given in parentheses.
‡*I* denotes either the indicator *A*, *B* or *C*.

Table 7. Estimated classification probabilities and standard errors by race for indicators *A*, *B* and *C*†

Classification		*Group*	*Estimated probabilities for the following indicators and data sets:*											
True	*Observed*‡		*Indicator A*				*Indicator B*				*Indicator C*			
			1994	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*
Yes	No	Hispanic	9.34 (3.02)	8.77 (2.91)	11.08 (3.32)	10.84 (3.16)	1.02 (0.35)	0.99 (0.35)	1.20 (0.48)	1.89 (0.72)	7.44 (1.14)	8.08 (1.67)	6.28 (1.49)	9.65 (1.91)
($X = 1$)	($I = 2$)	Black	7.53 (2.19)	6.40 (2.00)	12.58 (2.83)	10.16 (2.31)	0.61 (0.21)	0.59 (0.19)	0.51 (0.20)	1.06 (0.35)	7.54 (1.07)	7.51 (1.35)	8.73 (1.80)	9.01 (1.54)
		White or other	7.05 (0.84)	6.84 (0.83)	8.23 (0.93)	8.09 (0.89)	1.27 (0.36)	1.28 (0.36)	0.92 (0.32)	1.40 (0.37)	6.37 (0.71)	7.03 (0.77)	5.55 (0.73)	7.14 (0.78)
No	Yes	Hispanic	0.03 (0.02)	0.02 (0.02)	0.01 (0.01)	0.06 (0.03)	0.90 (0.25)	0.94 (0.25)	0.64 (0.21)	0.67 (0.20)	3.75 (0.47)	1.15 (0.23)	1.08 (0.24)	1.09 (0.22)
($X = 2$)	($I = 1$)	Black	0.03 (0.02)	0.04 (0.02)	0.01 (0.01)	0.06 (0.03)	1.40 (0.28)	1.52 (0.29)	1.33 (0.29)	1.18 (0.26)	3.74 (0.42)	1.27 (0.22)	0.77 (0.17)	1.18 (0.21)
		White or other	0.04 (0.02)	0.04 (0.02)	0.01 (0.01)	0.08 (0.02)	0.62 (0.07)	0.63 (0.07)	0.72 (0.08)	0.82 (0.08)	4.15 (0.17)	1.25 (0.09)	1.24 (0.09)	1.42 (0.10)

†Standard errors are given in parentheses.
‡*I* denotes either the indicator *A*, *B* or *C*.

Table 8. Estimated classification probabilities and standard errors by age for indicators *A*, *B* and *C*†

Classification		*Group (years)*	*Estimated probabilities for the following indicators and data sets:*											
True	*Observed*‡		*Indicator A*				*Indicator B*				*Indicator C*			
			1994	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*	*1994*	*1994′*	*1995*	*1996*
Yes	No	12–17	4.91 (1.62)	4.53 (1.52)	7.23 (1.69)	10.57 (2.02)	1.76 (0.62)	1.74 (0.59)	1.52 (0.55)	2.27 (0.66)	6.88 (0.93)	7.70 (1.31)	6.31 (1.10)	10.02 (1.48)
($X = 1$)	($I = 2$)	18–25	8.19 (1.31)	7.64 (1.25)	7.14 (1.25)	8.88 (1.28)	0.71 (0.22)	0.68 (0.21)	0.29 (0.10)	0.74 (0.23)	5.23 (0.67)	5.20 (0.77)	4.11 (0.66)	5.26 (0.74)
		26–35	8.29 (1.62)	8.15 (1.57)	7.88 (1.58)	8.00 (1.59)	0.90 (0.29)	0.93 (0.29)	0.44 (0.17)	1.13 (0.36)	5.37 (0.69)	4.74 (0.72)	4.78 (0.79)	6.62 (0.97)
		> 35	6.51 (1.40)	6.24 (1.36)	13.92 (2.02)	7.34 (1.53)	1.69 (0.52)	1.74 (0.52)	1.71 (0.60)	2.00 (0.61)	9.31 (1.06)	11.61 (1.36)	9.55 (1.28)	10.44 (1.29)
No	Yes	12–17	0.05 (0.03)	0.05 (0.03)	0.02 (0.02)	0.06 (0.02)	0.59 (0.18)	0.64 (0.18)	0.49 (0.16)	0.64 (0.19)	4.71 (0.49)	1.40 (0.24)	1.39 (0.24)	1.30 (0.21)
($X = 2$)	($I = 1$)	18–25	0.03 (0.02)	0.03 (0.02)	0.02 (0.02)	0.07 (0.03)	1.42 (0.28)	1.56 (0.29)	2.41 (0.36)	1.82 (0.32)	5.88 (0.51)	1.98 (0.27)	2.09 (0.29)	2.34 (0.30)
		26–35	0.03 (0.02)	0.03 (0.01)	0.02 (0.01)	0.08 (0.03)	1.08 (0.20)	1.09 (0.20)	1.41 (0.23)	1.17 (0.21)	5.70 (0.43)	2.16 (0.26)	1.72 (0.23)	1.85 (0.23)
		> 35	0.04 (0.02)	0.03 (0.02)	0.01 (0.01)	0.09 (0.03)	0.53 (0.07)	0.54 (0.07)	0.37 (0.06)	0.62 (0.08)	3.18 (0.17)	0.81 (0.08)	0.83 (0.08)	1.08 (0.09)

