Assessing the reliability of key measures in the National Survey on Drug Use and Health using a test-retest methodology

Joel Kennet[1], Dicy Painter[1], Susan R. Hunter[2], Rebecca A. Granger[2] & Katherine R. Bowman[2]

[1]Substance Abuse and Mental Health Services Administration [2]RTI International

Address correspondence to joel.kennet@samhsa.hhs.gov

Abstract

The National Survey on Drug Use and Health (NSDUH) is a major source of information on substance use and mental illness prevalence in the United States. It is administered in households to approximately 67,500 individuals annually using a complex, multi-stage sampling design. Assessing the reliability of estimates produced by the NSDUH is of primary importance to those who use these data for research and in the making of policy decisions. In Quarters 1 and 2 of 2005, a pre-test was carried out, in which approximately 200 NSDUH respondents were re-interviewed in an effort to fine tune the methods to be used in conducting a large-scale reliability field test in 2006. This paper will discuss the design and procedural considerations that were taken into account in planning the pre-test and upcoming field test. These considerations included time interval between test and re-test, sample size needed for reliability estimates of low-prevalence behaviors, whether sample would be embedded or not embedded in the NSDUH main study, using the same vs. different interviewers for the re-interview, increased risk of loss of respondent privacy due to the provision of re-contact information, amount of incentive for the re-interview, and others. In addition, preliminary findings from the pre-test that may influence methods employed in the 2006 field test, such as response rates on the re-interview, and respondent feedback, will be presented and discussed.

Introduction

The reliability of survey data on sensitive topics

The federal government is a huge consumer of survey data, and in some cases, these data influence major policy decisions. Thus, information on the quality of such data should be a standard output of major surveys. Most do a good job reporting on response rates and sampling error, but not on other types of error, such as measurement error. Response rates, while capable of indicating potential sources of bias in data, can not truly measure the accuracy with which data are being gathered. Determining the reliability of survey measures by re-interviewing respondents provides a direct measure of response variance. In other words, the capability of the survey to provide accurate data, and consequent population estimates, can be determined by carrying out an assessment of its reliability. The reliability of survey data is of particular concern when the data reflect responses to questions that are sensitive in nature, i.e., when respondents are asked to reveal personal characteristics and behaviors that they feel might be embarrassing or damaging to themselves if certain others were to find them out.

The National Survey on Drug Use and Health (NSDUH) is a cross-sectional household survey of the civilian, non-institutionalized U.S. population aged 12 and older. The survey gathers data on the recency and frequency of use of alcohol, tobacco and illicit substances from approximately 67,500 respondents per year in all 50 states and the District of Columbia. Recognizing that drug use, and particularly illicit drug use, is likely to be considered a sensitive topic, NSDUH staff have incorporated state-of-the-art methods to assure respondents of their privacy and confidentiality in providing truthful data. For example, in order to minimize the possibility of identification, household screening data are collected on a hand-held computer and kept separate from the actual survey data, which are collected anonymously on a laptop. The majority of the survey is administered in ACASI mode so even the interviewers are not aware of responses to the more sensitive portions of the questionnaire. However, despite these and other precautions related to privacy and confidentiality, the reliability and validity of NSDUH data continue to be questioned. This is for good reason, since respondents' intent toward truthfulness is only one of the requirements in collecting reliable and valid data. In addition, the measurement instrument must be carefully designed in order to be capable of capturing the respondents' and researchers' intended communication.

The relation between reliability and data quality

At this point it makes sense to narrow the definition of reliability to one that is manageable within the scope of this paper. We refer to reliability as the extent to which respondents answer alike when the same questions are presented on two occasions separated by a specified time period. This definition reveals the main reason that good reliability must be viewed as but one requirement of a quality data set. To yield good reliability, respondents need only provide highly similar, *but not necessarily truthful*, answers on two occasions. For data to be accurate, there must be high reliability *and* the responses must reflect the objectively true state of affairs. Finding out the objective truth about respondents' drug use behaviors is expensive and problematic. Data on reliability provide evidence that can inform notions of its accuracy. High estimates of reliability would provide modest support to the notion that the data are accurate. On the other hand, estimates of low reliability would signal with relative certainty that the data are not to be trusted. If respondents provide disparate answers to a factual question presented on two occasions, for instance age of first use of a particular substance, one or both of their answers must be false, and therefore this question must be considered a likely source of data that are not valid. In essence, assessing the reliability of the NSDUH questionnaire is seen as an important step that must be taken in order to quantify the error in key survey estimates that is due to response variance, and eventually to use this information to improve the instrument where improvements are needed.

