Journal of Official Statistics, Vol. 12, No. 3, 1996, pp. 275–300

Estimation of Measurement Bias in Self-Reports of Drug Use with Applications to the National Household Survey on Drug Abuse

Paul Biemer[1] and Michael Witt[1]

Direct estimates of response bias in self-reports of drug use in surveys require that essentially error free determinations of drug use be obtained for a subsample of survey respondents. The difficulty of obtaining determinations which are accurate enough for estimating validity is well-documented in the literature. Methods such as specimen (hair, urine, etc.) analysis, proxy reports, and the use of highly private and anonymous modes of interview all have to contend with error rates which may only be marginally lower than those of the parent survey. Thus, any methodology for direct validity estimation must rely to some extent on approximations and questionable assumptions.

In this article, we consider a number of methods which rely solely on repeated measures data to assess response bias. Since the assumptions associated with these approaches do not require highly accurate second determinations, they may be more easily satisfied in practice. One such method for bias estimation for dichotomous variables that is considered in some detail provides estimates of misclassification probabilities in the initial measurement without requiring that the second measure be accurate or even better than the first. This methodology does require, however, that two subpopulations exist which have different rates of prevalence but whose probabilities of false positive and false negative error are the same.

The applicability of these methods for self-reported drug use will be described and illustrated using data from the National Household Survey on Drug Abuse. In the discussion of the results, the importance of these methods for assessing the validity of self-reported drug use will be examined.

Key words: Latent class models; reliability; repeated measures; validity.

1. Introduction

The self-report is an integral component of the research methodology for measuring the prevalence of substance abuse and other stigmatized behaviors. While there is a growing body of literature that supports the validity of the self-report, there are also studies that question its validity (see Mieczkowski 1991 for a review of validation research in this area). These studies suggest that response validity for drug use is highly dependent upon the construction of the questions, the procedures for administration, the perceived intentions of the investigators, and the cognitive fitness of the respondent. Given the importance of monitoring drug use prevalence, trends, and risk factors for the U.S. population, considerable research has been conducted to improve the validity of the self-report for sensitive topics, for example,

[1] Research Triangle Institute, Research Triangle Park, NC 27709, U.S.A.

Acknowledgements: The authors gratefully acknowledge the assistance of Michael Sinclair, who generously provided them with a copy of his Doctoral dissertation and the SAS software he developed for the Hui-Walter procedure.

by using more private or anonymous reporting methods, attempting to motivate honest reporting by incentives or personal appeals, and so on (see, for example, Lessler and O'Reilly, in press, and Tourangeau, Jobe, Pratt, and Rasinski, in press).

For comparing the accuracy of alternative data collection methodologies for obtaining self-reports, some information on the reporting error associated with the measurement processes is required. If the objective of a methodological study is to estimate the magnitude of the measurement bias, then error-free determinations of drug use are typically required for a sample of study subjects. For other methodological studies, it may only be necessary to obtain determinations which have better measurement error properties than the methodologies that are being evaluated. If no criterion data are available for estimating measurement bias, it is sometimes sufficient to know the direction of the reporting bias in order to select the best data collection method. As an example, in many cases it is reasonable to assume that stigmatized behaviors will be generally underreported by the study population. In such cases, the data collection methodology that produces the highest prevalence rate is deemed the most valid method. (See Biemer 1988 for a critique of this approach.)

As the preceding discussion affirms, measurement error evaluation methodology is critical for the improvement of the survey design and survey methods. In addition, evaluation methods are used to assess the components of total error in the reported estimates from drug use prevalence studies and these data help define the limitations of the survey results for policy decisions and other uses of the data. Yet, all methods for estimating validity, reliability, and response bias are themselves subject to questions of validity.

This article focuses on a number of methods for assessing the validity of self-reports of drug use. In particular, the discussion is confined to methods for estimating measurement bias that rely on repeated measurements of the same characteristics for the same individuals. Examples are: reinterview methods, test-retest, record check studies, and biological test validation methods. In Section 2, a number of measurement error indicators and measures are reviewed that have been used in the literature to describe the measurement accuracy and precision of survey data. In Section 3, several approaches for estimating these measurement error indicators using repeated measurements methods are presented. Using a general, two-measurement model for measurement error, each estimation approach is seen as a design for restricting the parameter space of the overspecified, general model by setting some parameters to zero and others to the same, unknown constant. These restrictions then impose requirements on the evaluation designs that must be met in order for the model assumptions to hold.

A substantial part of this article presents the results of an evaluation of a recently developed statistical method for estimating false positive and false negative reports from repeated measurement studies. The method was developed by Hui and Walter (1980) for the evaluation of medical diagnostic testing procedures. Sinclair and Gastwirth (1993) applied the method for the evaluation of survey measurements and extended the methodology in ways that enhance the method's applicability for survey evaluation. The method provides estimates of misclassification probabilities in the initial measurement without requiring that the second measurement be without error or even more accurate than the first measure. The methodology does require, however, that two domains can be defined that have different rates of prevalence but have probabilities of misclassification that are identical.

For this application, the method is applied to data from the National Household Survey on Drug Abuse (NHSDA) in order to estimate the misclassification errors associated with self-reports of alcohol, marijuana, and cocaine use. False negative and false positive probabilities (and their standard errors) by various demographic subgroups and geographic areas are presented in Section 4.

Finally, in Section 5, the results of the application of repeated measurement methods for the evaluation of self-reports of drug use are summarized and conclusions regarding their application to the NHSDA are presented.

2. Review of the Measurement Error Terminology

In this section, several measurement error concepts that are relevant to the study of self-reported drug use are reviewed. The study is restricted to the error in a single dichotomous response variable, denoted by y, since this type of response is quite often encountered in drug use measurement. As an example, y may denote a ''yes or no'' response regarding the use of specific drugs during some period or it may denote a response to a category of use in a multiple category response set. Let y_i denote the measurement for some characteristic associated with the ith survey respondent where $y_i = 1$ if the respondent possesses the characteristic and $y_i = 0$ otherwise. Let μ_i denote the corresponding true value for the ith respondent. Following Bross (1954) and Cochran (1968), assume that the response y_i is determined from μ_i by means of a random process governed by parameters θ and ϕ, as follows

$$\begin{aligned}\theta &= P(y_i = 0|\mu_i = 1)\\ \phi &= P(y_i = 1|\mu_i = 0)\end{aligned} \qquad (1)$$

where θ and ϕ are referred to as the *probability of a false negative* and the *probability of a false positive*, respectively. Further assume that the expected value of y_i given μ_i is

$$E(y_i|\mu_i) = \mu_i(1-\theta) + (1-\mu_i)\phi. \qquad (2)$$

Measurement Bias. Let $\pi = E(\mu_i)$, the true prevalence of the characteristic in the target population and let $P = E(y_i)$, the expected observed prevalence. Let the *measurement bias* of the measure, y, be defined as $B(y) = P - \pi$. Thus, from (2)

$$B(y) = -\pi\theta + (1-\pi)\phi. \qquad (3)$$

From (3), it can be seen that the measurement bias is 0 or small relative to π if either: (a) $\theta \approx \phi \approx 0$ or (b) $\pi\theta \approx (1-\pi)\phi$. Condition (a) implies that there is almost no chance for a misclassification error. Condition (b) implies that the expected number of false positive errors in the population approximately equals the expected number of false negative errors. As Cochran (1968) points out, this latter condition is quite unlikely in most applications so that zero measurement bias is usually an indication that condition (a) holds.

Note that for drugs with low prevalence rates such as cocaine and heroin, π will be many times smaller than $(1-\pi)$ and a relatively small false positive rate can have large consequences on the bias. As an example, suppose $\pi = .010$, $\theta = .30$, and $\phi = .010$. Then, $\pi\theta = .010 \times .30 = .0030$ while $(1-\pi)\ \phi = .99 \times .010 = .0099$. Thus, the contribution to bias due to false positives is 3.30 times larger than the contribution due to false negatives although the probability of a false negative is thirty times greater than the probability of a false positive. Using (3), $B(y) = -.0030 + .0099 = .0069$. The *relative bias*, defined by $RB(y) = B(y)/\pi$, is $.0069/.010 = .69$. This implies that the true prevalence rate, π, will be overstated by 69% on average. Thus, for rare drugs, the consequences of even a small false positive rate can be substantial.

