Journal of Official Statistics, Vol. 21, No. 2, 2005, pp. 287–308

Model-based Estimation of Drug Use Prevalence Using Item Count Data

Paul Biemer[1] *and Gordon Brown*[2]

The item count (IC) method for estimating the prevalence of sensitive behaviors was applied to the National Survey on Drug Use and Health (NSDUH) to estimate the prevalence of past year cocaine use. Despite considerable effort and research to refine and adapt the IC method to this survey, the method failed to produce estimates that were any larger than the estimates based on self-reports. Further analysis indicated the problem to be measurement error in the IC responses. To address the problem, a new model-based estimator was proposed to correct the IC estimates for measurement error and produce less biased prevalence estimates. The model combines the IC data, replicated measurements of the IC items, and responses to the cocaine use question to obtain estimates of the classification error in the observed data. The data were treated as fallible indicators of (latent) true values and traditional latent class analysis assumptions were made to obtain an identifiable model. The resulting estimates of the cocaine use prevalence were approximately 43 percent larger than the self-report only estimates and the estimated underreporting rates were consistent with those estimated from other studies of drug use underreporting.

Key words: Latent class analysis; NSDUH; randomized response; surveying sensitive topics; cocaine use.

1. Introduction

It is well-accepted that survey estimates of the prevalence of illicit drug use tend to be negatively biased (see, for example, Wright, Gfroerer, and Epstein 1997; Turner et al. 1992; Mieczkowski 1991). This is primarily due to the social desirability bias associated with the reporting of stigmatized behaviors (Fisher 1993; Turner et al. 1992). Private modes collecting self-reported drug use data such as Audio Computer Assisted Self Interviewing (ACASI) have been shown to reduce the bias (O'Reilly et al. 1994). These findings led the U.S. Substance Abuse and Mental Health Services Administration (SAMHSA) in 1999 to adopt the ACASI methodology for the National Survey of Drug Use and Health (NSDUH). Although the use of ACASI increased the level of honest reporting of illicit drug use and other sensitive behaviors, there is still concern that considerable underreporting of drug use persists in the NSDUH since no estimates of drug use prevalence changed appreciably with the implementation of ACASI (Chromy, Davis, Packer, and Gfroerer 2002).

[1] RTI International, P O Box 12194, Research Triangle Park, NC 27709-2194, and University of North Carolina at Chapel Hill, NC, U.S.A. Email: ppb@rti.org
[2] RTI International, P O Box 12194, Research Triangle Park, NC 27709-2194, U.S.A. Email: ggb@rti.org

Acknowledgments: The authors thank Douglas Wright at SAMHSA for his comments on this work and the provision of a number of helpful references. We also thank Randall Bender and Sharon Christ for their helpful comments. Finally, we are sincerely grateful to SAMHSA for their support of this research.

Therefore SAMHSA has continued to investigate survey methodologies that may improve the accuracy of the prevalence estimates produced from the NSDUH. A relatively new methodology that shows some promise for improving reporting accuracy is the “item count method.” This technique is designed to give respondents an enhanced perception of anonymity when reporting a sensitive behavior such as drug use. This is accomplished by including the sensitive behavior of interest in a list of other, relatively nonstigmatizing behaviors. The respondent reports the number of items in the list in which he or she has engaged. Only the number of behaviors is reported, not which specific behaviors apply. Since the report does not allow anyone accessing the data to know which specific behaviors are true for a respondent, there is no way to determine whether a respondent has admitted to the sensitive behavior. If the average number of nonsensitive behaviors can be estimated for the population, the prevalence of the sensitive behavior in the population can be estimated by the difference between the average number of behaviors reported for the population including and excluding the stigmatized behavior.

A large-scale test of the efficacy of the item count (IC) methodology for estimating cocaine use prevalence was implemented in the 2001 NSDUH and is described in Biemer et al. (2005). Extensive research was conducted to design clear and nonthreatening item count questions and administration procedures; nevertheless, the results of the 2001 study were disappointing. The accuracy of the IC-based prevalence rates was in fact somewhat worse than the accuracy of estimates of cocaine use obtained by directly asking respondents about their drug use. Although previous applications of the item count methodology to drug use prevalence estimation (see, for example Droitcour et al. 1991) produced similarly unsatisfactory results, the NSDUH study showed promise due to its large sample size (nearly 70,000 interviews) and careful attention to the design of the item count questions. Nevertheless, test–retest data collected in the 2001 NSDUH experiment confirmed that a major source of the bias in the IC estimates was the poor reliability of the IC questions.

In this article, we propose a new estimator of drug use prevalence that incorporates both the item count data and the responses to the direct question about drug use in which both types of responses are adjusted for the classification error. The new estimator is based upon a latent structure model in which the latent variables represent the true values associated with the item count question and the cocaine use question; indicators of these latent variables are the observed responses from the corresponding survey questions. In addition, classification error is estimated simultaneously using the same model by incorporating test-retest data for the IC questions. The new estimator may be conceptualized as the usual item count estimator where the counts have been corrected for measurement error. Under certain specified assumptions, the new estimator has smaller measurement bias than the estimators from either the direct questioning or the item count method alone.

The next section provides more details on the design of the NSDUH IC experiment and compares the usual IC and direct question prevalence estimates to illustrate the IC bias. Section 3 provides the details of the latent class model proposed for correcting the IC estimates for the bias and applies this model to the IC experimental data. Section 4 provides our conclusions from the study.

2. The NSDUH Item Count Design

The National Survey on Drug Use and Health (NSDUH) has been conducted since 1971 and serves as the primary source of information on the prevalence and incidence of illicit drug, alcohol, and tobacco use in the civilian, noninstitutionalized population aged twelve or older in the United States. Information about substance abuse and dependence, mental health problems, and receipt of substance abuse and mental health treatment is also included. Before 2002, the name of the survey was the National Household Survey on Drug Abuse (NHSDA).

2.1. *The item count estimator*

For the basic IC approach, a sample of $2n$ households is split completely at random into two subsamples of size n. One subsample receives the *short* IC question which asks respondents to indicate how many behaviors from the list of k behaviors apply to them. The other subsample receives the *long* IC question consisting of the same k items as in the short IC question plus the sensitive item of interest for a total of $k + 1$ behaviors. The IC estimator of the sensitive item's prevalence is

$$\bar{x}_{\text{diff}} = \bar{x}_L - \bar{x}_S \tag{1}$$

where $\bar{x}_L$ is the mean response to the long IC question and $\bar{x}_S$ is the mean response to the short IC question. The variance of this estimator is

$$Var(\bar{x}_{\text{diff}}) = Var(\bar{x}_L) + Var(\bar{x}_S) \tag{2}$$

where $Var(\bar{x}_L)$ is the variance of $\bar{x}_L$ and $Var(\bar{x}_S)$ is the variance of $\bar{x}_S$.

Rather than using one pair of IC questions (corresponding to the short and long lists), the precision of the IC estimator can be substantially improved using two pairs of IC questions where the two short IC questions consist of mutually exclusive lists of behaviors. With this approach, one subsample receives the first short IC question, denoted by ICQ1(S), and the other subsample receives the second short IC question, denoted by ICQ2(S). In addition, the first subsample receives the second long IC question, denoted by ICQ2(L), and the second subsample receives ICQ1(L), corresponding to ICQ1(S). Table 1 summarizes this design. In this way, two IC count estimates can be computed from the same sample – one for the pair ICQ1(S)/ICQ1(L) (referred to as Pair 1) and one from the pair ICQ2(S)/ICQ2(L) (referred to as Pair 2). The average of the two IC estimators is much more precise than either single pair estimator.

