Journal of Official Statistics, Vol. 20, No. 3, 2004, pp. 417–439

The Twelfth Morris Hansen Lecture
Simple Response Variance: Then and Now

Paul P. Biemer[1]

This article begins by reviewing the measurement error model proposed by Hansen, Hurwitz, and Pritzker (1964), with particular emphasis on their concept of simple response variance. More recent developments in the modeling of measurement error are linked to their model and are shown to be extensions and generalizations of their essential concepts. The index of inconsistency, which was first formally described in their paper, is shown to have at least three interpretations, depending upon the statistical framework adopted for describing the gross differences in an interview-reinterview study. Several examples illustrate and compare their classical methods with more modern approaches that employ latent class analysis to estimate the error parameters. It is shown that use of estimates of simple response variance for survey evaluation may obscure important error structures that are more visible using estimates of classification error probabilities to assess data quality.

Key words: Reinterview surveys; latent class analysis; kappa; nonsampling error; reliability.

1. Introduction

One of the most influential papers in the field of survey nonsampling errors is that of Hansen, Hurwitz, and Pritzker (1964). In this landmark paper, Hansen, Hurwitz, and Pritzker (HHP) consider the information on survey data quality that can be obtained from an analysis of "gross differences" – i.e., the differences or discrepancies obtained from repeating a survey or by replicating some questions in a survey. In their paper, the authors develop a model for describing and interpreting the gross differences for measurements of binary data in a survey. They derive an expression for the total mean squared error (MSE) of a sample proportion including components for bias, sampling variance, and response variance.

The authors are thorough and precise in their derivations of the MSE components as they consider both the within and between trial covariances of the response deviations. However, a primary focus of their paper is on a component they refer to as the "simple response variance." HHP define the simple response variance as the "basic trial to trial variability in response, averaged over the elements in the population . . . ". To better understand simple response variance, the authors use the analogy of tossing a coin with probability P_i of a head for the ith population element. If the outcome is "head," the element is classified as a "1" by the survey process; otherwise, the element is classified as a "0." Each survey repetition results in an independent flip of the coin with the same probability P_i of a head. Thus, the probability that the ith element will be classified as 1 is P_i and as 0 is $1-P_i$. Using the Bernoulli distribution, the variance of outcomes for a single element is

[1] RTI International and University of North Carolina at Chapel Hill, Durham, NC 27709, U.S.A. Email: ppb@rti.org

$P_i(1-P_i)$. Then simple response variance is the expected value of this variance over all population units, *viz.*, $\mathrm{E}[P_i(1-P_i)]$.

HHP show that the simple response variance is an important component of the total variance of the sample proportion. To gauge the magnitude of the simple response variance they introduce a measure called the "index of inconsistency," which they define as the ratio of the simple response variance to the total variance. Their estimator of the index of inconsistency is a function of the *gross difference rate*, which is the proportion of the replicate outcomes that do not agree with the original outcomes. Finally, HHP discuss a number of applications at the U.S. Census Bureau where they estimated the index of inconsistency and obtained important information on data quality that led to improved survey processes.

For more than 40 years, the theories and methods first described in HHP's seminal paper have been the basis of numerous reinterview evaluation studies at the U.S. Bureau of the Census and other survey organizations world-wide. Even today, the basic concepts described in their paper continue to influence the analysis and interpretation of the results from reinterview surveys. There have also been a number of important new developments in the analysis of remeasurement data in general. For example, HHP's model for gross differences in binary data has been extended and generalized in several ways to include:

- three or more repeated measurements,
- polychotomous response variables,
- components for classification error probabilities, and
- more general assumptions for the error distributions through the specification of latent class models.

The present article continues in the same spirit as Hansen et al. (HHP) (1964), sharing their conviction that important survey improvements can be achieved through the analysis of gross differences and replicated measurement data. One objective of the present article is to present a modern view of simple response variance and to relate this view to the original ideas developed by HHP. Specifically, we show that the more recently developed approach of latent class analysis for the analysis of repeated measurements can be viewed as an extension of HHP's model for gross differences. The universal appeal of the index of inconsistency as a measure of data quality is demonstrated by presenting three quite distinct yet highly useful interpretations of the measure from the measurement error literature. Some examples that illustrate the use of latent class analysis for evaluating estimating simple response variance as well as other variance and bias components of the total mean squared error are also presented. These examples clearly illustrate the limitations of the index for error evaluation studies and demonstrate that in some cases, classification error probabilities can be estimated that may be more informative for the purpose of identifying the sources of the measurement error.

2. Simple Response Variance

In this section, the total mean squared error of a sample proportion is derived under the HHP (sometimes referred to as the U.S. Census Bureau) model and the classification probability

model. It will be shown that the latter model can be obtained from the former one by simply decomposing the HHP response propensities into components for classification error probabilities – i.e., false positive and false negative probabilities in the dichotomous case. To facilitate the development of the theoretical results, we exploit an analogy between repeated measurements of population elements under identical survey conditions and two-stage cluster sampling. The next section briefly reviews the formulas for the estimator of the population proportion under two-stage cluster sampling and its variance.

2.1. A review of two-stage cluster sampling formulas

The formulas for two-stage cluster sampling with equal size clusters can be found in almost any text on sampling theory (see, for example, Cochran 1977). Simple random sampling at both stages and equal cluster sizes are assumed. We begin by defining the required population quantities and parameters.

Let the population consist of N clusters or primary sampling units (PSUs) each containing M secondary sampling units (SSUs). Let Y_{ij} denote the value of the characteristic of interest for the jth SSU in the ith PSU. To fix the ideas, we assume a dichotomous (i.e., 0 or 1) survey variable; however, later we will generalize these concepts to polychotomous categorical variables. Let P_i denote the proportion of 1's (positives) in the ith PSU, $i = 1, \ldots, N$ and let P denote the population proportion which is to be estimated, i.e.,

$$P = N^{-1}\sum_{i=1}^{N} P_i \tag{1}$$

Finally, we define S_1^2, the between PSU variance, as

$$S_1^2 = \sum_{i=1}^{N} \frac{(P_i - P)^2}{N-1} \tag{2}$$

and S_2^2, the within PSU variance, as

$$S_2^2 = \sum_{i=1}^{N} \frac{P_i Q_i}{N} \tag{3}$$

where $Q_i = 1 - P_i$ (cf. Equation 14 in HHP).

Next, we introduce the sample quantities and statistics needed for estimation. Let n and m denote the number of PSUs and SSUs, respectively, in the sample. Let p_i denote the sample proportion for primary $i, i = 1, \ldots n$ and let p denote the sample proportion given by

$$p = \frac{\sum_{i=1}^{n} p_i}{n} \tag{4}$$

Assuming simple random sampling, p is an unbiased estimator of P with variance

$$\mathrm{Var}(p) = (1-f_1)\frac{S_1^2}{n} + (1-f_2)\frac{S_2^2}{nm} \tag{5}$$

An unbiased estimator of this variance is

$$v(p) = \frac{1-f_1}{n}s_1^2 + \frac{f_1(1-f_2)}{nm}s_2^2 \tag{6}$$

where

$$s_1^2 = \frac{\sum_{i=1}^{n} (p_i - p)^2}{n-1} \tag{7}$$

$$s_2^2 = \frac{m}{n(m-1)} \sum_{i=1}^{n} p_i q_i \tag{8}$$

f_1 is the PSU sampling fraction, n/N, f_2 is the within PSU sampling fraction, m/M, and $q_i = 1 - p_i$.

2.2. The HHP measurement error model

In this section, we consider the estimation of the population proportion P when the survey observations are subject to measurement errors. Recall that the HHP model is analogous to the situation where a coin is associated with each individual in the population. With each survey trial, the coin is flipped to determine whether an individual in the sample is classified as a "1" or "0." In this sense, the ith individual represents a "cluster" of responses with probability, P_i, of a positive response. Thus, HHP's formulas for the mean and variance of an estimator of the true population proportion can be obtained directly by applying the formulas for cluster sampling.