Measurement bias is important in survey work because it is directly related to the bias in estimators of means, proportions, and totals. Let $p = \Sigma y_i/n$ denote the sample proportion for a simple random sample. Then the bias in p for estimating π, the true population proportion, is defined as $E(p - \pi)$ which is also given by (3).

Reliability. Roughly speaking, reliability refers to the degree of consistency of responses from independent, replicated measurements of the same characteristic. The statistical definition of the *reliability ratio, R,* is the proportion of the variance that is *not* measurement variance, where *measurement variance* is defined as $E[\text{Var}(y_i|i)]$; i.e., the average variance within respondents and between hypothetical, independently replicated measurements. Thus, R can be expressed mathematically as

$$\begin{aligned} R &= 1 - \frac{E[\text{Var}(y_i|i)]}{\text{Var}(y_i)} \\ &= \frac{\text{Var}[E(y_i|i)]}{\text{Var}(y_i)}. \end{aligned} \tag{4}$$

Biemer and Stokes (1991) show that, for dichotomous responses, R can be quite difficult to interpret since it is a complex function of the misclassification probabilities and π. They show that under the model in (2)

$$R = 1 - \frac{\pi\theta(1-\theta) + (1-\pi)\phi(1-\phi)}{PQ} \tag{5}$$

where $P = \pi(1-\theta) + (1-\pi)\phi$ and $Q = 1 - P$. Further, they show that:

a. For two domains or subpopulations having identical probabilities of misclassification, the reliability ratio for one domain can be substantially larger than the reliability for the other solely as a consequence of the difference in their respective prevalence rates.

   As an example, suppose that for Domain 1, $\pi = .50$ while for Domain 2, $\pi = .10$. Further, let $\theta = 0.00$ and $\phi = .10$ for both domains. Then, using (5), $R = .82$ for Domain 1 and $R = .47$ for Domain 2. On this basis, it would be wrong to conclude that the responses from Domain 1 are of "higher quality" than those from Domain 2. Thus, in this respect, R can be misleading as an indicator of data quality.

b. From (5), it can be shown that the reliability ratio can be very high although there is a large amount of misclassification error.

   As an example, suppose the false positive probability is zero ($\phi = 0$) while the false negative rate is high, say 10% ($\theta = .10$). Further suppose that $\pi = .050$. This situation is often encountered in drug use measurement for rarely used drugs. Then it can be shown that $R = .90$ suggesting very high reliability in the measure. Further, the relative bias in the measure is -10%, which is nontrivial.

While good reliability is not necessarily an indicator of good data quality, poor reliability is usually an indication that the measure is subject to a large measurement bias. This is especially true when the prevalence rate is small, as with cocaine or heroin use. As an example, consider the case where $\theta = .10$, $\phi = .025$, and $\pi = .050$. Here, $1 - R = .43$ and the relative bias of the measure defined above is $RB(y) = .38$ or 38%. This correspondence between $I = 1 - R$, called the *index of inconsistency*, and the relative bias for small prevalence rates is not coincidental. By comparing (5) to (3) divided by π, it can be verified that I and $RB(y)$ will be close whenever π is small. Further, when π is small, the cause of poor reliability is a high and disproportionate number of false positives compared with the number of false negatives in the population. To illustrate, in the example, the expected number of false negatives in the population is $N \times .050 \times .10 = .0050N$, where N is the population size. This compares with $N \times .95 \times .025 = .024N$ false positives – approximately five times as many false positives than false negatives. Thus, one may conclude that when the prevalence of the characteristic is small, poor reliability is usually an indication of a large positive bias in the estimator of the prevalence rate. By a similar argument, one can conclude that when the prevalence rate is large (say, $\pi > .90$), poor reliability is usually an indication of a large negative bias in the estimator due to a high false negative rate. For π between .10 and .90, poor reliability is an indication of a large expected number of false positives or false negatives. However, little can be said regarding the direction of the bias or whether the net effect of misclassification error results in either a small or large bias in the estimator of the prevalence rate.

To summarize, this discussion shows that in some situations, the reliability ratio can be a good indicator of measurement and estimator bias. Further, a large value for the estimator of R is no assurance of good data accuracy. A low value of R is an indication of large misclassification errors in the data. Finally, in some situations, R can help us to determine whether the misclassification error problem is a result of high false negative or high false positive probabilities.

Validity. Bohrnstedt (1983) states that validity is an indicator of ''the degree to which an instrument measures the construct under investigation.'' He discusses a number of alternative concepts of validity proposed in the psychometric literature for describing data quality. Some of these are *predictive validity, concurrent validity, empirical validity, and theoretical validity.* These concepts and others are discussed in some detail in Groves (1989). Of particular relevance to the present discussion is theoretical validity (TV) which, in terms of the model is defined as the correlation between the observed measure and the conditional expectation of the observed measure, called the *true score.* Thus

$$TV = \mathrm{Corr}[E(y_i|\mu_i), \mu_i] \tag{6}$$

Like most validity concepts, theoretical validity is defined as a correlation between two *constructs* (i.e., measures or true scores). Because validity does not depend upon the existence of a true value, it is the preferred indicator for describing the quality of measures of psychological states, attitudes, or knowledge. Biemer and Stokes (1991) show, under the error model proposed above for dichotomous data, that $(TV)^2 = R$, the reliability ratio. They further show that under more general models, $(TV)^2 \leqslant R$. Thus, reliability is an upper bound on theoretical validity and thus, a measure may be reliable but lack validity. This is similar to the result shown for measurement bias: a measure may be reliable and still substantially biased. This result further implies that an unreliable measure cannot be valid. Note, however, that an unreliable measure may still be unbiased. For the classification error models considered here, reliability and validity, while conceptually different, are mathematically equivalent indicators. Thus, as a measure of data quality for categorical variables, the limitations of the reliability ratio are also limitations of validity measures.

It is not uncommon to find the terms validity and measurement bias used synonymously. It is important to note that these concepts are quite different. As an example, if some positive number, *C*, is added to every measurement, the validity of the measure is unchanged while bias is increased by *C*. The advantage of validity as an indicator of data quality is that, unlike measurement bias, validity does not require that true values exist for the constructs under study.

Mean Squared Error. Whereas measurement bias, reliability, and validity attempt to describe the error in an individual response, the mean squared error (MSE) aims at describing the error in an estimator. The mean squared error of an estimator is a measure of accuracy that is often used for estimators of population parameters. Let $\hat{\pi}$ be any estimator of π, then

$$\mathrm{MSE}(\hat{\pi}) = \mathrm{Bias}(\hat{\pi})^2 + \mathrm{Var}(\hat{\pi}) \tag{7}$$

the sum of the square of the bias and the variance. Suppose that $\hat{\pi}$ is the simple expansion mean under simple random sampling, denoted by p. As mentioned above, the bias in p is $B(y)$ defined in (3). Biemer and Stokes (1991) show that for small samples from large populations

$$\mathrm{Var}(p) = PQ/n \tag{8}$$

where P was defined before as $E(y_i)$. The usual unbiased estimator of the variance when there are no classification errors is

$$v(\bar{y}) = pq/(n-1) \tag{9}$$

where $q = 1 - p$. Under Bross's model, the usual variance estimator is still unbiased in the presence of measurement error. To see this, note that $E(pq) = E(p) - E(p^2) = P - [\mathrm{Var}(p) + P^2] = PQ(1 - 1/n)$ by (8), and the result follows. It will be shown subsequently that this is not true under more general models.

These variance formulas show that misclassification error can, sometimes, result in smaller variances for estimators of proportions and totals. Consider the situation where $\pi = .5$. In this situation, the variance of the sample proportion is at its maximum. Thus, misclassification can only reduce the variance. One exception to this is when the misclassification errors are correlated, as happens with interviewer error. If interviewers exert influence over the misclassification error for respondents in their assignments, then misclassification errors are correlated and (7), which was derived under the assumption of unit to unit independent misclassification error, no longer holds. Under a more appropriate model for this situation, misclassification error always results in an increase in estimator variance. Further, the usual estimators of variance may be substantially biased. Biemer and Stokes (1991) discuss models that are appropriate for the study of interviewer errors and dichotomous response variables.