Design considerations for the NSDUH reliability field test

In an effort to assess the reliability of data obtained in the NSDUH, a full-scale field test is planned for 2006, wherein a fairly large number of respondents will complete the survey twice. In the first two quarters of 2005, a pre-test of the NSDUH reliability field test was carried out in order to test and fine tune the methods decided upon in the planning phase. There were many decisions to make in the planning of this work. This section will describe the planning and decision-making involved in determining the procedures used in the 2005 field pre-tests, and the resulting changes that were decided upon for implementation in the 2006 full-scale study.

Initial constraints included that respondents would not be informed of the re-interview until after completing the initial one. This was done in order to maximize the independence of the two interview occasions. Interviewers would also be kept unaware of which respondents would be asked to complete a second interview. This was to prevent interviewers from mentioning the possibility of a re-interview and potentially affecting response rates. It was also desirable that opportunities for respondents to reveal the re-interview component to other potential respondents would be kept to a minimum. Of course, materials and procedures for re-contact cases would be kept as similar as possible to those for non-recontact cases.

Sampling considerations

Perhaps the most difficult determinations related to the NSDUH reliability study involved the size and composition of the sample. The sample had to be large enough to gain relatively precise reliability estimates, but small enough so as not to impact data collection for the main study. The population of interest was the same as that in the main study, but certain exclusions were needed for the sake of practicality or in order to satisfy the constraints imposed by the basic design. These issues are discussed below.

The effect of prevalence on sample size. The major determinants of sample size were the prevalence of the most rare behaviors of interest, and the desired precision of an estimated value of Cohen's Kappa, the statistic used in most cases as the reliability index. Because some illicit substances are reported to be used by only tiny fractions of the population, determining the reliability of the questions that ask about those substances would require sufficient numbers of people who report positively. It is difficult to anticipate the values of Kappa that will be observed in this study, but a fairly high value (around 0.90) appears plausible given the short period between the first and second interviews. With an estimated Kappa = 0.90, prevalence measures for the three standard age groups of 12-17, 18-25, and 26+ that are greater than 0.04 would achieve the desired asymptotic standard error of Kappa less than 0.05 when the effective sample size is about 600 per age group. For this total effective sample size of 1800, an asymptotic standard error of Kappa less than 0.05 would be achieved for prevalence measures of the total population greater than 0.02. If the interviewer effect from the same versus different interviewer sub-study[1] is large, it may not be reasonable to combine the data from the two sub-studies. Considering the Same-Interviewer sub-study alone and assuming a Kappa of 0.90, reasonable estimates of Kappa (asymptotic standard error less than 0.05) can be achieved for dichotomous measures by age group with a prevalence rate greater than 0.06.

[1] The same vs. different interviewer sub-study is defined in a later section.

Number of respondents selected per household. In the main NSDUH study, selected households can have 0, 1 or 2 respondents selected to complete the survey. This presented a procedural problem in designing the reliability study. When two persons are selected, it is not feasible to interview them simultaneously. Thus, the first respondent, when asked to complete a re-interview, might inform the second respondent of this possibility and in doing so would violate the constraint that respondents not be informed of the re-interview until after completing the initial interview, which was put in place to avoid contamination of the data. In other words, if the respondent knows beforehand that the survey will be administered twice, he or she might provide different responses on the first occasion from those he or she would have provided without such knowledge. Foreknowledge of the re-interview could also affect likelihood of participation, which would also be undesirable. It was therefore decided that re-interview cases would only be carried out in situations where only one person in the household was selected. This constraint was also applied so that group quarters dwelling units were excluded from the sample.

Spanish interviews. Approximately 2100 - 2200 respondents each year since 1999 have chosen to complete the NSDUH in Spanish, an option that is carried out using a bilingual interviewer, translated instrument and a full set of translated auxiliary materials (informed consent, showcards, etc.). In many of these cases, the household is screened in English, with an interpreter present if necessary, and a bilingual interviewer is later called upon to administer the questionnaire. For the reliability study, this would add considerable cost in the event that a bilingual interviewer would be needed twice, since these individuals often travel long distances in order to carry out interviews in place of their non-Spanish-speaking peers. In addition, a considerable amount of the auxiliary materials used in the main study had to be revised at least slightly for reliability study cases, and these would need to be translated, tested and printed. Because the proportion of respondents who choose to complete the survey in Spanish is small, and the costs of carrying out the reliability study in Spanish were projected to be quite high, it was decided that respondents who requested a Spanish interview would not be included in the pre-test or in the full-scale reliability study.