Now suppose a simple random sample of size n is drawn from a population of size N and consider a sequence of hypothetical repeated observations on the ith unit denoted by y_{1i}, $y_{2i}, \ldots, y_{ti}$, and so on. Of course, in a typical survey, only one observation is obtained on each sample unit and these observations will be denoted by y_{1i}, $i = 1, \ldots, n$. In a reinterview study, the individuals in the original sample are reinterviewed, resulting in two observations on each individual denoted by y_{1i} and y_{2i}, $i = 1, \ldots, n$.[2] In general, we denote the tth measurement on the ith individual in the sample as y_{ti} for $t = 1, 2, \ldots, m$ and $i = 1, \ldots, n$.

HHP assume that the repeated measurements are obtained under identical survey conditions and that the values y_{ti} and $y_{t'i}$ for $t \neq t'$ are uncorrelated for each i. Thus, the repeated measurements y_{ti} for $t = 1, 2, \ldots$ are independent and identically distributed random variables (sometimes referred to as *parallel measurements*). Under these so-called *parallel assumptions*, the survey measurement process is analogous to the cluster sampling setup described in the previous section. Individuals in the population represent the clusters and the hypothetical repeated observations on each individual represent the units within a cluster. The response propensity parameter, P_i, is interpreted as the probability that the ith individual in the population will be classified as a positive by the survey process. For a study with m trials, p_i is the proportion of trials resulting in a positive classification for the ith sample individual.

To obtain the variance of the sample proportion, note that the within PSU sampling fraction f_2 in (5) is 0 since we assume that the number of trials available for each individual in the population is essentially infinite. Thus, from (5), we obtain HHP's expression for the

[2] Note we have tacitly assumed here that all original sample units are reinterviewed; i.e., the interview and reinterview sample sizes are both n. This assumption is made to simplify the discussion of the general ideas.

variance given by

$$\text{Var}(p) = (1-f_1)\frac{S_1^2}{n} + \frac{S_2^2}{nm} \tag{9}$$

HHP consider the important special case of a single observation on a sample of size 1, i.e., $n = m = 1$ in (9). In that case, the total variance is $(N-1)N^{-1}S_1^2 + S_2^2$. HHP refer to the first term in this expression as the *sampling variance* (SV) and to the second term (viz., S_2^2) as the *simple response variance* (SRV). The SRV is the trial-to-trial variation in response averaged over all individuals in the population. HHP further show that SV + SRV is simply PQ where $Q = (1-P)$. Further, if f_1 is negligibly small, then the variance simplifies to

$$\text{Var}(p) = \frac{PQ}{n} \tag{10}$$

which is the usual variance of the sample proportion in simple random sampling with replacement.

Note that if there is no measurement error, then P_i is either 1 or 0, S_2^2 is 0 and (9) is reduced to the usual formula for variance of the sample proportion under simple random sampling. In that case, total variance is equal to sampling variance. The SRV is at its maximum value and SV is 0 when P_i is the same for all i. In that case, the classification process is equivalent to flipping the same coin, with P(head) $= P_i = P$, n times to determine the n observations in the sample where a "head" generates a 1 classification and a "tail" a 0 classification and total variance is equal to the simple response variance.

This gives rise to a useful measure of unreliability or inconsistency referred to by HHP as the *index of inconsistency* defined as

$$I = \frac{SRV}{SV + SRV} = \frac{SRV}{PQ} \tag{11}$$

which may be interpreted as the proportion of the total variance that is the response variance for a sample of size 1. It is can also be expressed as the complement of the reliability ratio, R; i.e., $R = 1-I$. Thus, reliability is the proportion of total variance that is "true score" variance (see, for example, Fuller 1991). Note that I and R are both bounded by 0 and 1.

2.3. Classification probability model

Further insights into the structure and effects of classification error can be obtained by considering a classification probability model (Biemer and Stokes 1991). This model assumes that a true value of the characteristic exists for every member of the target population denoted by μ_i, $i = 1, \ldots, N$ which takes the value 1 or 0.[3] As before, y_i is a dichotomous variable denoting the observed value of the characteristic. Assume that the population can be stratified into groups which are homogeneous with regard to P_i; i.e., within a stratum or group denoted by g, individuals that are truly in Class 1 are classified as

[3] Note that the existence of a true value was not required for the HHP model.

1's with probability $1-\theta_g$ where $\theta_g = P(y_i = 0|\mu_i = 1)$. Likewise, individuals that are truly in Class 0 are classified as 1's with probability ϕ_g where $\phi_g = P(y_i = 1|\mu_i = 0)$. Thus, for all i in group g,

$$P_i = \mu_i(1-\theta_g) + (1-\mu_i)\phi_g \tag{12}$$

To simplify the notation, we initially consider the case of a single group, dropping the subscript g, and then generalize the results to multiple groups. In addition, we assume a single survey trial, dropping the index, t.

We will refer to the classification error probabilities ϕ and θ as false positive and false negative probabilities, respectively. The probabilities $1-\phi$ and $1-\theta$ are also referred to in the epidemiological literature as specificity and sensitivity, respectively (see, for example, Rothman and Greenland 1998).

Substituting (12) into S_1^2 and S_2^2 in Equations (2) and (3), we obtain the following expressions for one group (see Biemer and Stokes 1991, for details):

$$\left(\frac{N-1}{N}\right)S_1^2 = \pi(1-\pi)(1-\theta-\phi)^2 \tag{13}$$

and

$$S_2^2 = \pi\theta(1-\theta) + (1-\pi)\phi(1-\phi) \tag{14}$$

Letting π denote the true population proportion, we have the following expression for the bias, $\mathrm{E}(p-\pi)$:

$$Bias(p) = -\theta\pi + \phi(1-\pi) \tag{15}$$

2.4. *Estimation of SRV and the index of inconsistency*

Of particular interest in this article is the estimation of SRV. As noted above, the general estimator for m repeated observations on each unit is given by (8). For the important case where $m = 2$, we form the interview-reinterview table as in Table 1. In this table, p_{11} denotes the proportion of individuals in the sample that are classified as 1 in both the interview and the reinterview, p_{01} is the proportion classified as 0 in the interview and 1 in the reinterview, p_{10} is the proportion classified as 1 in the interview and 0 in the reinterview, p_{00} is the number classified as 0 in both the interview and the reinterview, and $p_{11} + p_{01} + p_{10} + p_{00} = 1$. In addition, we adopt the notation frequently used in the literature on latent class analysis and denote the original observation by A (previously denoted by y_{1i}) and the reinterview classification by B (previously denoted by y_{2i}). Then it can be easily shown that (8) can be rewritten as

$$s_2^2 = \frac{p_{01} + p_{10}}{2} = \frac{g}{2}, \quad \text{say} \tag{16}$$

where g is referred to as the *gross difference rate*. Note that g is the *disagreement rate* or proportion of the sample that is classified inconsistently in the interview and reinterview.

Table 1. An interview-reinterview table

Interview (A)	Reinterview (B)	
	1	0
1	p_{11}	p_{10}
0	p_{01}	p_{00}

The denominator of the index of inconsistency, PQ, can be unbiasedly estimated by the sample quantity pq and thus, a consistent estimator of the index of inconsistency is

$$\hat{I}' = \frac{g}{2pq} \tag{17}$$

An estimator that has somewhat better stability and incorporates information from both the interview and reinterview to estimate PQ is

$$\hat{I} = \frac{g}{p_A q_B + p_B q_A} \tag{18}$$

where p_A and p_B are the interview and reinterview proportions, respectively.