†Standard errors are given in parentheses.
‡I denotes either the indicator A, B or C.

The next two sections of Tables 6–8 present the corresponding results for indicators *B* and *C*. The notable findings are as follows:

(a) for both indicators, no significant differences in error rates between males and females for all three years;
(b) black respondents have a higher false positive error rate for indicator *B* than do whites or Hispanics;
(c) members of the 35 years and older age group have a higher false negative error rate for indicator *C* than do those in other age groups;
(d) for both indicators *B* and *C*, the false positive error rates were higher for members of the 18–25 and the 26–35 years age groups and the false negative error rates for these two age groups are lower than for the other two age groups.

Since we would not expect a large difference in classification error by sex, the null findings for this variable for all three indicators is plausible. The findings of the higher false negative and lower false negative error rates for the 35 years and older age group, particularly for indicator *C*, fit the pattern hypothesized in the previous discussion of the interpretation of group-by-indicator interactions for model 1. It is plausible that the 35 years and older group could have very different interpretations of some drug use questions than the younger age groups have. For example, a person who smoked marijuana in the past year but did not inhale the smoke may indicate no use of marijuana in the past year. (Recall President Clinton's response to this question several years ago.) A greater tendency to interpret the drug use questions narrowly in this manner for the 35 years and older age group would lead to the differences by age shown in Table 8.

3.4.5. Estimates of prevalence

Finally, we compared three types of prevalence estimates of the use of marijuana in the past year for all three years produced by model 1. Table 9 provides the results for 1996; the results for 1994 and 1995 are similar. The first column of estimates in Table 9 is based on the recency question; the second column of estimates is based on model 1 in Table 3 which essentially adjusts the recency estimates for false negative and false positive classification errors; the third set of estimates corresponds to the official NHSDA estimates based on indicator *T* described in Section 3.3. Table 9 contains overall estimates (first row) as well as estimates by race or ethnicity, age and sex domains.

Because the recency question is biased downwards by a substantial false negative error, we expect the estimates based on *A* to be uniformly lower than the model-based estimates. Further, since the NHSDA estimates are biased upwards because of the false positive error in *A* and *B*, we expect the estimates based on *T* to be uniformly higher than the model estimates. This ordering of the three sets of estimates is apparent in the 1996 results as well as for the other two years not shown in Table 9.

This analysis indicates that the population level estimates from the NHSDA are, on average, between 0.7 and 0.9 percentage points higher than the corresponding model-based estimates. However, for some subgroups, the difference may be as high as 1.5 percentage points. The national level estimates based on the single recency question are between 0.5 and 0.6 percentage points lower than the corresponding model-based estimates. At the subgroup level, the difference may be as high as 1.9 percentage points. Thus, the consequence of assuming no false negative results as for the recency prevalence estimator or no false positive results as for the NHSDA estimator can be substantial for some domains.

Table 9. Comparison of recency, latent class model and NHSDA prevalence rates for the use of marijuana in the past year, for 1996†

Domain	*Prevalence rates (%) for the following indicators:*		
	Recency	*Latent class model*	*NHSDA*
Total	7.16 (0.19)	7.75 (0.20)	8.60 (0.35)
Race or ethnicity			
Hispanic	5.72 (0.55)	6.36 (0.59)	7.04 (0.44)
Black	8.79 (0.62)	9.72 (0.66)	11.09 (0.65)
White or other	7.11 (0.21)	7.65 (0.22)	8.43 (0.45)
Age (years)			
12–17	11.27 (0.72)	12.54 (0.76)	12.99 (0.78)
18–25	20.40 (0.83)	22.34 (0.86)	23.85 (1.08)
26–34	9.44 (0.53)	10.19 (0.55)	11.33 (0.54)
>35	2.94 (0.16)	3.08 (0.17)	3.76 (0.34)
Sex			
Male	9.55 (0.31)	10.29 (0.33)	11.38 (0.58)
Female	4.95 (0.22)	5.41 (0.23)	6.02 (0.31)

†Standard errors are given in parentheses.