3. Estimation of Reliability and Bias

This section considers methods for estimating the components of error for the dichotomous measurement error model. These methods are test-retest, true value measurements, and repeated measurements. The assumptions underlying these methods will be discussed in terms of a general model for two measurements. Models for multiple measurements are essentially extensions of this basic model.

Let y_{ti} denote the tth measurement on unit i for $t = 1,2$ and $i = 1,..., n$. In analogy to the single measurement model, assume the following:

General Model Assumptions.

i. $P(y_{ti} = 0|\mu_i = 1) = \theta_t, t = 1,2$
ii. $P(y_{ti} = 1|\mu_i = 0) = \phi_t, t = 1,2$
iii. $P(y_{2i} = 0|y_{1i} = 1, \mu_i = 1) = \theta_{0|1}$
iv. $P(y_{2i} = 0|y_{1i} = 0, \mu_i = 1) = \theta_{0|0}$
v. $P(y_{2i} = 1|y_{1i} = 0, \mu_i = 0) = \phi_{1|0}$
vi. $P(y_{2i} = 1|y_{1i} = 1, \mu_i = 0) = \phi_{1|1}$

Note that $\theta_{0|1} = \theta_{0|0} = \theta_2$ and $\phi_{1|0} = \phi_{1|1} = \phi_2$ if and only if the false negative errors and the false positive errors, respectively, corresponding to measurements 1 and 2 are independent. Under the general model, the probabilities of misclassification may differ between trials (assumptions i and ii) and further, the second trial outcomes are not independent of the first trial outcomes (assumptions iii–vi). Including π, there are seven parameters associated with this model. However, only three degrees of freedom are available for estimation for a dichotomous variable with two measurements. Thus, as will be shown subsequently, additional assumptions are needed to estimate any of the parameters.

3.1. Test-retest methods

As discussed in the previous section, although the interpretation of the reliability ratio is difficult in the dichotomous case, estimates of reliability contain some information on bias that can be useful in studies of the accuracy of drug use measurement. The most common method of estimating reliability for self-reports is the *test-retest method.* This method includes studies that embed replicate measures of the same characteristic within a single interview and also reinterview studies. In reinterview studies, a subsample of the original respondents is recontacted for the purpose of obtaining a set of second measurements for the original interview characteristics.

Let $t = 1$ denote the first measurement and let $t = 2$ denote the second measurement or reinterview response. For the test-retest measurement model, the assumptions of the general model are replaced by the following:

Test-Retest Assumptions.

(i) Independence

$\theta_{0|1} = \theta_{0|0} = \theta_2$

$\phi_{1|0} = \phi_{1|1} = \phi_2$

(ii) Homogeneity

$$\theta_1 = \theta_2$$

$$\phi_1 = \phi_2$$

Assumption (i), which replaces assumptions (iii) to (vi) in the general model, essentially states that the errors in the two measurements are independent. That is, whether a false positive or false negative error is made for the second measurement does not depend upon whether an error was committed for the first measurement. For embedded replication there is a risk that respondents may simply repeat the erroneous response on the second measurement made on the first. When the second measurement is collected after some time has elapsed since the first measurement, this is less of a risk. Yet, as several researchers have shown, correlated errors can persist even when the reinterview is conducted weeks after the initial interview (Bailar 1968, O'Muircheartaigh 1991).

Assumption (ii), which replaces assumptions (i) and (ii) in the general model, states that the false positive and false negative probabilities are the same for both measurements. Thus, the aim of the design of the second measurement is to replicate the first measurement by, for example, using identical procedures, questions, interviewer competencies, etc. For reinterview surveys where the second measurement is obtained in a separate interview with the respondent, the reinterview design should replicate to the extent possible, the same essential survey conditions that existed in the first interview. For replicate measures embedded with the same instrument, this assumption is more easily satisfied. Despite the potential difficulties with the test-retest assumptions, the method remains the most commonly used technique in survey methodology for estimating reliability.

To define an estimator of the reliability ratio, R, for dichotomous data, let a, b, c, and d denote the cell counts for the 2×2 measurement cross-classification table, as follows:

		y_{1i}		
		1	0	
	1	a	c	
y_{2i}	0	b	d	
				n

$$p_1 = \frac{a+b}{n},\ p_2 = \frac{a+c}{n},\ \text{and } n = a+b+c+d$$

Fig. 3.1. Measurement 1 by Measurement 2 Cross-Classification

Then, an estimator of R is $\hat{R} = 1 - I$ where

$$\hat{I} = \frac{(b+c)/n}{p_1 q_1 + p_2 q_2}$$

where $q_t = 1 - p_t$, $t = 1, 2$ and $\hat{I}$ is an estimator of the index of inconsistency (see U.S. Bureau of the Census 1985). These estimators assume that respondents are sampled using a simple random sampling design; however, for more complex sampling designs, weighted cell counts are typically used to estimate R.

It has been shown (U.S. Bureau of the Census 1985) that when the test-retest assumptions are not satisfied, the estimates of R can be substantially biased. Violations of assumption (i) are usually due to errors that are positively correlated between trials. In this situation, $\hat{R}$ is an overestimate of R. In U.S. Bureau of the Census (1985), it is shown that the bias in R is approximately $\rho_T I$ where ρ_T is the between trial correlation. As an example, if $R = .70$ and $\rho_T = .20$ then the bias in $\hat{R}$ is approximately $.20(.30) = .06$ and, thus, $\hat{R}$ overestimates R by approximately 6%. When assumption (ii) is violated, $\hat{R}$ estimates a complex function of the reliabilities associated with each trial. Thus, interpretations of R based upon the above model can be misleading in these situations.

3.2. True value measurement methods

To estimate measurement bias and the misclassification probabilities for self-reports, the traditional methodology has relied upon true value measurements. For drug use measurement, true values have been obtained from:

- administrative records, such as arrest records and drug treatment reports,
- hair, urine, and other specimen analyses to detect the presence of drugs in the specimens, and
- reinterviews using better methods than were used in the first interview, such as more private modes of interview, neutral (out-of home) settings, and better question design.

With any of these methods, the usual modeling approach is to assume the following:

True Value Assumptions.

$$\theta_{0|1} = \theta_{0|0} = \theta_2 \equiv 0$$

$$\phi_{1|0} = \phi_{1|1} = \phi_2 \equiv 0.$$

That is, it is assumed that the second measurement is the true value, or mathematically, $y_{2i} = \mu_i$. Thus, an estimator of the bias in the measurement y_I, assuming simple random samples, is

$$\hat{B}(p_1) = p_1 - p_2. \tag{10}$$

If y_2 in Figure 3.1 now denotes the true value, then using the notation for the cell counts in that table the estimates of the false negative and false positive probabilities are, respectively,

$$\hat{\theta}_1 = \frac{c}{a+c}$$

and

$$\hat{\phi}_1 = \frac{b}{b+d}.$$

As before, weighted counts may be used for unequally weighted samples.

Occasionally, the assumptions for the true value model hold only approximately and a more appropriate set of assumptions is:

Improved Measurement Assumptions.

i. Independence

$$\theta_{0|1} = \theta_{0|0} = \theta_2$$

$$\phi_{1|0} = \phi_{1|1} = \phi_2$$

ii. Improved Second Measurement

$\theta_2 < \theta_1$ and $\phi_2 < \phi_1$

In words, it is assumed that the second measurement is not free of error, but that the probability of error in the second measurement is smaller than that for the first measurement. Furthermore, the errors in both measurements are independent. Under these assumptions, it can be shown that if $B(p_1)$ and $B(p_2)$ have the same sign

$$|E[\hat{B}(p_1)]| < |B(p_1)|$$

where $\hat{B}(p_1)$ is given by (10). Thus, the usual estimator of bias is biased downward. However, if $B(p_2) \approx 0$, then $\hat{B}(p_1)$ may still provide a useful approximation for $B(p_1)$.