Hess, Singer, and Bushery (1999) show that $\hat{I}$ is identical to $1-\kappa$ where κ is Cohen's kappa measure of reliability (Cohen 1960) and is given by

$$\kappa = \frac{P_0 - P_e}{1 - P_e} \tag{19}$$

where P_0 is the agreement rate between the interview and reinterview classifications (i.e., $1-g$) and P_e is an estimate of the expected agreement by chance alone, i.e., $P_e = p_A p_B + q_A q_B$. Kappa may be interpreted as a "chance corrected" agreement rate since it compares the agreement rate in excess of chance agreement (i.e., $P_0 - P_e$) relative to the maximum value of this quantity ($1-P_e$).

Thus, $\hat{I}$, has two very different interpretations: (1) it is an estimator of I, the ratio of response variance to total variance, and (2) $1-\kappa$ or the chance corrected agreement rate. As we will see subsequently, $\hat{I}$ also has a third interpretation when viewed from a latent class model perspective.

Recall that, in order for s_2^2 to be unbiased for S_2^2, the two replicate measures should be equivalent to a simple random sample of size $m = 2$ from an individual's response distribution, i.e., interview and reinterview measurements must be parallel. As HHP discuss, these assumptions are seldom satisfied in practice. For example, despite attempts to ensure that the general survey conditions are identical for both survey trials, the design of the reinterview survey may be altered or the survey operations associated with the reinterview may be somewhat different from those used in the original survey, violating the equal error distribution assumption. This can happen if the reinterview survey uses an abbreviated questionnaire and more experienced interviewers, or if reinterview respondents tend to be more knowledgeable about the subject matter of the survey as a result of the original interview.

The assumption of conditional (or local) independence of response errors is also unlikely to hold in practice: errors made during the interview are often correlated with those in the reinterview. For example, between trial correlation may be induced if respondents tend to simply recall their interview responses and repeat them rather than

regenerating a response through an independent cognition. However, even when new responses are regenerated, response errors may be correlated if respondents tend to misinterpret the survey questions in the same way on both occasions.

If the error probabilities for the interview and reinterview are different and the errors are correlated, then $\hat{I}$ will be biased. HHP show (cf. Equation 30 in their paper) that in general,

$$\mathrm{E}(g) = \mathrm{SRV_A} + \mathrm{SRV_B} - \rho_{AB}\sqrt{SRV_A SRV_B} + D_{AB}^2 \tag{20}$$

where $\mathrm{SRV_A}$and $\mathrm{SRV_B}$ denote simple response variance for the interview and reinterview, respectively, ρ_{AB} is the between trial error correlation, and D_{AB} denotes the expected difference between the interview and reinterview responses. Suppose ρ_{AB} and D_{AB} are both 0 and let I_{A} and I_{B} denote the indexes of inconsistency for the interview and reinterview, respectively. HHP show that $\hat{I}$ estimates $(I_{\mathrm{A}} + I_{\mathrm{B}})/2$. Thus, if $I_{\mathrm{A}} < I_{\mathrm{B}}$, then $\hat{I}$ will overestimate I and if $I_{\mathrm{A}} > I_{\mathrm{B}}$, then $\hat{I}$ underestimate I (see also U.S. Census Bureau 1985). Now supposing that SRV is the same for both trials, $D_{AB} = 0$ and $\rho_{\mathrm{AB}} \neq 0$, HHP show that $E(g) = 2SRV(1-\rho_{\mathrm{AB}})$. Thus, if response errors are positively correlated (as is their general tendency), $g/2$ will underestimate the simple response variance under these conditions. However, in general, if the parallel assumptions do not hold, then the bias in g can be unpredictable. We will see some examples of that in the illustrations in Section 4.

3. Latent Class Models

An approach which addresses some of the shortcomings of traditional reinterview analysis is latent class analysis (LCA). When only two repeated measures are available, the LCA method must assume local independence to arrive at an identifiable model; however, the assumption of equal error probabilities can be relaxed. In this way, LCA can provide estimates of the misclassification probabilities associated with both the interview and the reinterview. This obviates the need to maintain the same essential survey conditions for the interview and the reinterview. In order to relax the equal error probabilities assumption, LCA with only two measurements requires additional assumptions which may still be problematic in some survey situations. However, for many reinterview surveys they may be more easily satisfied than the traditional parallel assumptions. In the next section, we describe the LCA model when two measurements are available and when both latent and manifest variables are dichotomous. Subsequently, we will address the situation of polychotomous variables and three or more measurements (see also McCutcheon 1987).

3.1. LCA with two measurements

Let X denote the true but unobserved (latent) classification for an individual in the sample, where $X = 1$ if the individual is a true positive and $X = 0$ if a true negative; let A denote the interview response and B the reinterview response, also assumed to be dichotomous variables taking value 1 for an observed positive and value 0 for an observed negative. Let π denote the true population proportion, i.e., $\pi = \mathrm{P}(X = 1)$; let θ_A and θ_B denote the false negative probability for the interview and reinterview, respectively, and ϕ_A and ϕ_B denote the false positive probability for the interview and reinterview, respectively. That is, $\theta_A = \mathrm{P}(A = 0|X = 1)$ and $\phi_A = \mathrm{P}(A = 1|X = 0)$ with analogous definitions for B.

Under the assumption that A and B are conditionally independent given X (sometimes referred to as local independence), we can write the expected cell counts associated with Table 1 in terms of the five parameters π, θ_A, θ_B, ϕ_A and ϕ_B. For models which are identifiable – i.e., a unique maximum value of the likelihood exists – maximum likelihood estimation can be used to estimate the parameters. A necessary condition of identifiability in LCA is that the number of parameters does not exceed the number of degrees of freedom for the model. With five parameters and three degrees of freedom in the AB cross-classification table (Table 1), the LCA model for two measurements is not identifiable. However, we can employ a device suggested by Hui and Walter (1980) to achieve an identifiable model.

Let G denote a grouping variable having two categories. For example, G may denote gender where $G = 1$ for a male and $G = 2$ for a female. Extending the LCA model to the three-way classification table, GAB, there are now 8 cells or 7 degrees of freedom, but 10 parameters to estimate: π_g, θ_{Ag}, θ_{Bg}, ϕ_{Ag} and ϕ_{Bg}, for $g = 1, 2$. Hui and Walter (1980) show that an identifiable model with seven parameters can be obtained by introducing the restrictions:

(a) $\theta_{A1} = \theta_{A2} = \theta_A$, say,
(b) $\theta_{B1} = \theta_{B2} = \theta_B$, say,
(c) $\phi_{A1} = \phi_{A2} = \phi_A$, and
(d) $\phi_{B1} = \phi_{B2} = \phi_B$

That is, the classification error probabilities are the same for the two groups. In addition to these restrictions, a necessary condition for the model to be identified is that $\pi_1 \neq \pi_2$; i.e., the prevalence rates in the two groups are different. The resulting model is fully saturated so there are no residual degrees of freedom for assessing lack of fit.

The Hui-Walter model can be expressed as a hierarchical log-linear model with three terms {GX, AX, BX}(see Hagenaars 1993). Under this model, the true prevalence in the population varies by group (hence, the GX term) while the error probabilities (represented by AX and BX) do not. Any software that can fit log-linear models with latent variables can be used to obtain the MLEs of the parameters for this simple model. The software used in the illustrations to follow is ℓEM software developed by Dr. Jeroen Vermunt (Vermunt 1997).

3.2. Models for three measurements

Extension to three locally independent measurements is straightforward. Let A, B, and C denote the three measurements of the latent variable, X. As for the Hui-Walter model, the three measurements of X need not be obtained under identical survey conditions, i.e., the measurements can have distinct error distributions. However, unlike the case of two measurements, the likelihood associated with the three-measurement model is identifiable without any grouping variable restrictions. Estimation proceeds as before using maximum likelihood estimation.