Alaska and Hawaii. Respondents in these states were excluded from the reliability study for reasons of cost and feasibility. In both cases, a small number of interviewers cover a large amount of territory that is difficult and costly to traverse. Administering interviews often involves flying between islands, and in the case of Alaska, deep into wilderness. Adding to the caseloads of interviewers in these places is simply not feasible, and bringing in additional interviewers to handle reliability study cases was considered to be too costly.

Field considerations

In addition to sample size and composition, several other factors needed to be taken into account. The decision on whether the reliability study sample would be drawn from within the NSDUH main sample was particularly difficult. Additionally, the specific interviewers carrying out the reliability cases were a consideration. Finally, the specific temporal parameters of the data collection were considered. These topics are discussed in the following paragraphs.

Embedded vs. non-embedded sample. One of the main considerations in designing the full reliability study was whether to re-contact a subset of respondents from the main study, or draw a separate sample. Obviously, cost was a major consideration here. If main-study respondents could be used for the reliability study as well, the cost of fielding the reliability study would essentially be cut in half compared with the non-embedded alternative. However, there were reasons to consider this decision carefully. Using an embedded sample held the danger of contamination. In spite of the decision to only use respondents from households in which one person was selected, there was still the danger that a respondent selected for the reliability study could inform a neighbor whose household was also selected. Again, this would endanger the data obtained in the "informed" household, even if it was selected only for the main study. Because of these risks, it was decided that, for the pre-test at least, a non-embedded sample would be drawn. However, this decision was re-examined for the 2006 full scale study, and the current plan is to use an embedded sample. The sample will be distributed across the nation for the full study so it is less likely that contamination will occur in the selected areas than with the small pre-test sample in only four states.

Same vs. different interviewers. One factor that could influence the results of the reliability assessment using test-retest methodology was the possible effect of individual interviewers. Prior research on the NSDUH (Chromy, Eyerman, Odom, McNeeley & Hughes, 2005; Hughes, Chromy, Giacoletti & Odom, 2001, 2002) revealed an effect of interviewer experience on respondent likelihood of reporting substance use. For reasons that continue to be investigated, prevalence rates among respondents interviewed by veteran interviewers tend to be lower than rates among respondents interviewed by field interviewers with less than a year of experience. Curiously, these effects continue to be observed, although they have been attenuated somewhat, after the switch from paper-and-pencil to ACASI administrative mode. Since it was clear that

interviewers were capable of affecting the data, it was decided that a portion of the reliability study sample would receive their second interview from a different person than the one who conducted the first one. In this manner, comparison of reliability estimates between same interviewer vs. different interviewer cases could be carried out. One third of the sample was thus randomly allocated into the different-interviewer condition, with the remaining two thirds in the same-interviewer condition. This manipulation could help to tease out the causes of some of the observed response variability.

Time interval between occasions. Another decision that required careful consideration was the length of time between interview occasions. Several factors were considered. Foremost among these was the NSDUH definition of current use, which is any use within the past 30 days. If the interview occasions were to be separated by too much time, reliability estimates for these measures could be expected to be quite low. In addition, prior experience with longitudinal data collection indicates that as the amount of time increases between interview occasions, the amount of personal information needed for re-contact of respondents increases. This was particularly relevant for the NSDUH, which asks many sensitive questions and goes to great lengths to assure respondents of confidentiality. On the other hand, placing the interviews too close together in time would reduce the independence of the two sets of observations, since respondents could rely on their memory of previous responses in order to complete the second interview. A brief search of the most relevant literature revealed that a minimum of five days between occasions was needed, so that interval was chosen.

Time window for second occasion. Another consideration involved the length of time during which respondents could complete the second interview. While a short window would be helpful in yielding a tight measure of reliability, fielding the study under such a constraint was foreseen to be difficult, especially when the interviewer on the second occasion was different from the first. A longer window would yield the ability to measure the effect of the length of the T1-T2 interval on reliability, and would make it easier to accommodate the different-interviewers condition. Thus it was decided that the second interview would need to take place 5 to 15 days after the initial interview. In the pre-test, the average length of time between interviews turned out be 8.34 days, with a range of 4 to 16 days.