One reason why the NHSDA estimator may be preferred over the model-based estimator despite its upward bias is that it may at least partially compensate for false negative errors which are not accounted for by the LCA. For example, respondents who use marijuana yet deny with probability 1 each time that the question is asked in the survey (so-called certain deniers) are not accounted for in the false negative probability estimates. Our limited simulation studies to date indicate that the bias in the false negative error estimates may be substantial when the proportion of users who are certain deniers exceeds 15%; however, more study is needed to understand this effect fully. Still, it is possible that the NHSDA estimator is less biased than the model-based estimator when the consistent denier bias is considered.

It is also possible that the patterns of false positive bias for population subgroups of interest are very different from the patterns of certain denier bias for these subgroups. For example, if we assume that false positive error occurs as a result of the complexity of the questionnaire and its difficulty, then the population subgroups that are most prone to false positive bias are low literacy populations and individuals who are careless in completing the NHSDA answer sheets regardless of age, race, sex, etc. By using the false positive error to compensate for the certain deniers, the tacit assumption is that the same population subgroups that inadvertently answer that they used marijuana when they did not are the same groups that deny their use consistently. Although it may be true that the use of marijuana is higher for some groups who have low literacy, to base a bias adjustment on the assumption that false positive and consistent denial patterns are similar is highly questionable.

4. Discussion

LCA methods have considerable potential for providing more valid estimates of self-reported drug use; however, we see even greater potential for exploiting these methods for

identifying problems in the design of questionnaires and the wording of questions. The problems in the questionnaire identified in this paper would have been much more difficult to discover by using other means of analysis. For example, the inconsistency analysis of Table 2 suggests that, for the 1994 NHSDA, indicator C is considerably more inconsistent with the other two measurements of the use of marijuana in the past year. However, it is not apparent that the problem is due to false positive error in indicator C and/or false negative error in the other two indicators. The LCA estimates of false positive and false negative error for the three indicators quickly and unequivocally identified the problem as false positive error in C.

Similarly, an analysis of the inconsistencies for indicators A and B clearly demonstrates a problem for one of the two indicators, but it is not apparent that the source of the inconsistency was the response of low frequency users to the recency question. LCA quickly led us to suspect that indicator A was the source of the inconsistency. This led to a further investigation using a more traditional analysis which uncovered the source of the problem. These analyses are illustrative of the utility of LCA for indicating problems in the execution of surveys.

Because the three indicators in our analysis are all obtained in one interview, we had expected a strong correlation between the errors in the indicators, yet this was not so. Even this null finding is interesting in that it suggests the absence of social desirability influences on responses to questions on the use of marijuana in the past year in the NHSDA. However, we expect that local dependence models will be quite important for the analysis of embedded indicators of more stigmatized drug use behaviours.

It should be noted that the NHSDA is currently undergoing a redesign and a conversion to computer-assisted self-interviewing (CASI). Questions on drug use that were in use before 1999 have been revised in the CASI version, so the problem identified for the recency question and the potential biases that are associated with the former NHSDA estimation approaches are not directly applicable to the post-1999 NHSDA design. In the CASI implementation, the marijuana sequence uses a 'gate' question that asks respondents whether they have ever used marijuana in their entire lifetimes. Only those who respond positively to this gate question are asked more detailed questions about their use of marijuana and recency. The analytical techniques described in this paper could provide a means for evaluating the error in the new design and determining whether the gains in accuracy expected from CASI administration are realized.

Appendix A: Definition of indicators *A*, *B* and *C* for 1994, 1995 and 1996

A.1. Indicator A for all three years—the recency-of-use question

From NSHDA answer sheet 3:

How long has it been since you last used marijuana or hashish?

$A \equiv$ 'yes' if either 'within the past 30 days' or 'more than 30 days but within past 12 months'; $A \equiv$ 'no' if otherwise.

A.2. Indicator B for all three years—the frequency-of-use question

From NHSDA answer sheet 3:

Now think about the past 12 months from your 12-month reference date through today. On how many days in the past 12 months did you use marijuana or hashish?

$B \equiv$ 'yes' if the response is 1 or more days; $B \equiv$ 'no' otherwise.