In practice, three measurements of the same survey characteristic are difficult to obtain, particularly by using a reinterview approach. Reinterview methods risk problems such as respondent burden and resistance, the response conditioning effects of prior contacts, and

high costs associated with repeat contacts with the respondent. Another method for obtaining three measurements of the same variable is that of *embedded replication*, in which replicate measurements are embedded in a single survey instrument and collected at the same interview. Embedded replication has an even greater risk that respondents will impose a false consistency on their responses by repeating the same response each time the question or a similar question is asked. In this situation, the possibility that the errors are locally dependent must be explicitly considered in the analysis. Altering the wording of the replicated items may help to conceal item redundancy from the respondent. This can help avoid respondent resistance to the burden of answering the same questions repeatedly while reducing the risk of correlated errors due to memory effects. However, it can also introduce additional complexity in the modeling process as a result of nonparallel measurements.

Latent class models for three measurements that assume three locally independent measurements having unequal error probabilities are saturated, leaving no degrees of freedom for modeling correlated errors. Thus models which introduce additional terms for local dependence are not identifiable unless further restrictions are placed upon the model (Hagenaars 1988). For example, by imposing the restriction that the classification error probabilities for A, B, and C are identical, two degrees of freedom are saved (in the case of dichotomous measurements) which can be used to estimate the two additional parameters introduced by relaxing the independence assumption for two of the three indicators, – for e.g., $\pi_{b|ax}$. However, this equal error probability restriction is not plausible and is likely to be violated if the methods (i.e., question wordings) for obtaining A, B, and C vary within the questionnaire.

As we described for the Hui-Walter methods, 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 models across the L groups to free-up enough degrees of freedom for estimating the correlated error parameters often results in more plausible restrictions on the model parameters than are possible without the grouping variable. This will be discussed in more detail in Section 4.

3.3. Estimation of simple response variance

Latent class analysis can be employed to obtain estimates of *SRV* and *I* under more general assumptions than are made in traditional analysis. In this section, two types of generalizations will be described. Each method is introduced assuming that only two measurements of the same dichotomous survey characteristic are available. Extensions of the methods to three or more measurements and polychotomous characteristics are also briefly outlined.

One alternative estimator of I can be derived directly from the expression of *SRV* (i.e., S_2^2) in (14). For interview-reinterview data, the Hui-Walter method can be applied to estimate the parameters π, θ_A, θ_B, ϕ_A, and ϕ_B. Denote the MLEs of these parameters by the parameter's symbol with a "hat." Using the MLEs from the LCA, we can estimate I for the original interview by replacing the parameter in (13) by its MLE and dividing by the survey estimate of the total variance as follows:

$$\hat{I}_A = \frac{\hat{\pi}\hat{\theta}_A(1-\hat{\theta}_A) + (1-\hat{\pi})\hat{\phi}_A(1-\hat{\phi}_A)}{p_A(1-p_A)} \tag{21}$$

The index of inconsistency for the reinteriew can be estimated by the same formula replacing the estimates $\hat{\theta}_A$, $\hat{\phi}_A$, and p_A by $\hat{\theta}_B$, $\hat{\phi}_B$, and p_B, respectively.

In Section 2.4 we noted that I can have two interpretations: (a) the proportion of total variance that is due to simple response variance and (b) $1-\kappa$, where κ is the expected agreement rate beyond chance agreement. A third interpretation of I, given by Guggenmoos-Holzmann (1996) and Guggenmoos-Holzmann and Vonk (1998), will now be described.

Assume that the population consists of two types of individuals: those that are easily and unequivocally classified as either 1's or 0's by the survey process and those that are difficult to classify and may be classified randomly as 1's or 0's from trial to trial. Guggenmoos-Holzmann refers the former group as "conclusive" and the later group as "inconclusive." She assumes that conclusive elements are classified consistently across repetitions of a survey process; i.e., $\mathrm{P}(B = 1|A = 1) = \mathrm{P}(B = 0|A = 0) = 1$ for conclusive elements. For inclusive elements, this constraint is removed.

Note that it is not possible to accurately identify to which group an individual belongs since inconclusive persons can sometimes be classified consistently across trials purely by chance and are, therefore, indistinguishable from conclusive persons. Thus, group membership is latent. Further, in terms of the HHP, P_i is either 1 or 0 for persons in the conclusive group while for persons in the inconclusive group, $0 < P_i < 1$.

Let H denote an latent indicator variable for conclusive and inconclusive individuals in the population; i.e., $H = 1$ for individuals in the conclusive group and $H = 2$ for individuals in the inconclusive group. Let $\pi_{\mathrm{H}=1}$ denote the proportion in the conclusive group and let $\pi_{\mathrm{H}=2} = 1-\pi_{\mathrm{H}=1}$ denote the proportion in the inconclusive group. Further, let $\pi_{\mathrm{A}=1|\mathrm{H}=1}$ denote the proportion of the conclusive group that is classified as a 1 by the survey process, i.e., $\mathrm{P}(A = 1|H = 1)$, and let $\pi_{\mathrm{A}=1|\mathrm{H}=2}$ denote the probability that an individual in the inconclusive group is classified as a 1. Finally, we assume these probabilities hold for each replication of the survey process which is analogous to the parallel assumption of the HHP model.

Denote by π_{11}, π_{01}, π_{10}, and π_{00} the expected cell proportions in Table 1; i.e., $\mathrm{E}(p_{ij}) = \pi_{ij}$ for $i, j = 0, 1$. Under these assumptions, the expected cell proportions under this model are:

$$\begin{aligned}
\pi_{11} &= \pi_{H=1}(\pi_{A=1|H=1}) + \pi_{H=2}(\pi_{A=1|H=2})^2 \\
\pi_{01} &= \pi_{10} = \pi_{H=2}(\pi_{A=1|H=2})(\pi_{A=0|H=2}) \\
\pi_{00} &= \pi_{H=1}(\pi_{A=0|H=1}) + \pi_{H=2}(\pi_{A=0|H=2})^2
\end{aligned} \tag{22}$$

Gugenmoos-Holzmann refers to this model as the "agreement" model and the latent class model described in Section 3.1 as the "error" model. Note that, as in the HHP model, the concept of a true value is not required under the agreement model since the latent variable, X, representing the true status of an individual does not appear in the expressions for the cell probabilities. With three parameters and only two degrees of freedom

(since $\pi_{11} + \pi_{01} + \pi_{10} + \pi_{00} = 1$), the model is not identifiable in general; however, the model is identifiable when three measurements on each individual are available or when two measurements are available and the manifest variables are at least trichotomous. An important restriction that achieves identifiability for the two degrees of freedom model is to set

$$\pi_{A=1|H=1} = \pi_{A=1|H=2} \tag{23}$$

i.e., we assume that the probability of a positive classification is the same for both conclusive and inconclusive groups. This restriction will be referred to as the kappa-constraints since Guggenmoos-Holzmann shows under (23), $\kappa = \pi_{H=1}$; i.e., the proportion of the population belonging to the conclusive group is equal to Cohen's κ. Equivalently, $I = \pi_{H=2}$; i.e., the size of the inconclusive group in the population is identical to the index of inconsistency.

Thus, we see that a third interpretation of HHP's index of inconsistency is the proportion of the population that would be classified at random by the survey process. For the groups of inconclusive individuals, positive classifications are made at random with probability $\pi_{A=1|H=2}$. For the conclusive groups, positive classifications are made with certainty and consistently across all repetitions of the survey process with prevalence $\pi_{A=1|H=1}$.

The agreement model yields a more general definition of reliability than either the HHP model or Cohen's κ since under the agreement model κ is equal to $\pi_{H=1}$ only when the kappa constraints (23) are imposed. When the constraints are relaxed, $\pi_{H=1}$ may be interpreted more generally and is referred to as a κ-like measure or a generalized κ.

Gugenmoos-Holzmann and Vonk argue that, in general, (23) will not be satisfied by most survey processes; moreover, there is no compelling reason why it should be. For example, (23) will hold if A and H are independent random variables. However, that assumption seems untenable since individuals having ambiguous classifications may have quite different characteristics than those whose classifications are more easily determined. As an example, if the characteristic A is labor force status (in particular, employed or not employed), persons in the conclusive group may be predominantly employed whereas persons whose status is inconclusive may be not employed since determining that status is usually more difficult.