Time frame for data collection. In order to minimize opportunities for data contamination, it made sense to spread out the collection over time. This would leave little opportunity for respondents to inform potential respondents of the two-interview manipulation. Spreading the collection out over four quarters also allowed the option to continue collecting data for the study into 2007.

Collection of re-contact information. The amount of information needed in order to re-contact reliability study respondents was also brought into question. The NSDUH main study carries out its own counting and listing operation, and the screener does not obtain any single piece of information that would personally identify a respondent. Thus, addresses are available to interviewers for re-contact, but the individual respondent within the household would need to be identified to some extent. It was decided that additional identifying information should be kept to a minimum. Respondents in the different-interviewer condition would be asked for their telephone numbers, and this, combined with the age and gender data collected on the first occasion, was considered sufficient for the second interviewer to identify the respondent.

Amount of incentive for re-interview. At present, there are no measures in the NSDUH that assess respondent level of satisfaction with the $30 incentive that is presently provided in the main study. What is known is that response rates rose after the introduction of the incentive, and the cost of fielding the survey was reduced (Kennet, Gfroerer, Bowman, Martin & Cunningham, 2005). A high response rate for the second interview was seen as especially important if a non-embedded sample would be used, since cases where respondents completed only the first interview would not be used in the main study. High incentives were also seen as important for the different-interviewer condition, lest the initial respondent or respondent's parents become hesitant to allow an additional stranger to enter their household and ask a second set of personal questions. It was decided that $50 would likely be sufficient to maintain high response rates on the second occasion without incurring excessive costs, especially considering the 5 to 15 day time window for recontacting respondents. Evidence from the pre-test indicates that this amount was indeed sufficient to maintain good participation through the second interview. The $30 incentive was maintained for participating in the first interview.

Final sample design

Taking all of the above into consideration, the sample was allocated as follows. The sample will be an area-based probability sample of 400 (out of 876) State sampling (SS) regions. Eight area segments per SS region will be sampled over all four quarters of 2006. This allocation was designed to minimize the chances of contamination of the main study, and assure that

an adequate number of segments would be unaffected in the event of a finding that the conduct of the reliability study somehow did contaminate the main NSDUH data collection. Approximately two-thirds of the cases will have both interviews administered by the same interviewer and one-third will be administered by different interviewers at times one and two.

Like NSDUH, persons eligible for the study will be the civilian, noninstitutionalized population aged 12 or older. Unlike the main study, the reliability field test will exclude residents of noninstitutional group quarters (e.g., shelters, rooming houses, dormitories) and residents of Alaska and Hawaii. Addtionally, re-interviews will not be conducted in Spanish.

The reliability field test sample will be embedded within the main study. Only households in which one person is selected will be eligible. Tables 1 and 2 show that approximately 26,098 selected main study dwelling units will be needed to yield a total of 3,100 completed re-interviews. This is assuming a 91 percent screening response rate (SRR) among eligible dwelling units, an 82 percent interview response rate (IRR) for initial interviews, and a 92 percent IRR for re-interviews. The expected overall response rate (ORR) is 69 percent. Assuming a design effect of 1.7, the effective sample size is approximately 1,800.

Table 1. Reliability Field Test Design Parameters: Dwelling Unit Level

Total Sample	Rate	Number
State Sampling Regions (SSRs)		400
Segments		3,200
Selected Lines		26,098
Expected Eligible Dwelling Units	0.18	4,698[2]
Expected Completed Screening Interviews	0.91	4,275

Table 2. Reliability Field Test Design Parameters: Person Level

	Age Domain							
	Overall		12-17		18-25		26+	
Total Sample	Rate	Number	Rate	Number	Rate	Number	Rate	Number
Expected Selected Persons (First Interview)	1.00	4,275	1.00	1,288	1.00	1,351	1.00	1,636
Expected Completed Interviews (First Interview)	0.82	3,488	0.89	1,140	0.84	1,133	0.74	1,214
Expected Selected Persons (Second Interview)[3]	0.97	3,372	0.99	1,123	0.96	1,088	0.96	1,161
Expected Completed Interviews (Second Interview)[4]	0.92	3,100	0.92	1,033	0.95	1,033	0.89	1,033
Expected Sample Size Based on Assumed Design Effect	1.70	1,824	1.70	607.8	1.70	607.8	1.70	607.8

Description of the pre-test

[2] Based on prior NSDUH experience, 16 percent of the selected dwelling units are expected to be ineligible (institutional, nonresidential, etc.). Of the eligible dwelling units, 21 percent are expected to result in a single person selection. Thus, the 26,098 dwelling units are reduced to 4,698 eligible dwelling units.