It should be noted that, under the improved measurement assumptions, the estimators $\hat{\theta}_1$ and $\hat{\phi}_1$ given above for the true value model are both biased and the directions of the biases are unknown. However, in this situation the estimation method discussed in the next section can be used to estimate the misclassification probabilities associated with both the first and second measurements.

3.3. Repeated measurements: the Hui-Walter method

In some studies, two or more measurements of μ_i are available for a sample of respondents; however, the assumptions made for test-retest and true value models are not tenable. For example, the second measurement is not perfect or even better than the first measurement. Neither is it plausible to assume that the second measurement is a replication of the first. Hui and Walter (1980) consider this situation in the evaluation of diagnostic tests. In this situation, the presence or absence of a disease may be indicated by two tests, each having probabilities of misclassification that are nonzero, nontrivial, and procedure dependent. Sinclair and Gastwirth (1993) applied the Hui-Walter estimation methodology for estimating the measurement error in self-reports in the evaluation of labor force characteristics in the Current Population Survey (CPS). Here the method is considered for the estimation of the false positive and false negative probabilities for self-reported drug use.

Consider the case where two measurements are taken from each individual in two subpopulations or domains indexed by g. For each domain g, let A_g, B_g, C_g, and D_g denote the four cells in Figure 3.1 as follows: A_g = cell (1,1), B_g = cell (1,0), C_g = cell (0,1), and

$D_g =$ cell (0,0). Then the probability that a randomly selected individual from domain g is classified in each cell is as follows:

$$P(A_g) = \pi_g(1-\theta_{g,1})(1-\theta_{g,0|1}) + (1-\pi_g)\phi_{g,1}\phi_{g,1|1}$$
$$P(B_g) = \pi_g(1-\theta_{g,1})\theta_{g,0|1} + (1-\pi_g)\phi_{g,1}(1-\phi_{g,1|1})$$
$$P(C_g) = \pi_g\theta_{g,1}(1-\theta_{g,0|0}) + (1-\pi_g)(1-\phi_{g,1})\phi_{g,1|0}$$
$$P(D_g) = \pi_g\theta_{g,1}\theta_{g,0|0} + (1-\pi_g)(1-\phi_{g,1})(1-\phi_{g,1|0}).$$

Assuming independence in the classifications between the two domains, the probability of observing $a_g, b_g, c_g,$ and d_g for $g = 1,2$ is therefore

$$l = \prod_{g=1}^{2} P(A_g)^{a_g} P(B_g)^{b_g} P(C_g)^{c_g} P(D_g)^{d_g}.$$

This likelihood function contains 14 parameters and only (2×3 =) 6 degrees of freedom for estimation. To reduce the number of parameters Hui-Walter and Sinclair-Gastwirth assume the following:

Hui-Walter Independence Assumptions

i. Independence

$$\theta_{g,0|1} = \theta_{g,0|0} = \theta_{g,2}, \ g = 1,2$$
$$\phi_{g,1|0} = \phi_{g,1|1} = \phi_{g,2}, \ g = 1,2$$

ii. Homogeneity between domains

$$\theta_{1,t} = \theta_{2,t} = \theta_t, \ t = 1,2$$
$$\phi_{1,t} = \phi_{2,t} = \phi_t, \ t = 1,2.$$

In words, this assumption says that

- misclassification probabilities differ between the two measurements, but are the same for both domains ($g = 1,2$),
- the prevalence rates differ between domains, and
- misclassification errors are independent between trials.

These assumptions reduce the number of parameters to six, viz., θ_1, θ_2, ϕ_1, ϕ_2, π_1, and π_2. A solution for this formulation can be obtained using maximum likelihood estimation. This model shall be referred to as the Hui-Walter independence model.

The assumption of equal error rates across domains is easily justified for many diagnostic tests of the types discussed by Hui and Walter. Their example considers two tests for the detection of tuberculosis which exhibit the same error distributions across socio-economic subgroups. In the survey setting, the misclassification errors may be highly correlated with the prevalence rates. Therefore, it is important to choose the two domains carefully to ensure proper application of these methods.

For their application to the CPS, Sinclair and Gastwirth define the two domains based on race and sex: white males and white females. Thus, it is not necessary that the two domains partition the entire population. Although the results of their study only apply to these two domains, important insights may be gleaned for the entire population by studying this part of it. Sinclair and Gastwirth demonstrate the importance of defining the two domains such that their respective prevalence rates for the characteristic of interest are markedly different. Since the characteristic of interest in their study was labor force participation, their choice of race and sex would seem appropriate since labor force participation rates are considerably higher for white males (π_1=.75) than for white females ($\pi_2 = .55$). Further, the assumption of equal error probabilities for the two domains is also plausible: both domains are administered the same questions by the same interviewers using the same survey procedures. However, the assumption of independence between the errors for the two trials may not be justified. O'Muircheartaigh (1991) estimates that the between trial correlation for labor force participation varies in the interval [.3,.5] when the second measurement is obtained using a replicate reinterview survey. Sinclair and Gastwirth consider the effects of between trial correlations on the resulting estimates and conclude that failure of this assumption to hold can result in large biases in the estimates of the error probabilities.

In this application to self-reported drug use, the estimates using the Hui-Walter independence model as well as a ''dependent'' model are compared and evaluated. The latter model is similar to the one proposed by Vacek (1985); however, it uses fewer parameters and therefore requires fewer degrees of freedom to estimate. For the dependent model the following is assumed:

Dependent Model Assumptions.

i. Homogeneous False Negative Probabilities

$$\theta_{g,1} = \theta_1, \; g = 1,2$$
$$\theta_{g,0|1} = \theta_{0|1}, \; g = 1,2$$
$$\theta_{g,0|0} = \theta_{0|0}, \; g = 1,2$$

ii. Independent and Homogeneous False Positive Probabilities

$$\phi_{g,1|0} = \phi_{g,1|1} = \phi_2, \; g = 1,2$$
$$\phi_1 = \phi_2 = \phi, \; g = 1,2.$$

Thus, it is assumed that a single false positive rate applies to both trials and both domains and, further, that the false positive errors are independent between both trials. Finally, it is assumed that the false negative errors are correlated between trials and that these correlations are equal for the two domains. As with the independent model, the dependent model provides for six parameters, viz., θ_1, $\theta_{0|1}$, $\theta_{0|0}$, ϕ, π_1, and π_2, all of which are estimable.

The rather restrictive assumptions regarding the false positive errors are justified because, for most of the drugs in this study, the false positive rates are expected to be quite small. In this situation, it may be reasonable to assume that $\phi = 0$ rather than estimate ϕ. However, by allowing ϕ to be estimated, it is hoped that the likelihood function is increased and, thus, the estimates of the more important false negative probabilities are improved.

4. Application of the Hui-Walter Method to the National Household Survey on Drug Abuse

In this section, the Hui-Walter method is implemented to estimate the false negative and false positive probabilities associated with the so-called recency question in the National Household Survey on Drug Abuse (NHSDA). The recency question asks respondents about the most recent time they used a particular drug. For this study, the measurement bias for this question was evaluated for alcohol, marijuana, and cocaine. By design, the NHSDA contains many redundant questions regarding drug use recency, particularly lifetime use. Because of this redundancy, the application of the Hui-Walter method to estimate NHSDA misclassification error is possible. In this section, the use of this methodology for assessing the accuracy of self-reports is demonstrated and the characteristics exhibited by the Hui-Walter estimates are critically examined.

4.1. Description of the NHSDA

The NHSDA is a multistage, household survey designed to measure the population's current and previous drug use activities. The 1993 survey was the 13th study conducted in a series initiated in 1971. Since 1990, the survey is conducted annually, with distinct samples of households and persons selected each year. In October 1992, sponsorship of the survey was transferred from the National Institute on Drug Abuse (NIDA) to the Substance Abuse and Mental Health Services Administration, Office of Applied Studies (SAMHSA/OAS) where it currently resides.

For this research project, data from the 1991, 1992, and 1993 surveys were used in the analysis, a total of 88,000 interviews. Subsequent discussions of the NHSDA will be restricted to design and implementation issues related to these surveys.