Another way in which (23) is satisfied is if the classification process for the survey has a type of learning mechanism that classifies inconclusive individuals at the same rate as previously encountered conclusive individuals were classified. This situation might be plausible if a single interviewer conducted all the interviews. As the interviewer encounters conclusive elements, he or she learns that roughly $100\pi_{A=1|H=1}$ percent of the population is positive. Thus, when he or she encounters an inconclusive element, the interviewer "guesses" a category to assign. Since prior experience suggests that roughly $100\pi_{A=1|H=1}$ percent are positive, the interviewer assigns roughly the same proportion of inconclusives to the positive category.

Guggenmoos-Holzmann and Vonk (1998) conclude that for reporting the reliability of survey measurements, the assumption $\pi_{A=1|H=1} = \pi_{A=1|H=2}$ is not plausible and thus, κ (or $1-\hat{I}$) should not be used as a measure of reliability whenever the more general measure, $\pi_{H=1}$, can be computed. Another advantage of the agreement model formulation

for I is that, like $\hat{I}_A$ in (21), the agreement model estimate of I is always between 0 and 1, the parameter space of reliability and inconsistency measures, since it is obtained using maximum likelihood estimation. Both κ and $\hat{I}$ may sometimes lie outside this range, which creates additional problems in the interpretation of the estimates. Some examples illustrating the use of these indicators of classification error are provided in the next section.

4. Illustrations

4.1. Illustration 1: Reliability of a question on race

In 1997, the U.S. Office of Management and Budget released new standards for asking about race to better reflect the increasing racial and ethnic diversity of the population of the United States. Under the new standards, Federal agencies are required to offer individuals the opportunity to select one or more of the following five race categories: (1) American Indian/Alaska Native, (2) Asian, (3) Black/African American, (4) Native Hawaiian/Other Pacific Islander, and (5) White. The first nationwide implementation of these standards was in the 2000 decennial Census.

To test this and other census questions, in 1998 a dress rehearsal pretest of the census operations was conducted in three sites: Columbia, SC, rural SC, and Sacramento. Immediately following the dress rehearsal census in these areas, a reinterview (or post enumeration) (PES) survey was conducted to evaluate the quality of the census data. Reinterviews were conducted for $n = 40,519$ census dress rehearsal respondents. These data will be used in the following to illustrate the methods for estimating I described above. All estimates are based upon unweighted data and are for illustrative purposes only.

To illustrate the simple case of a dichotomous measure, consider the reliability of the census variable A where $A = 1$ if an individual is classified as belonging to two or more races and $A = 2$ if only one race. Define the corresponding PES variable, B, analogously. We wish to compare the three estimates of I for these data: the traditional estimate, $\hat{I}$, the latent class error model estimate, $\hat{I}_A$, and the agreement model estimate, $\hat{\pi}_{H=2}$. The interview-reinterview cross-classification, which provides the sufficient statistics for the preliminary analysis is given in Table 2.

Table 2. Census by PES multiple race response

Census classification (A)	PES Classification (B)		
	Multiple	Single	Totals
Multiple	429	2,328	2,757
Single	956	36,806	37,762
Totals	1,385	39,134	40,519

The inconsistency between the Census and PES classifications is apparent from the data in Table 2. Among the 2,757 persons who chose multiple races in the Census, 2,328 (or 84 percent) changed to a single race in the PES! One possible explanation for this is that different modes of data collection were used for the two surveys. The Census responses were obtained through self-administered, paper questionnaire while the PES was

conducted by face to face and telephone interviewing. Since the race questions in both questionnaires were essentially the same, it is possible that the inconsistency is the result of interviewer effects.

Note that 2,328 individuals classified as multi-racial in the Census, or 84 percent of the Census multi-racial group, were reclassified into a single race category in the PES. Clearly, the parallel repeated measures assumption does not hold for these data. The assumption of equal error distributions can be formally tested by the test H_0: $P_A = P_B$. If the test is rejected, the equal error probabilities assumption must also be rejected. For this table, $p_A = 6.8$ percent and $p_B = 3.4$ percent, which are highly significantly different and the test is rejected. Thus, $\hat{I}$ (or κ) as well as $\hat{\pi}_{H=2}$ will be biased for I and considerably so, judging from the magnitude of the difference. In this situation, the estimator $\hat{I}_A$ in (20) is likely to be a better estimator of I under these conditions.

The "true race" of an individual, which is needed for the latent error model, is conceptually difficult since in many cases race can be subjective and based upon personal preference or racial identity. To use the error model, one must conceptualize a preferred method of obtaining race data – one that is devoid of influences that would cause instability in responses to the race question. This preferred response is regarded as an individual's true race and deviations from this response are interpreted as error. One advantage of such a concept is that it allows an examination of the systematic errors in the determination of race that may be related to the mode of interview. In this illustration, the focus will be on simple response variance; however, Biemer and Woltman (2001) discuss the bias in the race classifications using this error concept.

The alternative estimators of I that can be computed from the data in Table 2 are shown in Table 3. To obtain the Hui-Walter model estimates, $\hat{I}_A$ and $\hat{I}_B$, we used a dichotomous grouping variable denoted by O where $O = 1$ if the individual is of Hispanic origin and $O = 2$ if not. Since the prevalence of multiple race responses differs markedly between Hispanics and non-Hispanics, this choice of grouping variable satisfies one of the Hui-Walter assumptions. However, the assumption that error rates are equal across groups is still questionable. An alternative grouping variable that may better satisfy this assumption is the site variable, S, with three levels: Columbia, SC, rural SC, and Sacramento. But this variable is not ideal either since the proportion of respondents classified in multiple race categories does not differ appreciably across the sites. Perhaps the best choice is a combination of O and S. This model will be discussed subsequently.

Table 3. Estimates of the index of inconsistency by various methods

	$\hat{I}$	$\hat{\pi}_{H=2}$	$\hat{I}_{A=B}$	$\hat{I}_A$	$\hat{I}_B$
Estimate	83.1	83.6	78.5	44.6	94.1

The estimator $\hat{I}_A$ may also be computed using the parameter estimates from a latent class error model that assumes parallel measures. This is easily accomplished by the latent class model {AX BX} with constraints AX = BX or equivalently $\theta_A = \theta_B$ and $\phi_A = \phi_B$. We refer to this estimator as $\hat{I}_{A=B}$. All four estimators of I are shown in Table 3 as well as $\hat{I}_B$, the error model estimate of the index of consistency for the PES (I_B). Standard errors of the estimates are not shown; however, they are quite small – less than 0.2 percentage point.

The estimates of index of inconsistency for the dress rehearsal census are extremely high, particularly for the three methods that assume parallel measures. The estimate of $\hat{I}_A$, which specifies separate error terms for census and PES, is the smallest. Since $\hat{I}_B$ is more than twice that of $\hat{I}_A$, it appears that the source of the large estimates of I for the first three estimates in the table is the PES with an index of 94.1. Note that $(\hat{I}_A + \hat{I}_B)/2$ is 69.3 which approaches the magnitude of estimates of I based upon the parallel measures assumption.