[3] Since the second interview will be conducted in Spanish, respondents who completed the first interview in Spanish will not be selected to complete the second interview. The selection rate for the second interview is based on the number of Spanish interviews completed in 2003.

[4] Rates are based on actual experience during Phases I and II of the 2005 Reliability Study Pretest.

The NSDUH reliability study was pre-tested in the first two quarters of 2005 on a small sample in the states of Florida, Maryland, North Carolina, and Texas. The pre-test was carried out in two phases, each lasting two months and involving about 100 cases in which the respondent was re-interviewed. A non-embedded sample was drawn for this purpose, using retired segments. The purposes of the pre-test were to examine and improve the materials and procedures intended for use in the full reliability study, to gather preliminary data on response rates, some very preliminary measures of reliability, and to gauge respondents' and interviewers' reactions to the study.

The two-phase design was intended to allow for adaptations to be made in time for the second phase. Since the Phase I pre-test results confirmed that the instrumentation and procedures developed for the full reliability study were feasible when used in the field, the only change implemented for Phase II was to the iPAQ instrumentation so that it would be similar to the instrumentation that would be used in the reliability field test. Since the 2006 reliability study was planned to be an embedded sample, non-reliability cases were included in each pre-test segment. For these few cases, the interview was similar to a main study interview, in that respondents were not asked to complete a re-interview.

Pre-test Results

A series of feedback questions for respondents and interviewers was designed and adapted for use in each of the possible respondent cooperation scenarios. The questions were designed to assist NSDUH staff with decisions regarding the size of the incentive, reasons for non-cooperation, and other aspects of the reliability study design. Responses to these questions are presented below, as are response rates from the pre-test, and some preliminary reliability estimates.

Respondent Follow-up Questions. There were two types of respondent follow-up questions implemented in the pre-test, one for respondents who completed the re-interview and the other for re-interview non-responders. Respondents who completed the re-interview were asked an additional set of ACASI questions at the end of the re-interview to gather their feedback on completing the two interviews. Most respondents (over 70 percent) reported they remembered most or all of their answers to the tobacco, alcohol, and marijuana questions from the first interview. Additionally, the majority (over 89 percent) of respondents reported that most or all of their answers to the tobacco, alcohol, and marijuana questions were the same for both interviews.

Respondents who refused to complete the re-interview or were unable to contact for the re-interview were asked the follow-up questions via telephone during the verification process to determine why they did not complete a second interview. There were eight people who refused at the end of the initial interview, four people who refused when the interviewer returned to complete the re-interview, and five people who were unable to be contacted for the re-interview. Of these seventeen non-responders, there is no verification data for six respondents who were unable to be contacted for the verification interview and one respondent who did not provide a phone number for verification contact.

All of the people who were unable to be contacted for the re-interview responded that they would have participated if they had been available. Of the seven refusal non-responders contacted, only one person responded that the $50 payment was not enough. For those respondents who refused at the end of the initial interview, three people responded that they could not take the time to do another interview and one person reported not wanting to complete a re-interview because the questions were too personal. However, two of the respondents who refused indicated that they would have participated if they had been available.

Interviewer Debriefing[4]. To assess respondent's reactions to the re-interview recruitment process and the re-interview, interviewers were asked a series of questions in the CAI immediately following the initial interview and the re-interview. In addition, after each pre-test phase, interviewers participated in debriefing calls to discuss their pre-test experiences. The results of these CAI debriefing questions and interviewer debriefing calls are reported below.

Overall, interviewers reported that the re-interview recruitment process flowed smoothly. Interviewers agreed that the parental consent recruitment script and the respondent recruitment script provided within the CAI worked well for recruiting the respondent for the re-interview. In the CAI interviewer debriefing questions, FIs reported that over 70 percent of parents of youth age 12-17 did not ask any questions about the second interview, 11 percent of parents asked about the content of the

second interview, and 7 percent asked why we wanted to do another interview. Regarding respondent's reactions to the re-interview, interviewers reported that over 80 percent of the respondents did not ask any questions about the re-interview, just over 6 percent asked about the content of the second interview, and 5 percent asked about the length of the second interview.