Table 1. Design of the NSDUH IC experiment

	Sample 1	Sample 2
Short-list question	ICQ1(S)	ICQ2(S)
Long-list question	ICQ2(L)	ICQ1(L)

Then the estimator of cocaine use for Pair 1 (shaded cells) can be written as

$$\hat{p}_1 = \bar{x}_{L(1)} - \bar{x}_{S(1)} \tag{3}$$

and that for Pair 2 (unshaded cells) as

$$\hat{p}_2 = \bar{x}_{L(2)} - \bar{x}_{S(2)} \tag{4}$$

The two estimators are averaged to produce the IC estimate for the entire sample as

$$\hat{p} = (\hat{p}_1 + \hat{p}_2)/2 \tag{5}$$

The variance of this estimator is

$$Var(\hat{p}) = 0.25[Var(\hat{p}_1) + Var(\hat{p}_2) + 2\rho_{12}\sqrt{Var(\hat{p}_1)Var(\hat{p}_2)}] \tag{6}$$

where ρ_{12} is the correlation between the estimators $\hat{p}_1$ and $\hat{p}_2$. Noting that $Cov(\bar{x}_{L1}, \bar{x}_{L2}) = Cov(\bar{x}_{S1}, \bar{x}_{S2}) = 0$ by design, it follows that $\rho_{12} = -c_1\rho_1 - c_2\rho_2$, where c_1 and c_2 are nonnegative constants, ρ_1 is the correlation between $\bar{x}_{L1}$ and $\bar{x}_{S2}$, and ρ_2 is the correlation between $\bar{x}_{L2}$ and $\bar{x}_{S1}$. Assuming ρ_1 and ρ_2 are nonnegative, which is expected in almost all practical applications, ρ_{12} will be nonpositive. Thus, the expression for $Var(\hat{p})$ of $0.25[Var(\hat{p}_1) + Var(\hat{p}_2)]$ will likely overestimate the true variance.

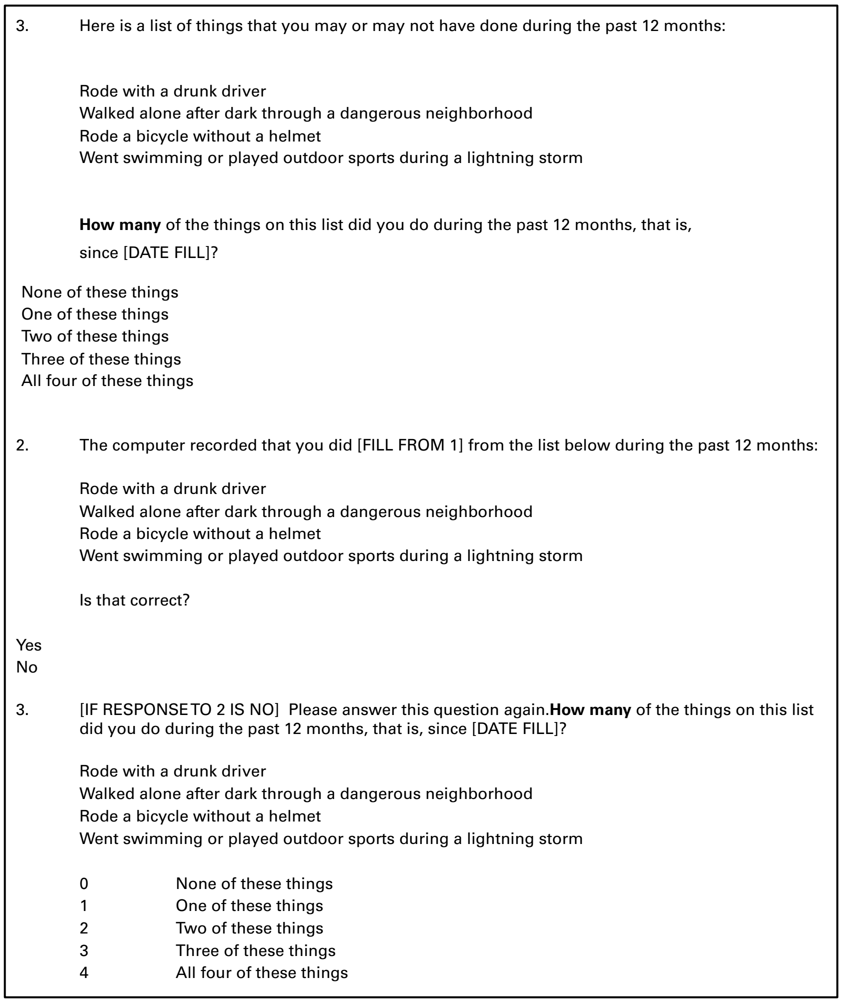
2.2. *Preliminary research to optimize the NSDUH item count design*

Considerable research and pretesting was conducted prior to the implementation of the IC procedures in order to adapt the approach to the specific content and data collection protocols of the NSDUH. For example, it was determined that the number of behaviors included in the short ICQ list is a key determinant of response accuracy. A list that is too short (say only one or two items not including the sensitive item) will cause respondents to fear that their privacy is not sufficiently protected since there are not enough innocuous behaviors to adequately "mask" counts that include the sensitive behavior. Deliberate exclusion of the sensitive behavior from the count could result. Conversely, a list that is too long (say, six or seven innocuous behaviors) substantially increases the difficulty of the task since respondents now have to think about more behaviors, determine their applicability, and keep a count for a longer list.

Cognitive laboratory experimentation suggested that the number of items in the short ICQ should not exceed five. Simulation studies designed to assess the precision of the IC estimates with lists of varying lengths suggested that four short ICQ items provided adequate precision for the NSDUH application. Therefore, the ideal number of behaviors for the short ICQ was determined to be four.

Preliminary testing also revealed that respondents were less suspicious of the IC tasks when the IC behaviors were consistent with the NSDUH's content, instrumentation, and target population. Cognitive research was conducted to determine which behaviors NSDUH respondents would find the least threatening in the context of a cocaine use question. Behaviors that are slightly counter to social norms were seen as more consistent with cocaine use behaviors and less likely to arouse suspicion than those that were perceived as completely innocuous relative to the use of cocaine. Reviews of the literature and multiple rounds of cognitive testing and revision produced the final lists of items shown in Figures 1 and 2.

These modules were presented to respondents during the ACASI portion of the NSDUH interview. The ACASI software queried respondents directly after they entered their



3. Here is a list of things that you may or may not have done during the past 12 months:

Rode with a drunk driver
Walked alone after dark through a dangerous neighborhood
Rode a bicycle without a helmet
Went swimming or played outdoor sports during a lightning storm

How many of the things on this list did you do during the past 12 months, that is, since [DATE FILL]?

None of these things
One of these things
Two of these things
Three of these things
All four of these things

2. The computer recorded that you did [FILL FROM 1] from the list below during the past 12 months:

Rode with a drunk driver
Walked alone after dark through a dangerous neighborhood
Rode a bicycle without a helmet
Went swimming or played outdoor sports during a lightning storm

Is that correct?

Yes
No

3. [IF RESPONSETO 2 IS NO] Please answer this question again.**How many** of the things on this list did you do during the past 12 months, that is, since [DATE FILL]?

Rode with a drunk driver
Walked alone after dark through a dangerous neighborhood
Rode a bicycle without a helmet
Went swimming or played outdoor sports during a lightning storm

0	None of these things
1	One of these things
2	Two of these things
3	Three of these things
4	All four of these things

Fig. 1. Item count practice question with verification. This practice series was administered to all respondents regardless of age or sample assignment. A similar series of questions was used for all item count questions

responses to give them an opportunity to change their counts if they entered a number in error. In this manner, all IC responses were verified by the respondents.

Although not a necessary component of the IC approach, respondents were asked directly about each behavior in the IC short list later in the questionnaire. The primary purpose of including a direct question for each short list behavior was to estimate the reliability of the IC questions. Using the responses from the individual item direct questions, a second, "pseudo-IC" response was formed by counting the number of positive responses to the individual questions for each respondent. Then, using test–retest methods, the reliability of the original IC question was estimated. As we will see in Section 3, these test–retest data are also necessary for applying the latent class model adjustments to the IC data.