Since the models underlying the estimates $\hat{\pi}_{H=2}$ and $\hat{I}_A$ (or $\hat{I}_B$) are saturated, the usual chi-square goodness of fit test cannot be used to assess fit. The latent class error model with the equal error distribution constraint (i.e., $AX = BX$) permits a test of fit with two degrees of freedom; however, the test is not particularly useful owing to the large sample size. When the sample size is as large as it is in this example, the chi-square goodness of fit criteria will reject models that fit the data well by most other fit criteria since the power of the test is near 1. An alternative measure of model adequacy that is often useful in such cases is the similarity index, d, defined as

$$d = \frac{\sum_k |n_k - \hat{n}_k|}{2n} \tag{24}$$

where n_k is the count in cell k of the cross-classification table and $\hat{n}_k$ is the expected cell count under the model. The index d may be interpreted as the proportion of observations misclassified by the model. A d of 0.01 or lower is usually considered an indication of a well-fitting model since it indicates that less than 1 percent of the data is inconsistent with the model. The highest value of d for the estimates in Table 2 is $d = 0.029$ for $\hat{I}_{A=B}$ while the lowest value, $d = 0.0015$, was associated with the estimates $\hat{I}_A$ and $\hat{I}_B$.. For the agreement model estimate, $d = 0.017$. Thus, the error model with separate error terms provides the best fit under this criterion.

As mentioned previously, the Hui-Walter model can easily be extended to two or more grouping variables; for example, to incorporate both the site variable (S) and Hispanic origin (O). In addition, rather than restricting the analysis to dichotomous race variables we will next consider the reliability of a race classification variable having five categories.

Let A denote the census response with 1 = White, 2 = Black, 3 = API, 4 = Some Other Race, and 5 = More than One (or Multiple) Race, respectively, and let B correspond to the PES response defined analogously. The "Some Other Race" category contains all persons who marked any single race category other than White, Black, and API or wrote-in a single other race. The More than One or Multiple Race category contains all persons who marked two or more race categories or one category and wrote-in one or more other categories. API is formed by collapsing Asian, Native Hawaiian, and other Pacific Islander categories.

The polychotomous form of $\hat{I}$ is the so-called aggregate index of inconsistency defined in U.S. Census Bureau (1985) as

$$\hat{I}_{AG} = \frac{1 - \sum_k p_{kk}}{1 - \sum_k p_{k\cdot} p_{\cdot k}} \tag{25}$$

where p_{kk} is the observed agreement for category k and $p_{k.}\,p_{.k}$ is the product of the marginal probabilities of assigning category k. For the agreement model, we again assume the population consists of two types of individuals: conclusive and inconclusive. Let $\pi_{H=2}$ denote the proportion of inconclusive individuals in the population, let $\pi_{A=a|H=1}$ denote the proportion of individuals in category a in the conclusive group, and let $\pi_{A=a|H=2}$ denote the corresponding probability for the inconclusive group. Under these assumptions, the agreement model is identifiable and has nine parameters since $\sum_{a=1}^{5} \pi_{A=a|H=h} = 1$ for $h = 1, 2$.

The estimate of $\pi_{H=2}$ under this model may be interpreted as a generalized, aggregate index of inconsistency. Guggenmoos-Holzmann (1996) shows that with the constraints $\pi_{A=k|H=1} = \pi_{A=k|H=2}$, for $k = 1, \ldots, 5$, $\hat{\pi}_{H=2}$ is equivalent to $\hat{I}_{AG}$ in (25). However, as noted in the discussion for the dichotomous case, the constraints are rather implausible and not likely to hold in most survey situations. For the general model, the cell probabilities for the interview-reinterview table are

$$p_{kk'} = \pi_{H=1}\pi_{A=k|H=1} + \pi_{H=2}\pi_{A=k|H=2}\,\pi_{A=k'|H=2,\,k,\,k'=1,\ldots,5} \tag{26}$$

With the constraints $\pi_{A=a|H=2} = \pi_{A=a|H=1}$ for $a = 1, \ldots, 5$, we have $\hat{\pi}_{H=2} = \hat{I}_{AG} = 0.776$ with dissimilarity index $d = 0.153$. Removing these constraints, $\pi_{H=2} = 0.826$ with $d = 0.0426$. Note, however, that neither model provides an adequate fit of the data.

For the latent class error model, several alternative models are feasible and can be explored in the model selection process. One identifiable model that also appears to fit the data well is, in hierarchical model notation, $\{SOX, AOX, BOX, AS, BS\}$. The SOX term in the model specifies race prevalence rates across all six groups formed by crossing S and O variables. Note that the absence of the AB-interaction implies that independent classification error (local independence) is assumed.

Since there are 150 cells in the $SOAB$ table and 126 parameters in the model, 24 degrees of freedom are available to test the fit of the model. Again, the model was rejected using the standard chi-square test criterion. However, d for the model was 0.0009, indicating a well-fitting model.

This latent class model provides estimates of π_g, θ_{Ag}, θ_{Bg}, ϕ_{Ag} and ϕ_{Bg}, where $g = 1, \ldots, 6$ denotes the site by Hispanicity groups. These estimates were then used to estimate $I_A$and I_B, using (21). The model fit the data quite well with $d = 0.0017$. The results are given in Table 4.

Table 4. Reliability for the Census and PES by race and hispanicity

Race	Overall		Non-Hispanics		Hispanics	
	$\hat{I}_A$	$\hat{I}_B$	$\hat{I}_A$	$\hat{I}_B$	$\hat{I}_A$	$\hat{I}_B$
White	15.0	20.1	6.9	8.0	84.7	95.8
Black	14.4	9.4	3.5	6.6	94.2	33.9
API	20.0	20.1	18.2	16.4	68.1	85.3
Some Other Race	58.6	39.3	86.4	55.2	60.5	49.1
Multiple	67.1	91.7	63.0	72.8	77.1	99.1

The equality of the terms *AOX* and *BOX* (i.e., equal error rates for the Census and the PES) can be tested using a likelihood ratio test in which the restricted model sets the conditional probabilities P($A|SOX$) equal to P($B|SOX$) and the unrestricted model removes this restriction. The restricted model was rejected with $p < 0.001$ hence, the hypothesis $AOX = BOX$ must also be rejected. This implies that $I_A$and I_B are not equal.

The largest differences between I_A and I_B in these tables occur for the Multiple Races and Some Other Race categories. Note that in some cases, census inconsistency is considerably less than PES inconsistency.

An aggregate index can be computed from the latent error model estimates in Table 4 by noting that the aggregate index in (25) is equivalent to

$$\hat{I}_{AG} = \sum_k W_k \hat{I}_k \tag{27}$$

where $\hat{I}_k$ is the dichotomous index of inconsistency computed from the 2×2 table with categories k and $k' \neq k$, $W_k = \text{denom}(\hat{I}_k)/\sum_j \text{denom}(\hat{I}_j)$, and $\text{denom}(\hat{I}_k)$ is the denominator of the index. For the estimates of I in Table 4, the denominator of $\hat{I}_k$ is $p_k(1-p_k)$ where p_k is the proportion in category k (from (21)). Thus, the corresponding latent class aggregate index computed from (27) is 35.3 percent for the Census and 22.3 percent for the PES. These values are considerably lower than the estimates previously obtained under the equal error distribution assumption, illustrating the heavy reliance of the estimates on the assumed model.

4.2. Illustration 2: Reliability of self-reported marijuana use

Biemer and Wiesen (2002) consider the case of three measurements obtained in a single interview in an application to the National Household Survey on Drug Abuse (NHSDA). The NHSDA is a multistage household survey designed to measure the U.S. population's current and previous drug use activities. Before 1999, the NHSDA was primarily a self-administered interview using a paper and pencil questionnaire. A number of drug use questions are repeated in the questionnaire since research has shown that some respondents who indicate that they never used the drug when asked directly, will later answer an indirect question about the drug in a way that implies use of it. The multiple measurements of drug use can therefore be used to improve the accuracy of drug use prevalence estimates. This redundancy in the questionnaire provides the basis for constructing three remeasurements of past year marijuana use which will be used in an LCA evaluation of these questions.

Biemer and Wiesen define three indicators of past year marijuana use (referred to as *A*, *B*, and *C*) in terms of the questions asked at various points during the interview. Indicator *A* is the response to the so-called recency of use (or recency) question which asks about the length of time since marijuana or hashish was last used. Past 12-month use was coded as "yes" for responses of "within past 30 days" or "more than 30 days but within past 12 months" and was coded "no" otherwise. The 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. This indicator was coded "yes" for any response of one or more days and was coded "no" otherwise. Indicator *C* is a

composite of a number of questions on the so-called drug answer sheet that involved the use of marijuana in the past 12 months. An affirmative response to any one of these is coded as "yes" for C, otherwise the code is "no."