4.1.1. Survey design and data collection

The NHSDA is based on a national probability sample of dwelling units in the United States. For the 1991, 1992, and 1993 studies, approximately 118 primary sampling units (PSUs) were selected at the first stage of sampling. These PSUs represent geographic areas in the United States; generally defined as counties, groups of counties or Metropolitan Statistical Areas (MSAs). At the second stage of selection, smaller geographic areas within each PSU called *segments* were selected. The NHSDA segments were defined by joining adjacent census blocks within each PSU. At the third stage of selection, a sample of dwelling units was selected within each segment and a resident of each occupied, sampled dwelling unit was asked to participate in a screening interview for this survey. Results from this personal visit, screening interview are used to randomly select up to two members of each household. Each selected person was then asked to participate in the personal visit, interview phase of the survey. Data on a person's current and previous drug use activities are collected during this interview phase of the survey.

The target population includes persons ages 12 years or older who live in households, certain group quarters (e.g., college dormitories, homeless shelters) and civilians living on military installations. Active military personnel and most transient populations, such as homeless people not residing in shelters, were not included. The annual sample for the 1991, 1992, and 1993 surveys is approximately 30,000 persons. Hispanics, Blacks, younger persons and the residents of six of the MSAs are oversampled to ensure that the sample sizes are adequate to produce the subpopulation estimates of interest.

Drug and demographic data are collected from each respondent during the interview phase using a combination of interviewer administered and self-administered instruments. On average, the interview takes about an hour to complete. The interview begins with a set of interviewer administered questions designed to collect data on the respondent's current and previous use of cigarettes and other forms of tobacco. These initial questions allow the respondent to become familiar with the format of the NHSDA.

The remainder of the questionnaire is divided into sections corresponding to each drug of interest: alcohol, the nonmedical use of sedatives, tranquilizers, stimulants and analgesics, marijuana, inhalants, cocaine, crack, hallucinogens, and heroin. For each section, the interviewer gives the respondent an answer sheet and asks him/her to record his/her responses on it. Depending on the complexity of an answer sheet, the interviewer will either read the questions to the respondent or, if preferred, the respondent can read the questions. Upon 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 and, more importantly, to guarantee response privacy.

Most of the answer sheets are designed so that even respondents who have never used a particular drug still need to answer each question about the drug. Since both users and nonusers 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 nonuser based on the time 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. This redundancy in the questionnaire, therefore, provides additional information regarding drug use that can be used to compensate for underreporting for the direct question.

4.1.2. Data editing and estimation

The raw NHSDA data are extensively edited to ensure the internal consistency of drug use responses. For the 1991, 1992, and 1993 surveys, this editing was based on a ''most recent indication of use'' rule. As described in the previous section, all respondents are required to respond to essentially the same questions regardless of their drug use. Consequently, use of a particular drug during a particular reference period can be logically established from responses to various questions. These questions include those items presented on the specific drug answer sheet, as well as several items on the latter answer sheets asking about general drug use activities.

For any particular drug, the logical editing begins with the *drug recency question*, a question at the beginning of each drug answer sheet that asks the respondent about the most recent time he/she used a particular drug. As an example, on the alcohol answer sheet the recency question is:

When was the most recent time that you had an alcohol drink, that is, of beer, wine, or liquor or a mixed alcoholic drink?

Within the past month (30 days)
More than 1 month ago but less than 6 months ago
6 or more months ago but less than 1 year ago
1 or more years ago but less than 3 years ago
3 or more years ago
Never had a drink of beer, wine, or liquor in your life

Thus, the recency question is used to establish the most recent time a drug was used. At this first stage of editing, the recency response categories are collapsed and for each drug, the respondents are classified into one of the following mutually exclusive categories: a past month user, past year user, lifetime user or not a lifetime user of the drug under question. Under these editing rules, that past year users do not include past month users and lifetime users do not include past year or past month users.

After this recoding is completed, it is checked against the responses to all other questions from which drug recency can be implied. These questions include drug use related questions which are asked on the specific drug answer sheet, as well as questions asked on the latter, drug use activities answer sheets. For example, alcohol use can be implied from other questions on the alcohol answer sheet such as:

- *About how old were you when you first began to drink beer, wine or liquor once a month or more often?* [This question can be used to establish lifetime use of alcohol]
- *On the average, how often in the past 12 months have you had any alcoholic beverage, that is, beer, wine or liquor?* [This question can be used to establish past year use of alcohol]
- *What is the most you had to drink on any one day you drank beer, wine or liquor during the past 30 days?* [This question can be used to establish past month use of alcohol]

And alcohol use can be implied from questions on the drug use activities answer sheets such as:

- *During the past 12 months, have you gotten any treatment for drinking – such as from a clinic, self-help group, counselor, doctor or other professional?* [From the Treatment Answer Sheet]
- *During the past 12 months, for which drugs have you consciously tried to cut down on your use?* [From the Drugs Answer Sheet]
- *In the past 12 months, I felt aggressive or cross while drinking? (Y/N)* [From the Drinking Experiences Answer Sheet]

In almost all cases where there is disagreement between the recency response and the responses to the ''other'' questions, the NHSDA editing rules dictate whether the respondent's final status should be changed to the most recent indication of use. If a response to some other question indicates use in a later recency period then generally the response to the other question is deleted and a ''Bad Data'' indicator response is put in its place. Because of this editing phase, a person's most recent use of any drug is determined by looking at all related questions and selecting the response for the most recent use. Unless otherwise noted, drug use estimates produced from the NHSDA are created using these ''most recent indication of use,'' edited responses.

By the nature of this editing process, there is the potential for overcorrecting for the negative bias in recency estimates and actually overestimating drug use prevalence for some subgroups. Work is currently underway in 1994 NHSDA to re-evaluate the effects of the editing procedures. In addition, comparisons of the Hui-Walter estimates of prevalence, which are adjusted for both false negative and false positive responses, with the usual NHSDA estimates will provide important information regarding the net biases in the NHSDA estimates.

4.2. *Results of the Hui-Walter estimation*

This analysis of the 1991–1993 NHSDA data is confined to three drugs – alcohol, marijuana, and cocaine. For each analysis, y_1 and y_2 are defined as follows:

$$y_1 = \begin{cases} 1 & \text{if lifetime use was indicated by the recency question response} \\ 0 & \text{if otherwise} \end{cases}$$

and

$$y_2 = \begin{cases} 1 & \text{if lifetime use was indicated by a response to any other question} \\ 0 & \text{if otherwise.} \end{cases}$$

As required by the Hui-Walter procedure, two domains were defined for estimation: smokers and nonsmokers. This partitioning of the population seems to satisfy the dual criteria that: (a) the difference between the drug prevalence rates for the two domains is large – drug use among smokers tends to be considerably greater than among nonsmokers, and (b) the assumption of equality of misclassification probabilities between the two groups is tenable.

Since y_1 and y_2 are collected in the same interview, the Hui-Walter independence model would not seem appropriate since respondents who intentionally falsify their responses to the recency question would likely consistently falsify their reports throughout the questionnaire. However, because subsequent questions regarding lifetime use are less direct than the recency question, it is possible that some lifetime users who falsify on the recency question may unintentionally indicate lifetime use. Then, too, some recency question falsifiers may find the less direct questions on drug use less intimidating and may respond truthfully. There is also the potential that some lifetime users who responded ''no lifetime use'' in the recency question due to forgetfulness may remember later in the interview and then indicate some use.

Even accepting that some inconsistencies in the responses y_1 and y_2 are likely, the assumption that these inconsistencies satisfy the independence assumption is still questionable. Therefore, these data have been analyzed using both the Hui-Walter independence model and the dependent model assumptions and both set of results will be reported subsequently.

This development of the Hui-Walter methodology for self-reported drug use is still very much in its preliminary stages. In the analyses presented here, the main objective is to investigate some capabilities and limitations of the methodology and demonstrate its use for surveys such as the NHSDA where repeated measures are available. For this objective, the usefulness of the methodology for estimating measurement bias is critically evaluated and additional applications in the field of drug use measurement are suggested.

It is possible that while the Hui-Walter false positive and negative rates are biased, their relative magnitudes still provide important insights regarding the causes and remedies of measurement error by identifying the socioeconomic subpopulations, data collection procedures, and survey designs that are most prone to measurement error.