Biemer et al. (2005) provides a more complete discussion of the NSDUH item count design and the findings from the IC pretests.

Youth Modules

ICQ1(S) Items

Ran away from home and slept on the street
Gained or lost more than 50 pounds
Were seriously injured in a fight
Used steroids to become more muscular

ICQ2(L) Items

Crossed railroad tracks when a train was coming and almost got hit by the train
Used laxatives or vomited on purpose in order to keep your weight down
Hacked into a government computer system
Used cocaine, in any form, one or more times
Was careless and set a large or serious fire with a cigarette or a match

Adult Modules

ICQ1(S) Items

Drove a car more than 100 miles per hour
Gained or lost more than 50 pounds
Were injured when you tried to stop a fight or an assault
Used steroids to become more muscular

ICQ2(L) Items

Crossed railroad tracks when a train was coming and almost got hit by the train
Used laxatives or vomited on purpose in order to keep your weight down
Passed another vehicle when you knew it was not safe to pass
Used cocaine, in any form, one or more times
Was careless and set a large or serious fire with a cigarette or a match

Fig. 2. Youth and adult items used for Random Sample I. These items were used in the item count questions like those shown in Figure 1 for Random Sample I. For Random Sample II, the questions were identical except the cocaine item was deleted from ICQ2(L) and added to ICQ1(S) to form ICQ2(S) and ICQ1(l), respectively

2.3. Item count estimates of cocaine use prevalence

As previously described, a verification question was included in the NSDUH implementation to provide respondents with an opportunity to correct their responses to the IC questions (see Figure 1). Both initial and verified or corrected responses were recorded, which allowed the calculation of IC estimates both before and after verification. If the verification approach was successful at reducing measurement error, the IC estimates based upon corrected data should be more accurate.

Past year cocaine use was estimated from both unverified and verified data according to the two-pair IC estimator in (5). These estimates as well as the estimates based upon the direct cocaine use question only (labeled as NSDUH) are provided in Table 2. All estimates are weighted and the standard errors reflect the weighting as well as complex survey design

Table 2. Item count estimates of past year cocaine use prevalence (in percent) by age and gender before and after verification

Age	Gender	Before verification	After verification	S.E.	NSDUH	S.E.
12–17	Total	0.05	0.73	0.49	1.5	0.10
	Male	− 0.61	0.19	0.75	1.4	0.14
	Female	0.73	1.28	0.63	1.5	0.15
18+	Total	− 0.44	− 0.08	0.39	1.9	0.09
	Male	− 0.29	0.42	0.63	2.8	0.15
	Female	− 0.60	− 0.55	0.45	1.1	0.08

effects. The estimates are provided for two age groups – 12 to 17 and 18 or older. As noted in Figures 1 and 2, the item count questions for the younger age group were slightly modified to be more age-appropriate.

An unexpected finding from Table 2 is that all of the IC estimates of past year cocaine use are smaller than the NSDUH estimates based upon direct questioning. This is disappointing since the IC methodology was designed to reduce underreporting, which should produce estimates that are no smaller than the estimates from direct self-reports. The results suggest that our pretest research efforts to develop item count modules that minimized the nonsampling error usually associated with this methodology did not adequately improve the item count estimates.

The estimates in Table 2 suggest that verification may have been successful at reducing underreporting since the verified estimates are slightly larger than the corresponding unverified estimates. Before verification, four IC estimates are negative, whereas only two estimates are negative after verification.

IC estimates will be biased downward if true cocaine users tend not to include cocaine use in their IC counts. Since cocaine use is a highly stigmatized, illegal behavior, underreporting of it may be a problem for IC questioning just as it is for direct questioning. The fact that the verified data estimates are larger than original data estimates provides some evidence that IC estimates of cocaine use, at least in this application, are negatively biased. The results in the table suggest that the verification questions reduced the negative bias in the IC estimates to some extent, but it is evident from the comparison of these estimates with the NSDUH estimates that a substantial amount of bias still remains.

To further examine the effect of verification on the IC response, the proportion of original responses that were changed and the direction of the change were estimated. Table 3 summarizes the results. Overall, about 1 percent of all responses changed in the verification process. Further, twice as many respondents decreased the originally reported count as increased it. Changes from a count of 0 to 1 or 1 to 0 accounted for more than half of the revisions. The patterns are similar for both IC pairs and both long- and short-form versions.

Table 3. How ICQ responses were revised following verification

Item count question	Percent changed in verification	Percent revised downward	Percent revised upward
ICQ1(S)	0.93	70.9	29.1
ICQ1(L)	0.69	56.5	43.5
ICQ2(S)	1.11	73.8	26.2
ICQ2(L)	0.74	64.5	35.5
Average	0.87	66.4	33.6

2.4. Reliability of the item counts

Using the data from the direct queries of each item count behavior, the reliability of both IC questions can be computed. The usual method for assessing the reliability of a question is through a test–retest design. A form of test–retest data is available from the IC experiment since the items making up the IC short list questions were also asked individually for the same respondents. The responses to the individual items can be used to form a second count of the IC short list items.

Let y_k for $k = 1, \ldots, 5$ denote a response to question k corresponding to item k in one of the IC short list questions, where $y_k = 0$ if the response is "no" and 1 if the response is "yes." Then, if there is no measurement error in either the response to the IC short list question or the corresponding individual question, y_k, then $\sum_{k=1}^{4} y_k$ should be equal to the response to the IC short list question. A departure from equality of the IC and the pseudo-IC counts is evidence of measurement error in either or both counts.

Using these data, the reliability of the IC short list questions count was estimated by Cohen's κ statistic. The values of κ for ICQ1(S) and ICQ2(S) responses are 0.48 and 0.43, respectively, indicating rather poor reliability. This suggests that measurement error is a serious problem with the IC approach and may be an important contributor to the failure of the approach to produce valid estimates of cocaine use. Biemer and Stokes (1991) show that unreliability increases the variance of an estimator by a factor of $1/\kappa$. This suggests that the variance of the item count estimator increased by more than 200% as a result of measurement variance.

To further investigate the test–retest reliability of ICQ1(S) and ICQ2(S), we compared the IC and the pseudo-IC responses (see Table 4). Table 4 indicates a considerable amount of inconsistency between the two responses, as only 84.9 percent of the responses are in agreement (i.e., in the diagonal cells of the table). Among the disagreements (the off-diagonal cells), the pseudo-IC response is higher than the IC response for approximately 75 percent of the cases. Many of the differences are quite extreme. For example, 1,393 persons responded "none" (0) to the IC question but answered "yes" (1) to all four items when the questions were asked individually. This large inconsistency is even more puzzling when compared with the number of persons who responded "yes" to two or three individual items–718 and 263, respectively.

Table 4. Item count response by pseudo-item count response for both short IC questions

Pseudo-IC response	Item count response				
	0	1	2	3	4
0	51,015	1,641	286	49	47
1	4,392	6,333	447	48	19
2	718	607	622	53	9
3	263	114	48	44	7
4	1,393	96	37	9	8

If the primary cause of the inconsistencies is memory error, the number of persons who respond "0" to the IC question and then answer with three positives to the individual questions should be larger than the number of persons who answered with four positives to the individual questions. This suggests that 1,393 individuals in the 0–4 cell of the table may have been confused as to how to respond to the IC question; for example, they may have indicated the number of behaviors that did *not* apply to them rather than the number that did.

Biemer et al. (2005) speculated that the primary cause of the disappointing performance of the IC method was probably task difficulty. Respondents have difficulty in accurately counting up the number of activities on a list that apply to them. Some respondents may become confused by the question and may record a number corresponding to the order of the item in the list for a particular activity instead of "1" if only one item applies. As previously mentioned, some may have misunderstood what count was needed: is it the number of applicable or nonapplicable behaviors? In addition, the process of reading the items in the list, recalling whether they ever engaged in each one over the last twelve months, keeping a running tally of those that apply and then recording the final tally accurately is difficult for some respondents.