Biemer and Wiesen analyzed three years of NHSDA data – 1994, 1995, and 1996. Their research was primarily focused on estimating the false positive and false negative probabilities separately for A, B, and C in order to determine the accuracy of each method. In this section, we use these data to illustrate the three methods of estimating I described in Section 3.3 for the case of three dichotomous measures.

It is unlikely that the assumptions of traditional analysis hold for the three measures just described. The assumption of equal error distributions also does not seem plausible since the three measures are based upon very different questions. The local independence assumption is also likely to be violated since all three measurements are obtained in a single interview and respondents may remember their earlier responses and try to respond consistently to similar questions. These dependencies can be modeled to some extent using latent class analysis. However, the introduction of additional terms into the model to reflect correlated errors will result in an unidentifiable model unless some restrictions on the model parameters are made (Hagenaars 1988).

Therefore, to achieve an identifiable locally dependent model, Biemer and Wiesen used the Hui-Walter grouping variable technique to increase the model degrees of freedom and then equated some parameters of models across groups to free-up enough degrees of freedom for estimating the correlated error parameters. They represented the between trial correlations in their models by AB, BC, and AC interaction terms in accordance with the ideas of Hagenaars (1988). Since sufficient degrees of freedom are a necessary but not sufficient condition for model identifiability, Biemer and Wiesen verified the identifiability of the models using the method of Goodman (1974).

As an example, for the case of two groups, say $G = 1$ for younger adults and $G = 2$ for older adults, the GAB table has a total of $2(2^3)$ or 16 cells. Denoting the conditional classification probabilities for indicator A in group g by $\pi_{a|gx}$, we assume that

$$\pi_{a|G=1,X=x} = \pi_{a|G=2,X=x} = \pi_{a|X=x} \tag{28}$$

i.e., the classification error probabilities for younger and older adults are equal. For a hierarchical linear model, this is represented by setting the interaction terms GAX and GA to 0. The analogous assumptions are made for indicators B and C as well.

To account for potential local dependence, a causal ordering of errors in the indicators can be assumed that reflects the temporal ordering of the indicators in the interview. That is, we assume local dependence between chronologically adjacent indicators in the NHSDA questionnaire so that the error in B depends upon A and the error in C depends upon B, and the error in C conditional on B is independent of A. Thus, the interaction terms AB and BC are introduced to model the correlation and the interactions AC and ABC are assumed to be 0.

These ideas can be extended in a number of ways. Additional grouping variables can be added to the model which may be desirable, not only to provide additional degrees of freedom for parameter estimation, 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

inter-relationships between the latent variable, the indicators, and the subgroups. However, the quantity of grouping variables was restricted to only a few to avoid problems with model instability and over-fitting.

Biemer and Wiesen explored a number of latent class models with three grouping variables: age (G) with two categories, race (R) with four categories, and sex (S). Interestingly, the best-fitting model in their analysis did not include the local dependence interaction term since it was only marginally significant. The best model was a path model with structural term $XGRS$ and three measurement terms corresponding to the three indicators: $A|XGRS = \{AX\ AG\ AR\ AS\}$, $B|XGRSA = \{BX\ BG\ BR\ BS\}$ and $C|XGRSAB = \{CX\ CG\ CR\ CS\}$. The reader is referred to their paper for an interpretation of the model terms.

Biemer and Wiesen did not consider the simple response variance associated with the three measures. However, error probability estimates from their model can be used to estimate the index of inconsistency for each indicator using the estimator in (21). These estimates appear in Table 5. The estimates in the columns labeled $\hat{I}_{\text{gen}}$ and $\hat{I}_{\text{trad}}$ are computed from the agreement model. The estimate $\hat{I}_{\text{gen}}$ allows the proportion of positives for the conclusive and inconclusive domains to differ while $\hat{I}_{\text{trad}}$ constrains these to be equal using kappa-like constraints. The latter estimate may be viewed as an extension of the traditional index of inconsistency (or Cohen's kappa) to three measures. Gugenmoos-Holzmann and Vonk (1998) consider a further generalization of $\hat{I}_{\text{gen}}$ to allow the conclusive populations to differ for each indicator; however, those estimators are not considered here. The last column is $1-\kappa$, where κ is the chance corrected agreement rate for three indicators.

Table 5. Index of inconsistency for three measures of marijuana use (Entries are percentages)

Year	$\hat{I}_A$	$\hat{I}_B$	$\hat{I}_C$	$\hat{I}_{\text{gen}}$	$\hat{I}_{\text{trad}}$	$1-\kappa$
1994	8.30	9.87	41.53	11.6	22.6	26.8
1995	10.27	11.38	23.14	4.2	13.3	18.2
1996	8.30	10.30	19.74	4.9	15.0	20.6
Average	8.96	10.52	28.14	6.9	17.0	21.9

Several observations can be made from Table 5. Focusing first on the columns labeled $\hat{I}_A$, $\hat{I}_B$, and $\hat{I}_C$, note that the inconsistency rates for C are considerably larger than those for A and B, particularly in 1994. Also, the estimates of I from using the agreement model, $\hat{I}_{\text{gen}}$ and $\hat{I}_{\text{trad}}$ (or equivalently, $1-\kappa$), are quite different from those using the latent class error model. Interestingly the estimates using the more constrained estimate, $\hat{I}_{\text{trad}}$ and $1-\kappa$, tend to agree more closely with the estimates assuming different error distributions for the three measures.

Table 6 provides the false negative and false positive probability estimates from the latent class model that were used to generate the estimates $\hat{I}_A$, $\hat{I}_B$, and $\hat{I}_C$. These estimates provide additional information about measurement errors that is not apparent in Table 5. For example, note that the false negative rates for measures A and C are quite large compared to those for B. However, in Table 5, $\hat{I}_{A,}$ and $\hat{I}_B$ were quite similar across years,

Table 6. Misclassification probabilities for three measures of marijuana use (Entries are percentages)

Year	$\hat{\phi}_A$	$\hat{\theta}_A$	$\hat{\phi}_B$	$\hat{\theta}_B$	$\hat{\phi}_C$	$\hat{\theta}_C$
1994	0.03	7.29	0.73	1.17	4.07	6.60
1995	0.08	8.61	0.84	1.38	1.36	7.59
1996	0.03	7.29	0.78	0.90	1.17	5.99
Average	0.05	7.73	0.79	1.15	2.20	6.73

masking the relatively high false negative probabilities for A. This occurs because, as shown in (21), $\hat{\theta}_A$ is multiplied by $\hat{\pi}$, which is approximately 0.08 for past year marijuana use; thus, the effect of high false negative error on response inconsistency is small.

Note also from Table 6 that the false positive rate for C is very high in 1994 (viz., 4.07 percent) and then drops considerably in 1995 and 1996. This explains the very high estimate of I for C in 1994 in Table 5. The higher estimates of both $\hat{\phi}_C$ and $\hat{\theta}_C$ in Table 6 for all three years explain the generally higher levels of inconsistency for C in Table 5. These results clearly illustrate the advantages estimates of error probabilities have over estimates of variance components in data quality investigations.

Biemer and Wiesen conducted additional analyses of the NHSDA data for these three years and discovered that the high false negative rate in measure A was the result of infrequent marijuana users who responded falsely to the recency question (A) but responded honestly to the frequency question (B). A possible explanation for this is that, through B, infrequent users are able to note their limited use of the drug (e.g., once or twice in last 12 months) rather than to code themselves simply as "users." Further analysis of the data provided additional support for this hypothesis. Likewise, further investigation of the high false positive rate for measure C in 1994 led to the discovery of a questionnaire problem that year that was repaired in the following years. Confusing question instructions and complex question wording led many nonusers of marijuana to classify themselves as users in 1994.