All interviewers agreed that the $50 incentive payment made respondents eager to complete the re-interview. In the CAI interviewer debriefing questions, interviewers reported that over 85 percent of the respondents made no comment about the $50 incentive for the second interview. Of those who did comment, 80 percent thought the amount was "about right." Interviewers noted in debriefing calls that the $50 incentive led respondents to schedule the re-interview as soon as possible, which was helpful considering the 5 to 15 day time window.

Generally, respondent's reactions to the re-interview were positive. Interviewers indicated in debriefing calls that many respondents mentioned that the interviews were the same or similar, but that respondents were not annoyed by completing the interview a second time. In the CAI interviewer debriefing questions, FIs reported that almost one-fourth of the respondents commented that they thought the initial and re-interviews were the same. However, of this group, less than 14 percent made any comment about a strategy for answering the re-interview questions.

The pre-test interviewer feedback confirmed that the procedures developed for the 2006 Reliability Study were feasible when used in the field. Interviewers agreed that they had no problems following the procedures as long as they read the screens verbatim and used the materials provided to them.

Response Rates. Table 3 provides the anticipated response rates for the initial interview (T1) and the re-interview (T2) by age group, and Table 4 shows the actual T1 and T2 response rates by same and different interviewer and by age group, gender, and race. These pre-test response rates, which are the combined rates from Phase I and Phase II, will be used to adjust sample selection for the full 2006 Reliability Study.

The anticipated T1 overall response rate from Table 3 is similar to the observed response rate in Table 4 (0.82 anticipated, 0.81 actual). This response rate is consistent with prior NSDUH data. While all of the T2 response rates are higher than anticipated, the only significant difference between the anticipated and observed response rates is that the observed overall T2 response rate (0.92) is significantly higher than the expected overall T2 response rate (0.86). Based on feedback received during pre-test interviewer debriefing sessions, the $50 incentive payment had a large impact on the response rate because it made respondents eager to complete the re-interview, especially respondents under 26. The same and different interviewer sub-studies had similar response rates, with overall rates of 0.92 and 0.90, respectively. Thus, there is no clear evidence to show that using a same or different interviewer for the re-interview will affect response rates. These high response rates may allow fewer persons to be selected for the 2006 Reliability Study.

Table 3. Anticipated Response Rates by Age Group

Age	T1*	T2
12-17	0.89	0.92
18-25	0.85	0.88
26+	0.76	0.78
Overall	0.82	0.86

* T1 response rates are based on actual experience in DC, MD, TX, and FL in the 2003 NSDUH

Table 4. Phases I and 2: Combined Unweighted Response Rates at T1 and T2 with T2 by Same or Different FI, by Demographics

	T1			T2								
	Total			Same FI			Different FI			Total		
Category	**Sel**	**Resp**	**Rate**	**Sel**	**Resp**	**Rate**	**Sel**	**Resp**	**Rate**	**Sel**	**Resp**	**Rate**
Age												
12-17	88	77	0.88	50	47	0.94	27	24	0.89	77	71	0.92
18-25	73	66	0.90	42	40	0.95	24	23	0.96	66	63	0.95

26+	125	88	0.70	67	60	0.90	21	18	0.86	88	78	0.89
Gender												
Male	153	125	0.82	82	76	0.93	43	38	0.88	125	114	0.91
Female	133	106	0.80	77	71	0.92	29	27	0.93	106	98	0.92
Race												
Hispanic	49	37	0.76	24	23	0.96	13	13	1.00	37	36	0.97
NonHisp Black	55	47	0.85	32	31	0.97	15	14	0.93	47	45	0.96
Other	182	147	0.81	103	93	0.90	44	38	0.86	147	131	0.89
Total	**286**	**231**	**0.81**	**159**	**147**	**0.92**	**72**	**65**	**0.90**	**231**	**212**	**0.92**

Measures of Agreement for Drug and Age of First Use Variables. Preliminary estimates for the reliability study have been obtained during the pre-test that was fielded in Quarters 1 and 2 of 2005. For key categorical measures, the table below shows substance prevalence rates at T1 and T2, percent agreement, a calculated value for Cohen's Kappa, and a 95% confidence interval for Kappa, as well as an indicator for a significant McNemar's Test of homogeneity of the marginal distributions. Note that only one substance, lifetime cigarette use, has a *p* value below 0.05 for McNemar's Test. A significant *p* value indicates that the marginal distributions of the 2 by 2 table are significantly different.