3. The Model-based Item Count Estimation

The above results suggest that the failure of the IC methodology to produce estimates of past-year cocaine use that are less biased than the direct cocaine use question is likely due to error in the IC responses. In this section, we consider a latent variable model that simultaneously represents the relationships between short- and long-form item count responses, the pseudo-IC response, and the direct cocaine use response, including parameters for the measurement error associated with all four responses. Our modeling approach treats the four types of response variables as indicators of corresponding latent variables that represent the true values of variables. A latent class model (see, for example Vermunt 1997; Heinen 1996) will be employed to specify the joint likelihood of these data. The parameter estimates will be obtained by maximizing the likelihood subject to constraints that reflect the relationship between the short- and long-form IC counts.

To fix the ideas, consider a single pair of IC questions – say, Pair 1 defined in Table 1. Let $X(=0,\ldots,4)$ denote the unobserved (latent) true ICQ1(S) response and let $Z(=0,\ldots,5)$ denote the unobserved true ICQ1(L) response. Let $A(=0,\ldots,4)$ and $D(=0,\ldots,5)$ denote the observed responses to ICQ1(S) and ICQ1(L) and $B(=0,\ldots,4)$ denote the pseudo-item count variable. Let C denote the response to the direct past year cocaine use Question (1 if "yes," 0 if "no") and let $Y(=1 \text{ or } 0)$ denote the unobserved (latent) true status of past year cocaine use. Thus, C is an indicator of Y.

Denoting the means of A and D by $\bar{A}$ and $\bar{D}$, respectively, the item count estimator of cocaine use prevalence in Equation (3) becomes

$$\hat{p}_{IC} = \bar{D} - \bar{A} \tag{7}$$

The measurement bias in this estimator can be eliminated if $\bar{A}$ and $\bar{D}$ were replaced by $\bar{X}$ and $\bar{Z}$, the means of the unobserved true responses X and Z, respectively. Thus, a somewhat oversimplified description of the model-based IC approach is (a) estimate the bias in $\bar{A}$ and $\bar{D}$, (b) adjust $\bar{A}$ and $\bar{D}$ for this bias, producing estimates of $\bar{X}$ and $\bar{Z}$ and (c) "unbiasedly" estimate cocaine use prevalence by $\bar{Z}-\bar{X}$. This description is oversimplified since (a)–(c) are accomplished simultaneously by maximizing the joint likelihood of the observations A, B, C and D. In fact, although $\bar{Z}-\bar{X}$ can be computed from the model estimates of P(X) and P(Z), it can be shown that this estimate is identical to the model estimate of $P(Y=1)$.

Applying this approach independently to each IC pair will produce two model-unbiased estimates of cocaine use prevalence (one for each pair) which can be combined to obtain a

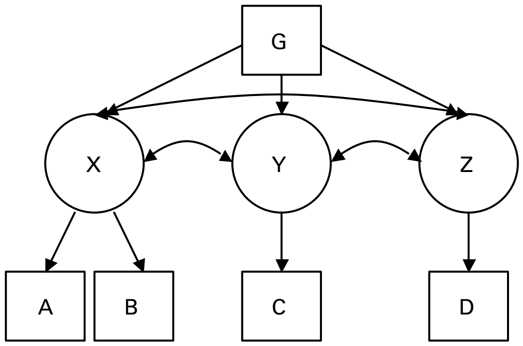


Fig. 3. Path diagram for the basic latent class item count model

single estimate of cocaine use prevalence. Additionally, if the same model is used for both pairs, the two sets of estimates can be compared for the purpose of cross-validating the model.

A simple model for the relationships between the latent variables *X*, *Y* and *Z* and indicator variables *A*, *B*, *C*, and *D* is represented by the path diagram in Figure 3. The variable *G* is a general grouping variable (e.g., respondent gender) that will be discussed subsequently. The true counts of IC behaviors (*X* and *Z*) and cocaine use (*Y*) are mutually correlated as indicated by the arrows between the variables while local independence is assumed for their corresponding indicator variables since there are no arrows between the indicators. Recall that for two indicators of *X*, say *A* and *B*, local independence means that $P(A = a, B = b|X = x) = P(A = a|X = x)P(B = b|X = x)$. Local independence essentially implies that the measurement errors in *A* and *B* are independent.

For indicators *A* and *D* local independence is assured by the random assignment of short- and long-list questions to the two half samples. The assumption that *A* and *B* are locally independent is not assured since the same respondents provide both responses in the same 60-minute interview. Still, the assumption seems plausible if one considers the differences in the response processes generating *A* and *B*. To obtain *A*, respondents read the list of items, tally the number, then apply and record the count. This task does not encourage deep cognitive processing of each item on the list. It requires time to understand the meaning of each item, to search one's memory for any occurrence of the behavior in the past twelve months and to accurately tally and record the correct number. The meanings they attribute to each item could also be influenced by the other items in the list. Respondents could easily commit comprehension, forgetting and counting errors if they answer too quickly.

Since *B* is a composite score consisting of responses to direct questions about the applicability of each item, the types of errors associated with *B* are quite different. Since each item is asked separately, responses to each item may reflect deeper cognitive processing. Evidence of this can be seen in Table 4 where the pseudo-IC count tends to be larger than the IC count indicating less forgetting. Further comprehension errors are likely to be less frequent. Such errors are unlikely to be correlated with the errors in *A*.

Evidence in support of the plausibility of this assumption is provided by Biemer and Wiesen (2002) for past year marijuana use. Biemer and Wiesen considered two questions that appeared consecutively in the NSDUH, both assessing past year marijuana use. One was a direct question about whether marijuana was ever used in the past year and the other was a question on the frequency of marijuana use in the past twelve months. The response processes for these questions are arguably quite different. In their latent class analysis, Biemer and Wiesen used a formal statistical test (not available for our study) to reject a hypothesis of local dependence for the two indicators. Thus, it is plausible that the local independence assumption can hold for two indicators of drug use asked in the same interview but by very different methods.

For the same reasons, it seems plausible to postulate that the errors in C (self-reports of cocaine use) are uncorrelated with the errors in the item count indicator A or D. It is debatable, however, whether the errors in C and D are uncorrelated for individuals whose true response to ICQ(L) is “5” i.e., persons who engaged in all five IC behaviors including the use of cocaine. If these individuals indicated “No cocaine use” in their response to the direct question, they are likely to enter a number less than “5” in their response to the ICQ(L), thus inducing local dependence. However, $P(Z = 5) \approx 0$, by design, and thus will have little effect on the joint conditional distribution of C and D given Z.

One complication in the analysis is that A and B are obtained for half the sample and D is obtained for the other half-sample. Since the half-samples were formed by randomization, the responses for A, B, and D are missing completely at random (MCAR) and can be easily represented in the likelihood of the ABC and CD tables, as shown in the next section.

3.1. Data likelihood

In this section, we derive the joint likelihood for the item count data under a model like that represented in Figure 3. Initially, we assume that the sample is selected by simple random sampling from the population. Methods for dealing with the unequal, clustered sample design of the NSDUH will be discussed subsequently.

The data can be summarized by two cross-classification subtables, denoted by GABC and GCD, corresponding to the split-sample design. Using the results in Rubin and Little (1987, p. 91) for MCAR data, in can be shown that the joint likelihood, $\mathsf{L}(GABC,GCD)$, is proportional to the product of the two subtable likelihoods, i.e., $\mathsf{L}(GABC)\mathsf{L}(GCD)$, where $\mathsf{L}(GABC)$ is the likelihood for the GABC table and $\mathsf{L}(GCD)$ is the likelihood for the GCD table. Next, we show how each log-likelihood can be expressed in terms of the conditional cell probabilities under the model in Figure 3.