Over 88,000 interviews were collected in the 1991–1993 NHSDA surveys and these data were the object of these analyses. The Hui-Walter method was applied to cell proportions formed by cross-classification of y_1 and y_2 and by smokers and nonsmokers. These cell proportions were weighted for the unequal probabilities of selection; however, the standard errors of the estimates were estimated assuming a simple random sampling design. Thus, while the multistage sample design is accounted for in the classification error estimates, it is not reflected in the standard errors of the estimates and caution must be exercised in applying significance tests to these results.

Table 1 gives the results of the analysis for alcohol, marijuana, and cocaine. In this table, we report estimates of the false positive probability associated with the recency question, ϕ_1, the corresponding false negative probability, θ_1, and the overall prevalence rate for smokers and nonsmokers combined, π. First note that in this table, the false negative rates for the dependent model are generally larger than those for the independent model. This is expected as Vacek (1985) has shown that positive between trial correlations result in a downward bias in the estimated error rates under the independence model. Recall that for the dependent model, only between trial independence for the false positive errors was assumed. Further the dependent model provides only one parameter, ϕ, for the false positive rate. The result is a rate that is an average of $\phi_1$and ϕ_2. Since in the independent model, ϕ_2 is usually much smaller than ϕ_1, the result that ϕ for the dependent model is usually less than ϕ_1 for the independent model is also expected.

Secondly, note the pattern exhibited by the prevalence estimate, viz. in almost all cases

$$\hat{\pi}_{RECENCY} \leqslant \hat{\pi}_{INDEP} \leqslant \hat{\pi}_{DEP} \leqslant \hat{\pi}_{NHSDA}.$$

As anticipated, the estimate of π from the recency question appears to be biased downward, the bias being largest when the false negative rate is largest. Since estimates of θ for the dependent model are usually larger than for the independent model, that $\hat{\pi}_{INDEP}$ is usually less that $\hat{\pi}_{DEP}$ is also anticipated. Note also that since the NHSDA estimator does not take into account the possibility of false positive errors, it is not surprising that $\hat{\pi}_{DEP} \leqslant \hat{\pi}_{NHSDA}$. Finally, it is possible that $\hat{\pi}_{DEP} > \hat{\pi}_{NHSDA}$. Let y_i denote the final edited classification for respondent i. Recall that the NHSDA estimator assigns $y_i = 1$ to any individual i for whom either y_l or y_2 is 1. Further, if both y_l and y_2 are 0, the NHSDA estimator assigns $y_i = 0$ to the respondent. However, the Hui-Walter estimator estimates the proportion of respondents in the population who are truly 1's though both y_l and y_2 are 0. Thus, when these respondents are added to the number of 1 responses, it is possible for the Hui-Walter estimator to produce estimates that are larger than the NHSDA estimates, as can be observed from Table 1.

Finally, the validity of the Hui-Walter estimates is considered; i.e., the degree to which the Hui-Walter estimates of measurement bias are themselves biased. Unfortunately, the evaluation of the bias in the estimators of the error probabilities and π requires knowledge of the true error probabilities that is not available. Sinclair (1993) and Sinclair and Gastwirth (1993) examine the sensitivity of the estimates to violations in the model

Table 1. Comparison of independent and dependent Hui-Walter estimates for the 1991–1993 NHSDA

Characteristic	False negative rate				False positive rate				Estimate prevalence rate (as a percent)			
	Independent		Dependent		Independent		Dependent		$\hat{\pi}_{INDEP}$	$\hat{\pi}_{DEP}$	$\hat{\pi}_{RECENCY}$	$\hat{\pi}_{NHSDA}$
	%	S.E.	%	S.E.	%	S.E.	%	S.E.				
Lifetime alcohol use												
Total	1.343	(0.106)	1.780	(0.053)	0.002	(0.001)	0.083	(0.279)	83.94	84.31	82.82	84.33
Race/ethnicity												
Hispanic	1.941	(0.294)	3.149	(0.142)	0.193	(0.082)	0.212	(0.089)	76.28	77.23	74.85	77.33
Black	1.971	(0.303)	3.228	(0.157)	0.000	(0.000)	0.764	(0.151)	75.74	76.56	74.25	76.91
White/other	1.207	(0.134)	1.450	(0.063)	0.000	(0.000)	0.000	(0.000)	86.12	86.33	85.08	86.33
Age group												
12–17	0.881	(0.427)	5.671	(0.278)	0.328	(0.085)	0.413	(0.069)	39.97	41.99	39.82	42.47
18–25	0.703	(0.164)	1.482	(0.092)	0.000	(0.000)	0.000	(0.000)	87.50	88.19	86.88	88.19
26–34	1.131	(0.173)	1.114	(0.077)	0.000	(0.000)	0.606	(0.238)	92.72	92.66	91.67	92.75
35+	1.725	(0.101)	1.725	(0.102)	0.000	(0.000)	0.000	(0.000)	88.09	88.09	86.57	88.09
Sex												
Male	1.390	(0.146)	1.620	(0.072)	0.353	(0.109)	0.368	(0.112)	88.26	88.46	87.07	88.55
Female	1.288	(0.159)	1.924	(0.072)	0.000	(0.000)	0.000	(0.000)	79.95	80.46	78.92	80.46
Lifetime marijuana use												
Total	0.594	(0.229)	3.384	(0.108)	0.011	(0.006)	0.014	(0.008)	33.21	34.17	33.02	34.18
Race/ethnicity												
Hispanic	1.538	(0.841)	5.439	(0.000)	0.003	(0.018)	0.003	(0.021)	27.11	28.23	26.70	28.23
Black	0.997	(0.638)	3.769	(0.261)	0.082	(0.032)	0.099	(0.037)	31.68	32.60	31.42	32.73
White/other	0.480	(0.334)	3.101	(0.142)	0.001	(0.312)	0.003	(0.012)	34.60	35.53	34.43	35.54
Age group												
12–17	0.523	(0.912)	4.711	(0.468)	0.003	(0.013)	0.006	(0.024)	10.68	11.15	10.62	11.16
18–25	0.940	(0.235)	*	*	0.016	(0.024)	0.023	(0.036)	48.72	55.41	48.27	49.10
26–34	0.355	(0.293)	1.175	(0.098)	0.037	(0.032)	0.043	(0.036)	60.23	60.73	60.03	60.76
35+	*	*	0.000	(0.000)	0.007	(0.017)	1.007	(0.060)	25.19	25.16	25.17	26.66
Sex												
Male	1.208	(0.320)	3.338	(0.151)	0.016	(0.013)	0.019	(0.015)	38.59	39.44	38.13	39.46
Female	0.000	(0.000)	0.000	(0.000)	0.007	(0.027)	0.720	(0.033)	28.36	28.36	28.37	29.39

Lifetime cocaine use												
Total	3.652	(0.471)	5.314	(0.230)	0.005	(0.004)	0.006	(0.004)	11.62	11.83	11.20	11.84
Race/ethnicity												
Hispanic	1.631	(1.699)	6.915	(0.553)	0.000	(0.000)	0.000	(0.000)	10.17	10.75	10.01	10.75
Black	6.027	(1.177)	8.983	(0.672)	0.003	(0.201)	0.005	(0.076)	9.67	9.98	9.09	9.99
White/other	3.471	(0.634)	4.717	(0.000)	0.005	(0.007)	0.005	(0.007)	12.29	12.45	11.87	12.46
Age group												
12–17	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.190	(0.023)	1.36	1.36	1.36	1.74
18–25	3.763	(0.671)	4.906	(0.000)	0.003	(0.015)	0.004	(0.017)	15.64	15.83	15.05	15.83
26–34	0.713	(0.723)	1.935	(0.196)	0.032	(0.019)	0.036	(0.021)	26.43	26.76	26.27	26.81
35+	6.182	(1.766)	8.299	(0.000)	0.000	(0.000)	0.000	(0.000)	8.03	8.21	7.53	8.21
Sex												
Male	4.904	(0.672)	5.271	(0.000)	0.011	(0.009)	0.011	(0.010)	14.75	14.81	14.04	14.83
Female	0.000	(0.000)	0.000	(0.000)	0.000	(0.000)	0.271	(0.018)	8.63	8.63	8.63	9.12

*Indicates estimate not available.

assumptions. For the independent model, they found that the estimates are highly sensitive to violations of the independence assumptions. Moderately large positive correlations between errors in the two measurements can lead to substantial negative biases in the estimates of the error probabilities. Similarly, violations of the between domain homogeneity assumption can also bias the Hui-Walter estimates; however, differences in the error rates as high as 20% between the two domains did not appear to bias the estimates of π appreciably. Since the dependent error model assumes homogeneity between domains but does not assume independence for the false negative errors, the results of Sinclair and Gastwirth support the claim that the dependent model estimates have greater validity than the independent model estimates.