The highest Kappa values occur with lifetime use of the most prevalent substances, lifetime cigarette use (0.921) and lifetime alcohol use (0.910), and the most rare substance, lifetime use of cocaine (0.915). High percent agreement tends to correspond with high Kappa, with the exception of less prevalent measures such as past year cocaine use. The low prevalence rate and small sample size causes past year cocaine use to have a lower Kappa even though past year cocaine had the highest percent agreement ($\kappa = 0.745$, percent agreement = 0.99). In Table 5, all reported Kappas are greater than 0.800 except past year cocaine use, which is greater than 0.700, indicating a high level of agreement among these measures.

Table 5. Prevalence Rates, Percent Agreement, Kappa, and 95% Confidence Bounds for Kappa for Key Measures (n=212)

	Prevalence Rates								
	T1		T2						
Variable	No.	%	No.	%	No. Agree	% Agree	Kappa	95% Lower Bound for Kappa	95% Upper Bound for Kappa
Lifetime Use									
Cigarettes*	125	0.59	133	0.63	204	0.96	0.9209	0.8673	0.9745
Alcohol	161	0.76	160	0.75	205	0.97	0.9102	0.8449	0.9755
Marijuana	87	0.41	82	0.39	201	0.95	0.8918	0.8297	0.9540
Cocaine	27	0.13	27	0.13	208	0.98	0.9151	0.8329	0.9973
Illicit Drugs	103	0.49	96	0.45	199	0.94	0.8770	0.8124	0.9417
Past Year Use									
Cigarettes	57	0.27	63	0.3	200	0.94	0.8607	0.7844	0.9369
Alcohol	134	0.63	133	0.63	195	0.92	0.8281	0.7498	0.9064
Marijuana	38	0.18	35	0.17	205	0.97	0.8842	0.8002	0.9682
Cocaine	4	0.02	4	0.02	210	0.99	0.7452	0.4046	1.0000
Illicit Drugs	49	0.23	46	0.22	201	0.95	0.8508	0.7653	0.9363

* Indicates that McNemar's test was significant at the .05 level for this measure.

The percent agreement for the continuous age of first use variables were calculated in two ways: 1) responses must be identical to be considered in agreement, and 2) the age of first use may vary by up to 1 year before the T1 and T2 responses are considered different (for Index of Inconsistency (IOI), if T1 and T2 values were within one year, T2 values were overwritten to equal T1 values before calculating IOI). The percent agreement that allows a one year variation in responses is high with most values ranging between 70 and 80 percent. Overall, as shown in Table 6, the index of inconsistency is quite high with all values lower than 0.20 and, consequently, all values of reliability greater than 0.80.

Table 6. Percent Agreement, Index of Inconsistency (I) and Reliability (R) for Age of First Use Measures (n=212)

Age of First Use	n	Identical Match				May Vary by 1 Year			
		No. Agree	% Agree	I	R	No. Agree	% Agree	I	R

Cigarettes	124	74	59.7	0.109	0.891	100	80.6	0.101	0.899
Alcohol	153	87	56.9	0.169	0.831	117	76.5	0.160	0.840
Marijuana	77	43	55.8	0.127	0.873	64	83.1	0.116	0.884
Cocaine	24	12	50.0	0.041	0.959	16	66.7	0.039	0.961
Illicit Drugs	90	44	48.9	0.157	0.843	68	75.6	0.152	0.848

Measures of Agreement for Demographic Variables Table 7 gives the frequency of differences for the first seven demographic variables at T1 and T2. Since marital and employment status questions are asked of respondents aged 15 or older, differences in marital and employment status responses were only counted if the respondent's calculated age was greater than or equal to fifteen during both the T1 and T2 interviews. These demographic variables are consistent between the T1 and T2 interviews except the family income variable, which is the only variable that requests information about someone other than the respondent. Of the forty respondents who had a difference between T1 and T2 family income, nineteen (or 47.5 percent) were age 12-17. Of the sixteen respondents aged 18 or older who answered the income question at both T1 and T2, only three were placed in a T2 income greater than one bracket different from their T1 response.