To simplify the notation, we let π_u and $\pi_{u|v}$ denote $P(U = u)$ and $P(U = u|V = v)$ for any two random variables U and V. For example, $\pi_G = P(G = g)$, $\pi_{xyz|g} = P(X = x, Y = y, Z = z|G = g)$, $\pi_{abc|xyz} = P(A = a, B = b, C = c|X = x, \quad Y = y, Z = z)$ and so on. For the joint probability $\pi_{xyz|g}$, the equality constraint $Z = X + Y$ must be imposed to represent the dependencies between the short- and long-lists. Note that a true cocaine user ($Y = 1$) who truly engages in x short-list behaviors ($X = x$) can be represented by $Z = x + 1$. Likewise, a true noncocaine user ($Y = 0$) with the same value of X corresponds to $Z = x$. This constraint will be imposed on $\pi_{xyz|g}$ in likelihood setting $\pi_{xyz|g} = 0$ whenever $z \neq x + y$ for $x = 0, 1, \ldots, 4$, $y = 0, 1$ for all g.

The path diagram in Figure 3 indicates that the grouping variable, G, satisfies the following conditions:

(1) structural probabilities depend upon G; i.e., $\pi_{xyz|g} \neq \pi_{xyz}$ and
(2) error probabilities do not depend upon G; i.e., $\pi_{a|xg} = \pi_{a|x}$, $\pi_{b|xg} = \pi_{b|x}\pi_{c|yg} = \pi_{c|y}$ and $\pi_{d|zg} = \pi_{d|g}$.

Hui and Walter (1980) show that for latent class models with two indicators of a single latent variable, these assumptions are sufficient for model identifiability. We extend this condition to the present situation where A and B are two indicators of the latent variable X. Likewise, C and D (through the constraint $Z = X + Y$) are indicators of Y. Thus, by extending the results of Hui and Walter, the model in Figure 3 is identifiable.

Assumptions (1) and (2) can often be adequately satisfied by a judicious choice of the grouping variable, G. A grouping variable that has worked well for this purpose in other studies (see, for example Hui and Walter 1980; Sinclair and Gastwirth 1996; Biemer and Wiesen 2002) is respondent gender. For the current application, Assumptions (1) and (2) imply that the prevalence of cocaine use and the item count behaviors depend upon the respondent's gender; however, the errors in reporting these behaviors do not. Although these assumptions seem plausible for the purposes of the present application, the available data do not permit a test of them.

Ignoring proportionality constants, the log-likelihood of each subtable can be expressed in terms of conditional cell probabilities given the latent variables as follows:

$$\log \mathsf{L}(GABC) \propto \sum_{xyz}{}' \sum_{gabc} n_{gabc} \log(\pi_g \pi_{xyz|g} \pi_{abc|xyz}) \tag{8}$$

and

$$\log \mathsf{L}(GCD) \propto \sum_{xyz}{}' \sum_{gcd} n_{gcd} \log \pi_g \pi_{xyz|g} \pi_{cd|xyz} \tag{9}$$

where $\sum'$ denotes summation over x, y and $z = x + y$.

By the assumption of local independence between A, B,C, and D, we can rewrite $\pi_{abc|xyz}$ as

$$\pi_{abc|xyz} = \pi_{a|x} \pi_{b|x} \pi_{c|y} \tag{10}$$

Note from (10), that the conditional distributions of A, B, and C do not depend on Z. Likewise, $\pi_{cd|xyz}$ can be rewritten as

$$\pi_{cd|xyz} = \pi_{c|y} \pi_{d|z} \tag{11}$$

where it is evident here that the conditional distributions of C and D do not depend upon X. Thus, from (8), the joint log-likelihood of the tables GABC and GCD is

$$\log \mathsf{L}(GABC, GCD) = \log \mathsf{L}(GABC) + \log \mathsf{L}(GCD) \tag{12}$$

where

$$\log \mathsf{L}(GABC) = \sum_{xyz}{}' \sum_{gabc} n_{gabc} \log \pi_g \pi_{xyz|g} \pi_{a|x} \pi_{b|x} \pi_{c|y} \tag{13}$$

and

$$\log\mathsf{L}(GCD) = \sum_{xyz}{}' \sum_{gcd} n_{gcd} \log \pi_g \pi_{xyz|g} \pi_{c|y} \pi_{d|z} \tag{14}$$

Goodman (1973) and Haberman (1979) provide a linkage between latent class models and log-linear models with latent variables and show that much of the statistical theory for log-linear analysis can be directly applied to latent class analysis. Using Goodman's notation, (13) and (14) can be written in hierarchical log-linear model notation as $\{\mathrm{GXYZ, AX, BX, CY|Z = X + Y}\}$ and $\{\mathrm{GXYZ, CY, DZ|Z = X + Y}\}$, respectively. We have altered Goodman's notation slightly with the addition of the $\mathrm{Z = X + Y}$ following the conditioning symbol "|" to emphasize these equality restrictions in the model specification. Combining these two models, the model in (12) can be written as $\{\mathrm{GXYZ, AX, BX, CY, DZ, R|Z = X + Y}\}$, where R is a random indicator variable denoting the random subsample. By the MCAR assumption, R is independent of the other variables in the model.

This model can be easily extended to incorporate additional grouping variables in order to improve the fit of the model. In the subsequent analysis, one grouping variable is considered for this purpose – respondent age (denoted by *H*). Consistent with the analysis of Section 2.3, two age groups are of interest: 12–17 years ($H = 1$) and 18 + ($H = 2$) and we consider models of the form $\{\mathrm{HGXYZ, HAX, HBX, HCY, HDZ, R|Z = X + Y}\}$ and its antecedents.

3.2. Latent class estimation of cocaine use prevalence

In this section, the LCM described in the previous section will be applied to the NSDUH item count data to obtain model-based estimates of cocaine use prevalence. The data for the analysis are the cell counts in the cross-classification tables HGABC and HGCD, where *H* denotes age with two levels, *G* denotes gender with two levels, *A* is the response to the short ICQ with five levels, *B* is the pseudo-ICQ response also with five levels, *C* is the response to the direct question on past 12-month cocaine use with two levels, and *D* is the long ICQ response with six levels. With 200 cells in the HGABC table and 48 in the HGCD table, the total number of degrees of freedom for modeling is 248.

Models were fit to both unweighted and weighted classification tables. The weighted classification tables were formed by summing the weights of the observations in each cell and then rescaling these cell counts so that their sum equaled the total number of observations in the analysis. Ultimately, the weighted analysis was abandoned since none of the models explored in the analysis produced an adequate fit to the data and likelihood maximization was beset by convergence problems such as local maxima and boundary solutions. We suspect this was largely due to the instability of the weighted counts since there were numerous sparse cells that carried very large weights. The unweighted data produced better fitting, more stable models with fewer convergence problems. In addition, the misclassification probability estimates in LCA are seldom influenced by survey weighting (see, for example Patterson, Dayton, and Graubard 2002). To correct the NSDUH estimates of cocaine use prevalence for measurement bias, the misclassification probability estimates from the unweighted analysis will be applied to the weighted (survey) estimates of cocaine use employing the following approach.