A.3. Indicator C for 1994—the composite question

From NHSDA answer sheet 16:

1. As you read the following list of drugs, please mark one box beside *each* type of drug to indicate whether you have used that drug *during the last 12 months*.
2. As you read the following types of drugs, please mark one box beside *each* type of drug to indicate whether you had a period of a month or more during the past 12 months when you *spent a great deal of time getting the drug, using the drug, or getting over its effects*.
3. As you read the following list of drugs, please mark one box beside *each* type of drug to indicate whether you have *used that kind of drug much more often or in larger amounts than you intended during the past 12 months*.
4. As you read the following types of drugs, please mark one box beside *each* type of drug to indicate whether you have *built up a tolerance for the drug so that the same amount of the drug had less effect than before* during the last 12 months.
5. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether you have *often been under the effects or after-effects of that kind of drug in situations where your physical safety was threatened (such as driving a car or motorcycle, using heavy machinery, or swimming)* during the past 12 months.
6. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether your use of that drug has caused you to have *problems with your family or friends, problems at work, school, or with the police, or any emotional or psychological problems* during the past 12 months.
7. As you read the following list of types of drugs, please mark one box beside *each* type of drug. On each line under *Column A*, mark the "YES" box on the left if you *wanted* to cut down or stop using that drug in the past 12 months. Mark the "NO" box on the right if you did *not* want to cut down or stop using that drug or if you did not use that drug during the past 12 months.

For each "Yes" box you mark in *Column A*, please indicate in *Column B* whether you were *able* to cut down on or stop your use of that drug every time you wanted to during the past 12 months. Mark the "YES" box in *Column B* if you were able to cut down on or stop your use of that drug *every time* you wanted to during the past 12 months. Mark the "NO" box if you were unable to cut down or stop your use of that drug when you wanted to during the past 12 months.

$C \equiv$ 'yes' if either questions 1, 2, … or 7 answered 'yes'; $C \equiv$ 'no' otherwise.

A.4. Indicator C for 1995 and 1996—the composite question

From NHSDA answer sheet 14 (1995) or answer sheet 16 (1996) (questions 1–4 are identical with those for indicator *C* for 1994):

5. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether your use of that drug has *often kept you from working, going to school, taking care of children, or engaging in recreational activities* during the past 12 months.
6. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether your use of that drug has caused you to have any *emotional or psychological problems—such as feeling uninterested in things, feeling depressed, feeling suspicious, feeling paranoid, or having strange ideas* during the past 12 months.
7. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether your use of that drug has caused you *any health problems—such as liver disease, stomach disease, pancreatitis, feet tingling, numbness, memory problems, an accidental overdose, a persistent cough, a seizure or fit, hepatitis, or abscesses* during the past 12 months.
8. As you read the following list of types of drugs, please mark one box beside *each* type of drug to indicate whether, during the past 12 months, *you have wanted or tried to stop or cut down on your use of that drug but found that you couldn't*.

$C \equiv$ 'yes' if either questions 1, 2, … or 8 answered 'yes'; $C \equiv$ 'no' otherwise.

Appendix B: Classification error rates for the model {*AX*, *AF*}

In this appendix we show that, for estimating the probability $P(A|XF)$ where X is a dichotomous latent variable, A is a dichotomous indicator of X and F is a grouping variable with f levels, the model $\{AX, AF\}$ implies that $\phi_f\theta_f$ is constant for all f, where ϕ_f is the odds of a false positive error and θ_f is the odds of a false negative error.

Let

$$\lambda_{xf} = \log(\pi_{a=1|xf}/\pi_{a=2|xf}). \tag{7}$$

When $x = 2$, $\lambda_{x=2,f}$ is the log-odds of a false positive error and, when $x = 1$, $-\lambda_{x=1,f}$ is the log-odds of a false negative error. It therefore follows that

$$\begin{aligned}\lambda_{x=2,f} - \lambda_{x=1,f} &= \log(\pi_{a=1|x=2,f}/\pi_{a=2|x=2,f}) + \log(\pi_{a=2|x=1,f}/\pi_{a=1|x=1,f}) \\ &= \log(\phi_f) + \log(\theta_f)\end{aligned} \tag{8}$$

where ϕ_f is the odds of a false positive error and θ_f is the odds of a false negative error. Now, the model $\{AX, AF\}$ for $\pi_{a|xf}$ implies that

$$\lambda_{xf} = \omega + \omega_x^X + \omega_f^F \tag{9}$$

for logistic regression variables ω, ω^X and ω^F. Thus,

$$\begin{aligned}\lambda_{x=2,f} - \lambda_{x=1,f} &= \omega + \omega_2^X + \omega_f^F - (\omega + \omega_1^X + \omega_f^F) \\ &= \omega_2^X - \omega_1^X \\ &= \lambda_0\text{, say,}\end{aligned} \tag{10}$$

where λ_0 is a constant for all f. Therefore, it follows from equation (8) that $\phi_f\theta_f$ is constant for all f.