Table 7. Frequency of Differences at T1 and T2 Among First Seven Demographic Variables (n=212)

Variable	Number Different	Percent
Date of Birth	2	0.9
Calculated Age	3	1.4
Gender	1	0.5
Marital Status	3	1.4
Education	16	7.5
Employment Status	15	7.1
5-Category Family Income*	40	18.9

* Twelve of these respondents are aged 12-17

Of the twenty-two race/ethnicity variables, only seven respondents selected different race choices in T1 than in T2 (3.3 percent). Of these seven respondents, six reported Hispanic origin at both T1 and T2. Overall, the race and other demographic variables tend to have a high level of consistency across T1 and T2.

Overall Consistency. As a measure of overall consistency, Table 8 shows the number respondents with zero to six variables different between T1 and T2 out of forty-four variables compared. Of these forty-four variables, twenty-nine were the demographic variables listed in the previous section (date of birth, calculated age, gender, marital status, education, employment status, family income, and twenty-two race variables), five were the age of first use variables of cigarettes, alcohol, marijuana, cocaine, and illicit drugs, and ten were substance use variables (lifetime and past year use of cigarettes, alcohol, marijuana, cocaine, and illicit drugs). Since marital and employment status questions are asked of respondents aged 15 or older, differences in marital and employment status responses were only counted if the respondent's calculated age was greater than or equal to fifteen during both the T1 and T2 interviews. For the substance use variables, a difference in past year use was counted as a difference only if the lifetime use measure was the same in T1 and T2. For the age of first use variables, only respondents with both T1 and T2 nonmissing could be considered for a difference, and among those respondents, only an age of first use difference of two or more years between T1 and T2 was considered a difference. Given these conditions, none of the 212 T1 and T2 respondents had more than five variables different.

Table 8. Frequency of Respondents with 0 to 6 or more Variables Different at T1 and T2 out of 44 Variables Compared

No. of Vars Different	Frequency of Respondents	Percent of Respondents	Cumulative Frequency	Cumulative Percent
0	76	35.9	76	35.9
1	58	27.4	134	63.2
2	37	17.5	171	80.7
3	26	12.3	197	92.9
4	14	6.6	211	99.5
5	1	0.5	212	100.0
6+	0	0.0	212	100.0

Conclusions

In conclusion, the reliability study pre-test achieved higher than expected re-interview response rates, successfully completed re-interviews within the 5 to 15 day window, displayed a high level of consistency in responses to drug and demographic questions between T1 and T2, received a positive response from respondents, and demonstrated that field interviewers will be able to follow the procedures and protocols in the 2006 reliability study.

Since the pre-test results showed that the instrumentation and procedures developed for the 2006 reliability study were feasible when used in the field, the only major difference between the pre-test and the field test will be the sample design. Because the pre-test was a stand alone sample and the 2006 field test will be embedded within the main study, the sampling algorithm for the field test will be fairly different. In the pre-test, only one person was selected per household; in the field test, either 0, 1, or 2 persons will be selected, but only one-person selections will be eligible for the reliability study. Because the field test will be spread out over four calendar quarters, contamination to the main study will be minimized and the sampling algorithm can be adjusted as needed to obtain the desired 3,100 interviews.

References

Chromy, Eyerman, Odom , McNeeley & Hughes (2005). Association between interviewer experience and substance use prevalence rates in NSDUH. In J. Kennet & J. Gfroerer (Eds.) *Evaluating and improving methods used in the National Survey on Drug Use and Health* (DHHS Publication No. SMA 05-4044, Methodology Series M-5). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Hughes, Chromy, Giacoletti & Odom (2001, August). Impact of interviewer experience on drug use prevalence rates in the 1999 NHSDA. In Proceedings of the American Statistical Association [CD-ROM]. Alexandria, VA: American Statistical Association.

Hughes, Chromy, Giacoletti & Odom (2002). Impact of interviewer experience on respondent reports of substance use. In J. Gfroerer, J. Eyerman & J. Chromy (Eds.) *Redesigning an ongoing national household survey: Methodological issues* (DHHS Publication No. SMA 03-3768). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.

Kennet, Gfroerer, Bowman, Martin & Cunningham (2005). Introduction of an incentive and its effects on response rates and costs in NSDUH. In J. Kennet & J. Gfroerer (Eds.) *Evaluating and improving methods used in the National Survey on Drug Use and Health* (DHHS Publication No. SMA 05-4044, Methodology Series M-5). Rockville, MD: Substance Abuse and Mental Health Services Administration, Office of Applied Studies.