Let $\boldsymbol{\pi}_{\mathbf{c}|\mathbf{y}} = [\pi_{c=i|y=j}]$ denote the 2×2 matrix of conditional response probabilities for C and let $\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}$ denote the LCM estimate of $\boldsymbol{\pi}_{\mathbf{c}|\mathbf{y}}$. Let $\hat{\boldsymbol{\pi}}_{\mathbf{c}(NSDUH)} = [\hat{\pi}_{c=1}, \hat{\pi}_{c=0}]'$ denote the 2×1 vector of weighted cocaine use prevalence estimates from the NSDUH and let $\boldsymbol{\pi}_{\mathbf{c}(NSDUH)}$ denote $\mathbf{E}(\hat{\boldsymbol{\pi}}_{\mathbf{c}(NSDUH)})$ with expectation taken with respect to both the NSDUH survey design and the LCM. Let $\boldsymbol{\pi}_{\mathbf{y}} = [\pi_{y=1}, \pi_{y=0}]'$ denote the 2×1 vector of true prevalence. Note that if $\boldsymbol{\pi}_{\mathbf{c}} = \boldsymbol{\pi}_{\mathbf{c}|\mathbf{y}}\boldsymbol{\pi}_{\mathbf{y}}$, it follows that a measurement bias corrected estimator of $\boldsymbol{\pi}_{\mathbf{y}}$ from the LCM is

$$\hat{\boldsymbol{\pi}}_{\mathbf{y}} = \hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}^{-1}\hat{\boldsymbol{\pi}}_{\mathbf{c}(NSDUH)} \tag{15}$$

As mentioned previously, LCMs will be fit separately for ICQ Pairs 1 and 2 to produce two estimates of $\boldsymbol{\pi}_{\mathbf{c}|\mathbf{y}}$ denoted $\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}(1)$ and $\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}(2)$, respectively. A combined estimate of $\boldsymbol{\pi}_{\mathbf{c}|\mathbf{y}}$ will be obtained by averaging the two estimates; i.e., $\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}} = 0.5\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}(1) + 0.5\hat{\boldsymbol{\pi}}_{\mathbf{c}|\mathbf{y}}(2)$.

Table 5 is representative of the range of models that were fit to the item count data. Model selection was confined to only hierarchical linear models, i.e., models that include all lower order interactions and main effects that make up the highest order interaction terms in the model. Several forms for the structural component of the model (i.e., the GHXYZ term) were explored from the most general – viz., {GHXY} – to simpler forms containing all three-way interactions – viz., {GHX, GHY, GXY, HXY}. (Note that the variable Z is redundant since Z = X + Y and therefore need not be included in the structural component.) Ultimately, the form {GHY, GXY, HXY} was selected as the most parsimonious model providing an adequate fit to the data. This model specifies that true cocaine use, *Y*, varies across the four combinations of gender and age. The GXY and HXY terms provide for the mutual dependence of cocaine use and the short-list behaviors, which also vary by gender and age. All the models presented in Table 5 use this form of the structural component.

Lin and Dayton (1997) provide three criteria for selecting the best LC model: (1) the model should be identifiable; (2) the likelihood ratio chi-square *p*-value for the model should be greater than 0.05, indicating that the model fits the data reasonably well, and (3) the Bayesian Information Criterion (BIC) should be the smallest among all competing models. The BIC is defined as $L^2 - (\log N)df$ where L^2 is the likelihood ratio chi-squared statistic and *df* is the model degrees of freedom computed as 248-(number of estimated parameters). It is used in the model selection process to determine the most parsimonious model that fits the data (i.e., satisfies Criterion 2). The dissimilarity index (*d*), which is the proportion of observations that would have to change cells for the model to fit perfectly, provides a fourth criterion. As a rule of thumb, models having $d \leq 0.05$ are considered to fit the data well (Vermunt 1997).

Criterion (2) is much too conservative in the present application since, with almost 70,000 observations, the power of the chi-square test at $p = 0.05$ is approximately 1. The criterion could result in model overparameterization or rejecting a model that still fits the data quite well. Thus, we advocate using a smaller value (say, $p = 0.01$) to allow consideration of models with expected cell probabilities that may only differ trivially from the observed data while satisfying the other selection criteria. The identifiability of each model considered was verified using a sufficient condition suggested by Dayton and

Table 5. Model diagnostics for alternative models by ICQ pair using structural component {GHY, XYG, XYH}

Model		d.f.	ICQ Pair 1				ICQ Pair 2			
			L^2	BIC	*d*	*p*	L^2	BIC	*d*	*p*
0	AX, BX, CY, DZ	144	228.7	−467	0.0092	0.000	211.2	−485	0.0094	0.000
1	AXH, BX, CY, DZ	124	279.3	−320	0.0106	0.000	191.3	−408	0.0097	0.000
2	AX, BXH, CY, DZ	124	174.8	−425	0.0076	0.002	142.6	−457	0.0072	0.121
3	AX, BX, CYH, DZ	142	226.6	−460	0.0091	0.000	199.7	−487	0.0086	0.001
4	AX, BX, CY, DZH	114	207.8	−343	0.0089	0.000	197.8	−353	0.0083	0.000
5	AXH, BXH, CY, DZ	104	136.1	−367	0.0070	0.019	130.8	−372	0.0082	0.039
6	AX, BXH, CYH, DZ	122	170.6	−419	0.0075	0.002	144.3	−446	0.0082	0.082
7	AX, BXH, CY, DZH	94	153.7	−301	0.0070	0.000	130.2	−324	0.0067	0.008
8	AXH, BXH, CYH, DZ	102	170.6	−322	0.0075	0.000	134.2	−359	0.0077	0.018
9	AXH, BXH, CY, DZH	74	156.4	−201	0.0069	0.000	121.6	−236	0.0071	0.000
10	AXH, BXH, CYH, DZH	72	139.1	−209	0.0063	0.000	120.7	−227	0.0070	0.000

Macready (1980), viz., that the variance–covariance matrix for the parameters should be of full rank.

Table 5 shows the model diagnostics for eleven models and for each ICQ pair. Note that, consistent with the Hui-Walter assumptions of Section 3.1, the measurement error components for these models do not include gender (*G*). Model 0 in Table 5, the simplest model considered, is included primarily for comparison purposes. This model assumes that the measurement error does not depend directly on either age or gender. Models 1–4 relax this assumption by adding one age by error interaction term to Model 0. These are the AXH, BXH, CYH, and DZH terms in Models 1–4, respectively. These results suggest that, among the four age by error interactions, BXH provides the most improvement in model fit ($p = 0.002$ and 0.121 for Pairs 1 and 2, respectively).

Models 5–7 all include BXH and add the terms AXH, CYH or DZH, respectively. The best model among these is Model 5 for Pair 1 ($p = 0.019$) and Model 6 for Pair 2 ($p = 0.082$). The remaining models add other interaction by error terms to Model 5 in an attempt to gain further improvement in the fit. Among the models considered in Table 5, only Model 5 satisfies the selection criteria set forth above for Pair 1 ($p = 0.019$). For Pair 2, the model with the smallest BIC having a *p*-value larger than 0.01 is Model 2 (BIC $= -457$ and $p = 0.121$). However, the fit for Model 5 is also quite adequate ($p = 0.039$). An important advantage of using the same model for both ICQ pairs is better comparability and consistency of the model estimates. Model 5 (with 144 parameters) was therefore selected for estimating cocaine use prevalence.

Model 5 suggests that the error rates for the short ICQ (*A*) and pseudo-ICQ (*B*) differ for the two age groups while the errors associated with the direct cocaine question (*C*) and the long ICQ (*D*) do not. To understand why this is plausible, note that both *C* and *D* involve questions about cocaine use (*C* directly and *D* indirectly through the item count) while *A* and *B* do not. It is possible that the errors in *A* and *B* are rooted more in miscounting or misinterpretation while the errors in *C* and *D* are related more to fear of disclosure and privacy concerns. It is also conceivable that the former type of error differs for younger and older age groups while the later type of error does not. As a consequence of the model, a single pair of accuracy rates for the cocaine use reporting will be produced and be applied to both age groups. Denote these estimates by $\hat{\pi}_{c=1|y=1}$ and $\hat{\pi}_{c=0|y=0}$.