Another indicator of the validity of the estimates is the degree to which the patterns of errors across demographic variables and the magnitudes of the estimated error rates agree with those in the published literature. There are many articles in the literature that will attest to the high potential of underreporting for drug use self-reports, particularly among arrestee reports (see, for example, Mieczkowski 1991; GAO 1993). These researchers would tend to support the higher estimates of false negative error observed for the dependent error model rather than the smaller estimates produced by the independent model. However, since the true false negative and false positive error probabilities for the NHSDA are unknown, the existing literature is insufficient for assessing the magnitudes of the biases in the error rates obtained from either the dependent or the independent model.

Besides the question of the bias in the estimates, one can, to some extent, investigate the question of the relative validity of the Hui-Walter estimates; that is, the extent to which the estimates of misclassification error provide information regarding the relative bias in self-reports across socioeconomic classes and geographic regions, and for alternate drugs of abuse. For this analysis, the results from Fendrich and Vaughn (1994), who estimated the denial rates for the National Longitudinal Survey of Youth (NLSY) cohort, were used. For nine socioeconomic variables, they computed the proportion of respondents who admitted to using a drug (marijuana or cocaine) in the 1984 survey and then denied ever using the drug in the 1988 survey.

The NLSY is a nationally representative longitudinal sample of 12,686 individuals who were ages 14 to 21 when they were first interviewed in 1979. Twelve waves of interviews were conducted between 1979 and 1990 for the sample analyzed by Fendrich and Vaughn. Retention rates averaged about 90% in each of the survey years. Questions about illicit substance use were asked in 1980, 1984, and 1988. In 1988, an experiment was conducted in which a half sample of subjects were randomly assigned to an interviewer assisted mode and the other half to self-administered mode.

The focus of Fendrich and Vaughn's study is on responses to the surveys administered in 1984 and 1988 since these two surveys included nearly identical questions about lifetime use for two illicit drugs – cocaine and marijuana. Their study considers two subsamples as follows: (a) all respondents who completed the questions about marijuana use in 1984 and 1988 and also reported lifetime use of marijuana in 1984 ($n = 6{,}204$) and (b) all respondents who completed the questions about cocaine use in 1984 and 1988 and also reported lifetime use of cocaine in 1984 ($n = 1{,}589$).

Although denial rates estimated by Fendrich and Vaughn provide direct evidence of false negative error in the NLSY, they should not be taken as estimates of the false negative probabilities since they refer only to respondents who reported any use of a drug in the first interview. Thus, the rates exclude persons who used the drug but did not report their use and respondents who never used the drug but reported that they did in the first interview. Further, the magnitudes of the Fendrich and Vaughn denial rates are not useful for predicting the magnitudes of the NHSDA false negative error rates for a number of reasons. First, they are denial rates, not false negative rates. Second, the interview setting and mode in the NLSY are quite different from the NHSDA. While the NLSY is a panel study in which the interviewer returns annually to reinterview the respondents and may become quite familiar with them, the NHSDA is a one-time cross-sectional survey in which the interviewer and respondent have never met before. In the NHSDA, great care is taken to preserve the anonymity of the respondents and to protect their responses from discovery by the interviewer. In the NLSY, this type of confidentiality is not possible due to the nature of the survey. Finally, in the NLSY, the two measurements were separated by a period of four years, while in the NHSDA the two measurements were separated by only a few minutes. Thus, in the NLSY, there is a greater chance that the respondent's response on measurement 1 will change by the time measurement 2 is taken.

Despite these limitations of comparisons between the NHSDA and the NLSY estimates, such comparisons may still be quite fruitful. First, to the extent that the denial rates estimated in the Fendrich and Vaughn study reflect general tendencies of various socioeconomic domains to underreport their drug use and second, to the extent that these tendencies and patterns for underreporting are stable over time, the estimates of false negative rates from the NHSDA study should be correlated, to some extent, with the denial rates from the NLSY for the same subpopulations. Lack of concordance between the two sets of estimates may not be evidence of the invalidity of either set of estimates for the reasons cited above. However, significant correlations between the two estimates are evidence of the validity of both sets of estimates as measures of the ''relative'' true false negative error in self-reported drug use in surveys.

In Table 2, Fendrich and Vaughn's NLSY denial rates, the NHSDA independent model false negative error estimates (NHSDA-IND), and the NHSDA dependent model false negative error estimates (NHSDA-DEP) are given. Note first that the NLSY denial rates are considerably larger than both sets of NHSDA estimates. However, what is important here is the correlation between the NLSY and the NHSDA estimates. Table 3 displays these correlations for all pairs of the three sets of estimates for marijuana and cocaine. The ''across variables'' correlation is Corr(NLSY, NHSDA) across all 29 variable categories shown in Table 2. The NHSDA-INDEP estimates exhibited highly significant correlation with the NLSY denial rates for both marijuana (.76) and cocaine (.58). Surprisingly, the ''across variables'' correlations for the NHSDA-DEP estimates are not significant. The ''within variables'' correlation is the average correlation between categories within each of the nine variables in Table 2. Here, both the NHSDA-INDEP and the NHSDA-DEP estimates exhibit highly significant correlations with the NLSY estimates for cocaine while for marijuana, the correlations are not distinguishable from 0. These results support the validity of the Hui-Walter estimates, when viewed as measures of relative bias (between socioeconomic domains).

Table 2. Comparison of NLSY denial rates and 1991–1993 NHSDA false negative rates. Lifetime use for 23–32 year olds

Characteristic	Marijuana (%)			Cocaine (%)		
	NLSY	NHSDA-IND	NHSDA-DEP	NLSY	NHSDA-IND	NHSDA-DEP
Total Total 23–32 Yr. Olds	11.7	0.77	1.38	18.9	1.72	48.07
Privacy						
Private interview	12.5	0.93	0.38	18.6	1.50	2.17
Others present	10.3	0.55	0.39	22.1	1.93	2.36
Race/ethnicity						
Hispanic	14.9	2.58	3.40	20.8	0.85	3.21
Black	19.3	2.66	2.87	33.2	3.99	6.79
White/other	8.0	0.38	0.99	15.0	1.55	1.66
Sex						
Male	11.3	0.90	*	19.4	1.79	*
Female	12.2	0.62	1.40	18.3	0.00	0.00
Income						
0–11,999	15.0	0.65	2.57	19.7	0.35	2.85
12,000–19,999	11.1	0.57	0.92	20.3	2.46	3.10
20,000–29,999	10.6	1.38	1.49	16.5	0.55	2.45
30,000–42,999	10.2	0.49	1.53	22.4	0.74	1.48
43,000+	9.0	0.16	0.61	22.6	0.11	2.00
Education						
<High school	15.4	1.56	2.09	26.6	1.99	3.18
High school	11.6	0.29	1.69	18.9	0.56	2.92
Some college	11.3	0.56	0.88	18.7	0.16	1.14
College graduate	8.3	0.00	0.00	12.8	0.19	0.00
Labor force						
Employed	11.2	0.68	1.40	18.3	0.22	1.86
Unemployed	12.3	1.26	1.26	22.8	3.74	3.74
Not in labor force	14.6	0.77	1.35	19.7	1.44	3.39

Marital status						
Single	11.7	0.89	1.66	17.5	2.21	2.39
Married	11.8	0.78	1.28	22.1	1.58	2.45
Widowed/div/sep	11.6	0.46	1.06	14.3	0.22	1.22
Residency						
Urban	11.7	0.86	10.56	17.9	1.78	2.46
Rural	11.7	0.57	0.90	25.0	1.27	1.27
Age						
23–25	11.7	1.22	1.81	21.4	3.09	3.09
26–27	12.4	0.49	0.63	19.7	0.92	2.36
28–29	11.4	0.72	1.16	17.6	1.40	1.94
30–32	11.3	0.52	1.62	16.6	0.00	0.00

*Indicates estimate not available.