References

Biemer, P. (1988) Measuring data quality. In *Telephone Survey Methodology* (eds R. Groves *et al.*), pp. 273–282. New York: Wiley.

Biemer, P. and Forsman, G. (1992) On the quality of reinterview data with applications to the current population survey. *J. Am. Statist. Ass.*, **87**, 915–923.

Biemer, P. and Trewin, D. (1997) A review of measurement error effects on the analysis of survey data. In *Survey Measurement and Process Quality* (eds L. Lyberg, P. Biemer, M. Collins, E. De Leeuw, C. Dippo, N. Schwartz and D. Trewin). New York: Wiley.

Biemer, P. and Witt, M. (1996) Repeated measures estimation of measurement bias for self-reported drug use. *J. Off. Statist.*, **12**, 275–300.

Biemer, P., Woltman, H., Raglin, D. and Hill, J. (2001) A latent class analysis of census coverage error. *J. Off. Statist.*, to be published.

Borhnstedt, G. W. (1983) Measurement. In *Handbook of Survey Research* (eds P. H. Rossi, R. A. Wright and A. B. Anderson), pp. 70–122. New York: Academic Press.

Cox, B., Witt, M., Traccarella, M. and Perez-Michael, A. (1992) Inconsistent reporting of drug use in 1988. In *Survey Measurement of Drug Use: Methodological Studies* (eds C. Turner, J. Lessler and J. Gfroerer). Washington DC: US Department of Health and Human Services.

Dayton, C. M. and Macready, G. B. (1980) A scaling model with response errors and intrinsically unscalable respondents. *Psychometrika*, **45**, 343–356.

Fuller, W. A. (1991) Regression estimation in the presence of measurement error. In *Measurement Errors in Surveys* (eds P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz and S. Sudman). New York: Wiley.

Goodman, L. (1973) The analysis of multidimensional contingency tables when some variables are posterior to others: a modified path analysis approach. *Biometrika*, **60**, 179–192.

——— (1974) Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika*, **61**, 215–231.

Haberman, L. (1979) *Analysis of Qualitative Data: New Developments*, vol. 2. New York: Academic Press.

Hagenaars, J. (1988) Latent structure models with direct effects between indicators: local dependence models. *Sociol. Meth. Res.*, **16**, 379–405.

——— (1990) *Categorical Longitudinal Data: Loglinear Panel, Trend, and Cohort Analysis*. Newbury Park: Sage.

——— (1993) *Loglinear Models with Latent Variables*. Newbury Park: Sage.

Hui, S. L. and Walter, S. D. (1980) Estimating the error rates of diagnostic tests. *Biometrics*, **36**, 167–171.

Jay, G., Belli, R. and Lepkowski, J. (1994) Quality of last doctor visit reports: a comparison of medical records and survey data. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 362–367.

Lin, T. H. and Dayton, C. M. (1997) Model selection information criteria for non-nested latent class models. *J. Educ. Behav. Sci.*, **22**, 249–264.

Marquis, K. (1978) Inferring health interview response bias from imperfect record checks. *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass.*, 265–270.

Martin, M. E. and Straf, M. L. (1992) *Principles and Practices for a Federal Statistical Agency*. Washington DC: National Academy Press.

McCutcheon, A. L. (1987) *Latent Class Analysis*. Newbury Park: Sage.

Mieczkowski, T. (1991) The accuracy of self-reported drug use: an evaluation and analysis of new data. In *Drugs, Crime and the Criminal Justice System* (ed. R. Weisheit), pp. 275–302. Cincinnati: Anderson.

Saris, W. and Andrews, F. (1991) Evaluating measurement instruments using a stuctural modeling approach. In *Measurement Errors in Surveys* (eds P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz and S. Sudman). New York: Wiley.

Turner, C., Lessler, J. and Devore, J. (1992) Effects of mode of administration and wording on reporting of drug use. In *Survey Measurement of Drug Use: Methodological Studies* (eds C. Turner, J. Lessler and J. Gfroerer). Washington DC: US Department of Health and Human Services.

Vermunt, J. (1996) *Log-linear Event History Analysis: a General Approach with Missing Data, Latent Variables, and Unobserved Heterogeneity*. Tilburg: Tilburg University Press.

———(1997) *lEM: a General Program for the Analysis of Categorical Data*. Tilburg: Tilburg University Press.