3.3. Results

Model 5 was fit to the full NSDUH data set – a total of 68,285 observations. A small number (less than 1 percent) of observations were not included due to item missingness. Table 6 shows the estimates of $\hat{\pi}_{c=1|y=1}$ (i.e., probability of a correct classification given a true cocaine user) and $\hat{\pi}_{c=0|y=0}$ (i.e., probability of a correct classification given a true nonuser) for Model 5 for each ICQ pair and the average of these estimates. The estimates are remarkably consistent for each pair, which gives support to the model's validity. In each case, the accuracy rate for reporting use is approximately 0.70, which equates to a false negative rate of about 30%. Both ICQ pairs estimate the false positive rate to be essentially 0, which is plausible given the stigma associated with cocaine use.

The subsequent discussion and analysis will focus on the average estimate in the second to last column. These accuracy rates can be directly applied to the NSDUH estimates of

Table 6. Estimates of classification accuracy (i.e., $\hat{\pi}_{c=1|y=1}$ and $\hat{\pi}_{c=0|y=0}$) from Model 5

	Pair 1	s.e.	Pair 2	s.e.	Average	s.e.
$\hat{\pi}_{c=1\|y=1}$	.6953	.1961	.7065	.2835	.7009	.1724
$\hat{\pi}_{c=0\|y=0}$	.9988	.0012	.9993	.0012	.9991	.0004

cocaine use prevalence to obtain measurement bias corrected estimates of cocaine use prevalence using Equation (15). Since $\pi_{c=0|y=0} \approx 1$, (15) simplifies to

$$\hat{\pi}_{y=1|k} = \frac{\hat{\pi}_{c=1|k(NSDUH)}}{\hat{\pi}_{c=1|y=1}} \tag{16}$$

for any group k defined by age and gender where $\hat{\pi}_{c=1|y=1}$ is the average of the estimates in Table 6 and $\hat{\pi}_{c=1|k(NSDUH)}$ is the survey weighted estimates of cocaine use prevalence from the NSDUH for group k (see, for example Table 2).

Approximate standard errors for the estimator in (16) can be estimated using the delta method assuming that the covariance between $\hat{\pi}_{c=1|y=1}$ and $\hat{\pi}_{c=1|k(NSDUH)}$ is negligible. Note that $\hat{\pi}_{c=1|y=1}$ and $\hat{\pi}_{c=1|k(NSDUH)}$ are likely to be positively correlated since higher reporting accuracy among true users produces larger estimates of cocaine use prevalence. Hence, the negligible covariance assumption will likely result in overstating the variance. The approximate variance of (16) is given by

$$Var(\hat{\pi}_{y=1|k}) \cong \left(\frac{1}{\pi_{c=1|y=1}}\right)^2 Var(\hat{\pi}_{c=1|k(NSDUH)}) + \left(\frac{\pi_{c=1|k(NSDUH)}}{\pi_{c=1|y=1}}\right)^2 Var(\hat{\pi}_{c=1|y=1}) \tag{17}$$

Estimates of $Var(\hat{\pi}_{c=1|k(NSDUH)})$ and $Var(\hat{\pi}_{c=1|y=1})$ are obtained by squaring the standard errors in Table 2 and Table 6, respectively. The LCM estimates of $\pi_{y=1|k}$ for all age and gender margins and combinations are reported in Table 7 (in percent) along with their standard errors.

There are very little external data available to test the validity of the model-based IC estimates in Table 7 or to even establish that they have smaller absolute bias than the NSDUH. Wright et al. (1997) compared the 1992 NHDUH estimates of drug use prevalence to estimates derived from various administrative systems (drug treatment programs data; parole, probation and arrest records, etc.) and regarded the administrative records as the gold standard. They found significant underreporting in the NSDUH for the drugs they evaluated. Unfortunately, past year cocaine use was not included in their evaluation so there is no direct comparison between our estimates and theirs. Moreover, the NSDUH has undergone several important design changes since 1992 aimed at improving reporting accuracy including the adoption of ACASI and the use of incentives (Wright, Barker, Gfroerer, and Piper 2002). The results still provide an indication of the magnitude of the underreporting in the NSDUH for stigmatized drugs such as cocaine.

Wright's et al. estimate of the NSDUH classification accuracy was 54.9 percent for past year heroin use and 89.4 percent for past year marijuana use. Our model-based IC estimate of 70 percent falls between their estimate of heroin and marijuana use. This is reasonable since currently the most stigmatized drug appears to be heroin, followed by cocaine and

Table 7. NSDUH and model-based IC estimates of past year cocaine use prevalence (in percent) by gender and age

	NSDUH	s.e.	Model 5	s.e.
Total	1.9	0.08	2.71	0.36
Gender				
Male	2.6	0.14	3.71	0.44
Female	1.1	0.08	1.57	0.28
Age				
12–17	1.5	0.1	2.14	0.33
18+	1.9	0.09	2.71	0.36
Gender by Age				
Male				
12–17	1.4	0.14	2.00	0.35
18+	2.8	0.15	3.99	0.46
Female				
12–17	1.5	0.15	2.14	0.37
18+	1.1	0.08	1.57	0.28

then marijuana. As Harrison (1997) notes, the more stigmatized the drug, the greater the under-reporting, a finding which has been replicated in several other studies (Harrison 1995; Fendrich and Xu 1994; Mieczkowski et al. 1991). The Wright et al. study suggests that the estimates in Table 7 are quite plausible, although it is not sufficient to establish the validity of the model-based IC approach.

Additional research is now underway to assess the validity of the model-based estimates. Biological specimens (hair and urine) were collected from a sample of about 4,500 respondents aged 12–25 from the 2000-2001 NSDUH surveys (Odum and Chromy 2003). These data are being analyzed to produce estimates of the bias in the NSDUH estimates for cocaine as well as other substances. The validity of both the model-based IC estimates and the biological specimen estimates will be assessed and reported in a subsequent article.

4. Summary and Discussion

The motivation for this work was the poor performance of the simple IC estimator despite considerable effort and research to refine the IC method for NSDUH use. Designed to remove the negative bias in estimates based on self-reported drug use, the simple IC estimator produced implausible cocaine use prevalence estimates. All the estimates were lower than those of the self-report estimator and in few cases were less than 0. Reliability analysis of the IC questions revealed that the responses contained considerable measurement error, which would explain the poor performance of the simple IC estimator. This article has investigated a new, model-based estimator that attempted to correct the IC estimates for measurement error, thus producing less biased estimates of the prevalence of the sensitive item.

The model-based approach combined data from the short-list, long-list, pseudo-item counts, and the direct cocaine use questions to obtain estimates of the classification error in the observed data. The data were treated as fallible indicators of (latent) true values and traditional latent class analysis assumptions were made to obtain an identifiable model. The key parameters estimated from the model were false positive and false negative rates for self-reported cocaine use. These error rates can be studied for their own analytical interest or be used to correct the NSDUH published estimates of cocaine use prevalence for reporting bias.

The validity of the estimates was addressed in two ways. First, separate models were fit to the data from both ICQ pairs as a way of cross-validating the estimates from each. The model selection process identified a model that fit the data from both ICQ pairs quite well and produced estimates that were remarkably similar. Since the two sets of estimates use different IC questions, the close correspondence of the estimates supports the validity of the estimation approach. In addition the model-based estimates were compared to the estimates from other studies. In particular our estimates of drug use reporting accuracy were consistent with the corresponding estimates from an administrative records study conducted by Wright, Gfroerer, and Epstein (1997) for the 1992 NSDUH.

The best model (Model 5 in Table 5) postulates that the errors in the short-list IC responses and the pseudo-IC responses depend upon respondent age while the errors in the direct cocaine use question and long-list responses do not depend upon age. We speculate that this is due to the nature of the errors for the two types of questions. Questions involving cocaine use are likely to elicit errors due to fear of disclosure, and such fear is present for both age groups. The more innocuous items contained in the short-list ICQ are less subject to disclosure concerns, allowing counting and comprehension errors to dominate, which are more likely to differ between the two age groups.