Table 3. Correlational analysis of NHSDA false negative rates and NLSY denial rates for characteristics in Fendrich and Vaughn (1994)

Correlation	Marijuana		Cocaine	
	Across Var. ($n=29$)	Within Var. ($n=9$)	Across Var. ($n=29$)	Within Var. ($n=9$)
NHSDA-IND with NLSY	0.76***	0.28	0.58***	0.57***
NHSDA-DEP with NLSY	0.06	0.01	0.02	0.55**
NHSDA-IND with NHSDA-DEP	0.15	0.41*	0.19	0.87***

*Significant at $\alpha = 0.05$ **Significant at $\alpha = 0.01$ ***Significant at $\alpha = 0.001$.

5. Summary and Conclusions

In this article, a general model for studying misclassification in self-reported drug use was presented and the model was then extended to the case where two measurements of the same characteristic are available for a sample of respondents. For the two measurements case, the general model requires seven parameters while only three degrees of freedom are available for estimation. Thus, some additional assumptions are required to reduce the set of unknown parameters to three or less. It was shown how the assumptions typically made for test-retest, true value, improved value, and Hui and Walter methods relate to the general model. Further, it was shown how the measures of reliability, measurement bias, estimator bias, mean squared error, false negative and false positive probability can be defined in the context of the general model and how they may be estimated under the appropriate study designs.

Finally, the use of the Hui and Walter method for estimating misclassification error based upon two erroneous reports was demonstrated. The reports may be self-reports, biological tests, administrative record values, or any other measure. For the general case of two measurements, the Hui-Walter method using maximum likelihood estimation to obtain estimates of the false negative and false positive probabilities associated with each measurement as well as the error adjusted estimates of prevalence based upon both measurements. The method requires that the population be divided into two domains that have markedly different prevalence rates and that satisfy the assumption of homogeneity of error probabilities.

To demonstrate the use of the Hui-Walter method for evaluating the error in self-reported drug use, the method was applied to the 1991–1993 NHSDA data. Two sets of model assumptions were evaluated: the independent model and the dependent model. The dependent model yielded estimates of false negative error that were generally larger than those for the independent model. Further, the dependent model produced estimates of drug use prevalence that were very nearly the same as the NHSDA published estimates. However, an important advantage of the Hui-Walter method is that it has a probability basis for the estimation that is lacking in the NHSDA estimation procedure. In addition, the Hui-Walter estimators are adjusted for false positive errors and consistent false negative errors while the NHSDA estimator ignores these errors.

To provide evidence of the validity of the Hui-Walter estimates, correlations between the NHSDA model-based estimates of false negative error and the NLSY denial rates were computed. The independent model exhibited highly significant average correlations across categories within the nine socioeconomic variables reported in Fendrich and Vaughn (1994). For cocaine, both models produced estimates that were significantly correlated with the NLSY within variables. This evidence suggests that the Hui-Walter method is at least useful for comparing false negative rates across socioeconomic subgroups within the same survey in order to identify which groups are most prone to false negative error. The available data were inadequate to determine whether the false positive and false negative error rates produced by Hui-Walter are unbiased for this application.

Future work in this area will include further study of the bias and validity of the Hui-Walter estimation method. As an example, in this application, the joint likelihood of smokers and nonsmokers was considered since this partitioning of the population seemed to fit the Hui-Walter criterion well. Other definitions for the two domains that also *a priori* seem to meet the Hui-Walter criterion will also be considered and the estimates produced by each definition will be compared. Finally, attempts will be made to relate the estimates as dependent variables to subpopulation characteristics using logistic models that predict the false negative rate from variables such as age, race, sex, income, etc. In this way, the concurrent validity and predictive validity of the Hui-Walter estimates can be investigated.

Finally, the Hui-Walter method should be considered for studies of drug use reporting error that use a biological test such as a hair, urine, or nail test to evaluate the error in the self-report. As reported in the literature (see, for example, Cone 1994), biological tests are themselves subject to considerable error, even when the period for drug use is restricted to maximize the accuracy of the test results. Self-report validity studies employing biological testing have assumed the true value or preferred value assumptions described in Section 3.2. However, the general two measurement model of Section 3.3 may be more appropriate for these studies. As mentioned in Section 3.3, when the second measurement is a biological test, the assumption of between measurement independence is likely satisfied and thus the Hui-Walter independence model can be used. Under this model, the procedure will provide estimates of false positive and false negative errors for both the self-report and the biological test result. In this way, the accuracies of both self-reports and biological tests for drug use measurement can be studied.

6. References

Bailar, B.A. (1968). Recent Research in Reinterview Procedures. Journal of the American Statistical Association, 63, 41–63.

Biemer, P.P. (1988). Measuring Data Quality. In: Telephone Survey Methodology, eds. R.M. Groves, et al. New York: John Wiley & Sons.

Biemer, P.P. and Stokes, S.L. (1991). Approaches to the Modeling of Measurement Errors in Surveys. In: Measurement Errors in Surveys, eds. P. Biemer, et al. New York: John Wiley & Sons, 487–516.

Bohrnstedt, G.W. (1983). Measurement. In: Handbook of Survey Research, eds. P.H. Rossi, et al. New York: Academic Press, 70–122.

Bross, I. (1954). Misclassification in 2×2 Tables. Biometrics, 10, 488–495.

Cochran, W.G. (1968). Errors of Measurement in Statistics. Technometrics, 10, 637–666.

Cone, E. (1994). New Developments in Biological Measures of Drug Prevalence. Paper presented at the National Institute on Drug Abuse (NIDA) Technical Review: The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates, Gaithersburg, MD.

Fendrich, M. and Vaughn, C. (1994). Diminished Lifetime Substance Use over Time: An Inquiry into Differential Underreporting. Public Opinion Quarterly, 58, 96–124.

Groves, R.M. (1989). Survey Errors and Survey Costs. New York: John Wiley & Sons.

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

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. Anderson Publishing Co. and the ACJS, Cincinnati, OH, 275–302.

Lessler, J. and O'Reilly, J. (in press). Mode of Interview and Reporting of Sensitive Issues: Design and Implementation of AUDIO Computer-Assisted Self-Interviewing. Proceedings of the 1994 NIDA Technical Review: The Validity of Self-Reported Drug Use – Improving the Accuracy of Survey Estimates, September 8–9, Gaithersburg, MD.

O'Muircheartaigh, C. (1991). Simple Response Variance: Estimation and Determinants. In: Measurement Errors in Surveys, eds. P. Biemer, et al. New York: John Wiley & Sons, 551–574.

Sinclair, M.D. (1993). Evaluating Reinterview Survey Methods for Measuring Response Errors. Unpublished doctoral dissertation, Department of Statistics, George Washington University, Washington, D.C., September 30, 1994.

Sinclair, M.D. and Gastwirth, J.L. (1993). Evaluating Reinterview Survey Methods for Measuring Response Errors. Proceedings of the Annual Research Conference of the U.S. Bureau of the Census, Washington, D.C., 771–738.

Tourangeau, R., Jobe, J., Pratt, W., and Rasinski, K., (in press). Mode Effects in Survey Results on Sensitive Topics. Proceedings of the 1994 NIDA Technical Review: The Validity of Self-Reported Drug Use – Improving the Accuracy of Survey Estimates, September 8–9, Gaithersburg, MD.

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

U.S. General Accounting Office (GAO) (1993). Drug Use Measurement: Strengths, Limitations, and Recommendations for Improvement. Report to the Chairman, Committee on Government Operations, House of Representatives, GAO/PEMD-93-18, June.

Vacek, P.M. (1985). The Effect of Conditional Dependence on the Evaluation of Diagnostic Tests. Biometrics, 41, 959–968.

Received October 1994
Revised January 1996