5. Discussion

When Hansen, Hurwitz, and Pritzker published their seminal work in 1964, the concept of simple response variance was as novel as it was revolutionary. HHP explained why two measurements of the same characteristics for the same individuals will often differ even though nothing has changed in the population. They provided the statistical methodology to allow statisticians to interpret these so-called gross differences so as to better understand what they imply about the quality of survey responses and survey estimates. Their simple statistical model extended theory of finite population sampling developed by Neyman and other early statisticians to include both randomization theory and nonsampling error theory. This greatly enlarged the contemporary view of the estimation process and provided survey methodologists with a framework for estimating and evaluating measurement errors. Their paper also emphasized the utility of measurement error variance estimates for the purpose of controlling and minimizing errors in surveys.

In Section 2 we showed how HHP's model for response error can be derived directly from the formulas for two-stage cluster sampling with equal size clusters. By recasting it

as a simple application of the cluster sampling, the sometimes challenging concept of response error modeling may be made more accessible to some students of nonsampling error theory. In viewing individuals as "clusters" of potential survey observations, simple response variance is analogous to within cluster variance. Its magnitude can be gauged by the index of inconsistency, which is the ratio of simple response variance to the total variance for a single random observation.

Since it was introduced in 1964, the index of inconsistency has been a key measure of data quality for many years. The importance of this parameter to survey methodologists is evidenced by the fact that no fewer than three interpretations of it exist: (a) as the proportion of total error that is simple response variance, (b) as $1-\kappa$, where κ is Cohen's agreement rate adjusted for chance, and (c) as Guggenmoos-Holzmann's parameter $\pi_{H=2}$, the proportion of the population belonging to an "inconclusive" domain under the agreement model with kappa constraints.

However, despite its importance, we have seen examples of its limitations. In the Census Dress Rehearsal race question example, we showed how the estimates of I based upon the parallel indicator assumptions are quite different from the latent class error model estimate that allows error distributions to differ by survey trial. HHP discussed in detail the consequences of the failure of the model assumptions to hold, but they did not have the statistical machinery for dealing with the problem. Latent class analysis provides a structure for estimating classification errors within a much broader range of survey conditions and in situations where conditions between the original and the reinterview surveys may differ in unknown ways. Although the basic concepts of latent class analysis were laid out by Lazarsfeld in the 1950's (Lazarsfeld 1950), the use of latent class analysis for obtaining estimates of inconsistency and reliability is a fairly recent development.

Although improved estimates of I are possible using an appropriate LCA model, I is quite limited as an indicator of data quality, and this was illustrated in the NHSDA past year marijuana use example. In that illustration, all estimates of I failed to identify important problems in the data – problems that were obvious when the individual classification error probability estimates that are combined to produce $\hat{I}_A$, $\hat{I}_B$, and $\hat{I}_C$ were considered. Those results beg the question: "Why bother to compute an estimate of I from classification probability estimates when the probability estimates themselves can provide more detailed information on data quality?"

We have seen that, even when only two indicators of a latent variable are available, it is possible to estimate the false positive and false negative probabilities using the approach of Hui and Walter (1980). When it is not possible to find one or more grouping variables that can satisfy the assumption of this model, the Guggenmoos-Holzmann agreement model can often be used, which provides a more general measure of inconsistency than I. Although the general agreement model is not identifiable with two dichotomous indicators, it is identifiable for two trichotomous or three dichotomous indicators.

This leads to the recommendation that analysis of reliability or inconsistency should not be limited to the traditional estimates of $\hat{I}$ or κ. There is much to be gained by exploring latent class analysis in situations that will allow the specification of plausible identifiable models. Further, LCA is a generalization of the traditional modeling approaches in the sense that, when the traditional assumptions hold, LCA and traditional analysis will often produce the same results. Guggenmoos-Hozmann's generalized kappa index is a good

example of this. Under certain parameter constraints, the parameter $\pi_{H=2}$ is equal to I; however, removing these parameter constraints produced a more general measure of inconsistency.

When the assumptions associated with traditional analysis do not hold, LCA may be the only valid method for assessing the error in the original measurements. The two illustrations clearly demonstrate that even when the repeated measurements were not intended either to be a replicate of the original measurement or to provide a gold standard measurement, LCA methods can be used to estimate the variance, bias, and the building blocks of these mean squared error components – viz., classification probabilities.

For data quality investigations, perhaps the best strategy is to use multiple approaches for assessing the magnitudes of the errors. In this regard, we recommend that the latent variable models be used to *supplement* traditional approaches rather than to *supplant* them.

6. References

Biemer, P.P. and Wiesen, C. (2002). Latent Class Analysis of Embedded Repeated Measurements: An Application to the National Household Survey on Drug Abuse. Journal of the Royal Statistical Society, Series A, 165(1), 97–119.

Biemer, P.P. and Stokes, L. (1991). Approaches to Modeling Measurement Error. In Measurement Errors in Surveys, P.P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). New York: John Wiley and Sons.

Biemer, P.P. and Woltman, H. (2001). Estimating Reliability and Bias from Reinterviews with Application to the 1998 Dress Rehearsal Race Question. Proceedings of the Federal Committee on Survey Methodology Conference, Washington, D.C.

Cochran, W.G. (1977). Sampling Techniques. 3rd Edition. New York: John Wiley and Sons.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurements, 20, 37–46.

Fuller, W.A. (1987). Measurement Error Models. NY: John Wiley and Sons.

Goodman, L.A. (1974). Exploratory Latent Structure Analysis Using Both Identifiable and Unidentifiable Models. Biometrika, 61, 215–231.

Guggenmoos-Holzmann, I. (1996). The Meaning of Kappa: Probabilistic Concepts of Reliability and Validity Revisited. Journal of Clinical Epidemiology, 49, 775–782.

Guggenmoos-Holzmann, I. and Vonk, R. (1998). Kappa-like Indices of Observer Agreement Viewed from a Latent Class Perspective. Statistics in Medicine, 17, 797–812.

Hui, S.L. and Walter, S.D. (1980). Estimating the Error Rates of Diagnostic Tests. Biometrics, 36, 167–171.

Hagenaars, J.A. (1988). Latent Structure Models with Direct Effects Between Indicators: Local Dependence Models. Sociological Methods and Research, 16, 379–405.

Hagenaars, J. (1993). Loglinear Models with Latent Variables. Sage University Paper Series, Quantitative Applications in the Social Sciences, 07-094. Newbury Park, CA: Sage.

Hansen, M., Hurwitz, W.N., and Pritzker, L. (1964). The Estimation and Interpretation of Gross Differences and the Simple Response Variance. In Contributions to Statistics

(presented to P.C. Mahalanobis on the occasion of his 70th birthday), C.R. Rao (ed.). Calcutta: Statistical Publishing Society.

Hess, J., Singer, E., and Bushery, J. (1999). Predicting Test-Retest Reliability from Behavior Coding. International Journal of Public Opinion Research, 11, 346–360.

Lazarsfeld, P.F. (1950). The Logical and Mathematical Foundation of Latent Structure Analysis. In Measurement and Prediction, S.A. Stouffer, et al. (eds). Princeton, NJ: Princeton University Press, 362–412.

McCutcheon, A. (1987). Latent Class Analysis, Sage University Paper 64, Sage, Newbury Park, CA.

Rothman, K.J. and Greenland, S. (1998). Modern Epidemiology, 2nd Edition, Lippincott-Raven, Hagerstown, MD.

U.S. Census Bureau (1985). Evaluation of Censuses of Population and Housing, STD-ISP-TR-5, Washington, D.C., U.S. Government Printing Office.

Vermunt, J. (1997). ℓEM: A General Program for the Analysis of Categorical Data. Tilburg University.

Received February 2004