For both ICQ pairs, the estimates of cocaine use underreporting and overreporting error were the same for both age groups: 30 percent underreporting and approximately 0 percent overreporting error. This suggests that the NSDUH prevalence estimates should be increased by the factor $(0.7)^{-1} = 1.43$. As an example, given the NSDUH estimate of 1.9 percent for cocaine use prevalence, the model-based IC estimate is (1.9 percent) × 1.43 = 2.7 percent.

The likelihood maximization process frequently encountered problems with local maxima, boundary solutions, and implausible estimates that presumably were due to survey weight variation, sparse cells and model complexity particularly with regard to the structural component of the model. Several steps were taken to alleviate these problems, including: (a) using unweighted data, (b) reducing the structural component to three-way interaction terms only and (c) running each model up to 100 times with different starting values and choosing the solution corresponding to the largest of the likelihood maxima.

Another potential option for addressing these difficulties is to incorporate additional grouping variables in the model that are more highly correlated with measurement error than our age variable. Although age explained a large proportion of the variation in the structural component, it was less useful for explaining measurement variance. Grouping variables such as education (for modeling comprehension error) and social-economic status (for disclosure or privacy concerns) may be better choices for modeling measurement variance. Survey designers should also consider collecting information in the survey for the specific purpose of modeling the measurement error. As an example, to aid in the modeling of errors due to deliberate misreporting of drug use, questions could be added to the questionnaire that directly assess the respondent's attitudes and concerns regarding privacy and confidentiality of the survey results.

Our work raises questions about the efficacy of the item count methodology for estimating the drug use prevalence in surveys. For the NSDUH application, considerable research was devoted to enhancing and adapting the IC methodology for past year cocaine use prevalence estimation. Despite these efforts, the methodology failed to produce estimates of cocaine use that were even at the level of those obtained by simply asking respondents directly about their cocaine use. In fact, our findings suggest that the simple IC estimator is even more biased than estimates based upon direct drug use questions. We are skeptical as to whether additional refinements of the IC methodology would produce more useable results. The complexity of the IC task and respondent concerns about privacy are inherent issues that will always lead to some amount (albeit acceptable in some cases) of measurement error. The model-based approach provides means for dealing with these unavoidable measurement errors by taking them into account in the estimation process.

Thus, we advocate the use of measurement error modeling as an integral part of the IC methodology. To that end, we recommend that the items in the short-list IC question be asked directly of the same respondents receiving the short-list IC question. These data can be used to compute the reliability of the IC responses as shown in Section 2.3 and can also be incorporated into the estimation process to correct for the measurement error using models such as the one in Figure 3. It is informative to compare the two sets of estimates. If they are consistent, the validity of the simple IC estimate is supported. Otherwise, the model-based approach provides the possibility of computing an alternative and possibly improved estimator of the prevalence rate.

5. References

Biemer, P., Jordan, B.K., Hubbard, M., and Wright, D. (In Press). A Test of the Item Count Methodology for Estimating Cocaine Use Prevalence. In J. Kennet and J. Gfroerer (eds). Evaluating and Improving Methods Used in the National Survey on Drug Use and Health. (DHHS Publication No. SMA O5-4044, Methodology Series M-5.) Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

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

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, 97–119.

Chromy, J., Davis, T., Packer, L., and Gfroerer, J. (2002). Mode Effects on Substance Use Measures: Comparison of 1999 CAI and PAPI Data. In J. Gfroerer, J. Eyerman, and J. Chromy (eds). Redesigning an Ongoing National Household Survey: Methodological Issues, DHHS Pulication No. SMA 03-3768. Rockville: SAMHSA, Office of Applied Studies, 135–157.

Dayton, C.M. and Macready, G.B. (1980). A Scaling Model with Response Errors and Intrinsically Unscalable Respondents. Psychometrika, 45, 343–356.

Droitcour, J., Caspar, R.A., Hubbard, M.L., Parsley, T.L., Visscher, W., and Ezzati, T.M. (1991). The Item Count Technique as a Method of Indirect Questioning: A Review of Its Development and a Case Study Application. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, and S. Sudman (eds). Measurement Errors in Surveys. New York: John Wiley and Sons.

Fendrich, M. and Xu, Y. (1994). The Validity of Drug Use Reports from Juvenile Arrestees. International Journal of Addiction, 29, 971–985.

Fisher, R.J. (1993). Social Desirability Bias and the Validity of Indirect Questioning. Journal of Consumer Research, 20, 303–315.

Goodman, L. (1973). The Analysis of Multidimensional Contingency Tables When Some Variables Are Posterior to Others: A Modified Path Analysis Approach. Biometrika, 60, 179–192.

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

Harrison, L.D. (1995). The Validity of Self-Reported Data on Drug Use. Journal of Drug Issues, 25, 91–111.

Harrison, L. (1997). The Validity of Self-Reported Drug Use in Survey Research: An Overview and Critique of Research Methods. In L. Harrison and A. Hughes (eds). The Validity of Self-Reported Drug Use: Improving the Accuracy of Survey Estimates. U.S. Department of Health and Human Services, National Institutes of Health, NIDA Research Monograph, 167, 17–36.

Heinen, T. (1996). Latent Class and Discrete Latent Trait Models. Thousand Oaks, CA: Sage Publications.

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

Lin, T.H. and Dayton, C.M. (1997). Model Selection Information Criteria for Non-Nested Latent Class Models. Journal of Educational and Behavioral Sciences, 22, 249–264.

Mieczkowski, T. (1991). The Accuracy of Self-Reported Drug Use: An Evaluation and Analysis of New Data. In R. Weisheit (ed.). Drugs, Crime and the Criminal Justice System. Anderson Publishing Co. and the ACJS, Cincinnati, OH, 275–302.

Mieczkowski, T., Barzelay, D., Gropper, B., and Wish, E. (1991). Concordance of Three Measures of Cocaine Use in an Arrestee Population: Hair, Urine and Self-report. Journal of Psychoactive Drugs, 23, 241–249.

Odom, D.M. and Chromy, J.R. (2003). 2000-2001 National Household Survey on Drug Abuse (NHSDA) Validity Study: Sample Design Report. Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

O'Reilly, J., Hubbard, M., Lessler, J., Biemer, P., and Turner, C. (1994). Audio and Video Computer-Assisted Self-Interviewing: Preliminary Tests of New Technologies for Data Collection. Journal of Official Statistics, 10, 197–214.

Patterson, B., Dayton, C., and Graubard, B. (2002). Latent Class Analysis of Complex Survey Data: An Application to Dietary Data. Journal of the American Statistical Association, 97, 721–729.

Rubin, D. and Little, R. (1987). Statistical Analysis with Missing Data. New York: John Wiley and Sons.

Sinclair, M. and Gastwirth, J. (1996). On Procedures for Evaluating the Effectiveness of Reinterview Survey Methods: Application to Labor Force Data. Journal of the American Statistical Association, 91, 961–969.

Turner, C.F., Lessler, J.T., and Devore, J. (1992). Effects of Mode of Administration and Wording on Reporting of Drug Use. C.F. Turner, J.T. Lessler, and J.C. Gfroerer (eds). Survey Measurement of Drug Use-Methodological Studies. Washington, DC: U.S. Government Printing Office, 221–244.

Vermunt, J. (1997). Log-linear Models for Event Histories. Thousand Oaks, CA: Sage Publications.

Wright, D., Barker, P., Gfroerer, J., and Piper, L. (2002). Summary of NHSDA Design Changes in 1999. In J. Gfroerer, J. Eyerman, and J. Chromy (eds). Redesigning an Ongoing National Household Survey: Methodological Issues. DHHS Publication No. SMA 03-3768, Rockville: SAMHSA, Office of Applied Studies, 9-22.

Wright, D., Gfroerer, J., and Epstein, J. (1997). Ratio Estimation of Hardcore Drug Use. Journal of Official Statistics, 13, 401–416.

Received March 2005
Revised May 2005