A SYSTEM FOR DETECTING INTERVIEWER FALSIFICATION

Joe Murphy, Rodney Baxter, Joe Eyerman, David Cunningham (RTI International)
and Joel Kennet (Substance Abuse and Mental Health Services Administration)

Presented at the American Association for Public Opinion Research 59th Annual Conference

Introduction

Interviewer falsification of survey responses introduces bias when falsified responses do not match values that would have been provided by eligible respondents (Schraepler and Wagner, 2003). The greater the number of falsified cases and the greater the difference between the true sample values and the falsified values, the greater the bias introduced into the survey statistics. In the National Survey on Drug Use and Health (NSDUH), interviewer falsification is a rare event. Instances of falsification have been detected in time to minimize the impact of the cost and quality of the survey data, but it remains a threat to data quality.

The increasing prevalence of monetary incentives in surveys may be magnifying the potential for frequent falsification. This is because the use of incentives may encourage falsification among unscrupulous interviewers. In general, interviewers are committed to data quality and honest reporting. However, a small subset of interviewers may be willing to falsify cases, but are deterred by the costs and risks of being caught, as established by the project verification and monitoring protocols. For these interviewers, an incentive payment may increase the value of falsification beyond the costs and risks established as the deterrent, if the interviewer keeps the incentive payment. Unless the risk associated with fraud is increased, unscrupulous interviewers may be encouraged to falsify cases when an incentive is used. Although there was no direct evidence of an increase in falsification after adoption of the incentive, this may be an issue with the NSDUH, which initiated an incentive payment of $30 in 2002.

Two general classes of falsification are item and unit. Item falsification occurs when an interviewer completes individual items on the survey without input from the respondents. This is generally not a problem on audio computer-assisted self-interview (ACASI) surveys like the NSDUH because the interviewer does not have contact with the instrument during the interview. Unit falsification, or curbstoning, occurs when an interviewer falsifies the entire survey. There are several types of curbstoning that can occur on surveys like the NSDUH:

Random infrequent curbstoning occurs when an interviewer randomly selects cases for falsification. An interviewer may engage in this type of falsification in order to increase productivity statistics or reduce workload. The impact on the bias is expected to be small because of the small number of cases being falsified. This type of falsification is very difficult to detect due to its random nature and the small number of cases affected. It can be detected through verification methods but is difficult to capture through analysis of response data and metadata, such as record of calls and timing data.

Systematic infrequent curbstoning occurs when an interviewer systemically selects cases for falsification based on the characteristics of the dwelling unit. An interviewer may engage in this type of falsification in order to bypass difficult to complete cases. This can include difficult respondents, dwelling units (DUs) in dangerous neighborhoods, DUs with barriers to entry, etc. Although this type of falsification affects only a small number of cases, it can have a large impact on bias for estimates of sub-groups or the overall bias of the survey if the falsified groups are heavily weighted. This type of falsification is difficult to detect, but can be identified through verification methods and through analysis of response data and metadata.

Frequent curbstoning occurs when an interviewer falsifies most or all of his or her assigned cases. An interviewer may engage in this type of behavior in order to receive income (wages, incentive payments, etc.) without completing the work. This type of falsification can have a significant impact on the survey statistics if the interviewer's workload represents a large share of the total surveys completed. Furthermore, this can lead to problems with small area estimates if interviewers are assigned cases based on geographic regions (as with the NSDUH). This type of falsification can be detected through verification methods and through analysis of response data and metadata.

In most cases, verification methods will reveal a frequent curbstoner. However, full verification can be very expensive and time consuming, as it requires staff to contact respondents after the data are collected. Sampling is often used in verification to reduce costs, unfortunately, sampling also increases the time required to detect a falsifier. Response data and metadata analysis can be used to supplement standard verification procedures to reduce detection time and costs. This can be achieved by identifying suspicious patterns in the data (response times, response patterns, calling patterns, etc.), and selecting suspected interviewers for full verification of all cases. Eventually, this process should allow for a reduction in the size of the sample of cases automatically assigned for verification, and detect fraudulent interviewers more quickly.

This paper provides an overview of the system for detecting interviewer falsification on the NSDUH, including recent enhancements that incorporate the review of response data and timing data in the process. This paper also describes the process of creating these measures based on data from known falsifiers. Finally, additional measures for future inclusion are discussed.

Background

The National Survey on Drug Use and Health (NSDUH) is an annual cross-sectional face-to-face household survey that gathers data on substance use and abuse among the non-institutionalized civilian population of the United States aged 12 and older. Data collection is conducted by RTI International for the Substance Abuse and Mental Health Services Administration (SAMHSA). Approximately 150,000

households and 67,500 persons are selected for the survey each year. Interviewers screen each household and then conduct a short computer-assisted personal interview (CAPI) with selected respondents in each respondent's household. The main portion of the survey data are captured through audio computer-assisted self-interviewing (ACASI).

As part of the NSDUH's Data Quality Monitoring System, field interviewers collect telephone numbers from selected households. These telephone numbers are used to call individuals and check on the quality of the interviewers' work. Verification is conducted on the first two non-interview cases and the first two interview cases completed by each interviewer in each calendar quarter. In addition, at least five percent of each interviewer's completed screenings and at least fifteen percent of each interviewer's completed interviews are selected for telephone verification. Trained professional telephone interviewers call to verify that the screening or interview occurred and was conducted in the correct manner. Those cases selected for telephone verification that do not have a telephone number are verified by mail.

If there is any suspicion about the performance of an interviewer, a greater proportion of his or her cases (up to 100 percent) can be "forced" into verification. Some data quality trends that may lead to increased verification include:

- Missing or refused verification telephone number for 30 percent or more of an interviewer's screenings and interviews completed
- Three or more errors of the same type for three different cases (such as "Roster Incorrect"), possibly indicating an emerging trend
- An interviewer report of his or her own or another staff member's phone number for verification
- Unexpected duplicate use of any particular telephone number
- Any problem that represents a serious protocol violation, such as an interview completed over the telephone

The results of the telephone and mail verification processes are summarized in weekly reports on the project web site and are reviewed by Data Quality and Field Management staff. Any indications of problematic behavior are investigated further and can lead to a closer review of other data quality measures.

In addition, The NSDUH collects timestamp data on each interview completed. These data are transmitted from the interviewers' computers nightly along with completed interview data. Any cases completed in less than 30 minutes or in more than 60 minutes automatically appear on a weekly Interview Length Report. This report lists the name of the interviewer completing the case, the case ID, the length of the overall interview, the length of the audio computer-assisted self-interviewing portion (ACASI) of the interview, the length of the FI-administered portion of the interview, and any comments entered by the interviewer in the debriefing portion at the end of the interview. Data Quality and Field Management staff examine these timing data in the context of any explanatory comments entered by the interviewer in the debriefing section of the interview, and ask the interviewer to provide an explanation for the unusual timing. If short interviews cannot be satisfactorily explained, other data quality measures are examined. If a problematic pattern emerges, the interviewer's cases may be forced into verification and examined for shortcutting or fraudulent behavior.

Several reports summarizing interviewer data quality performance are run and also posted on the project web site. Data Quality and Field Management staff have access to these reports and regularly review and analyze the information so corrective actions can be taken as necessary. The primary reports include:

Data Quality Summary Report: a count of verification data quality errors broken down by type of error

Data Quality Report: a list of questionnaire administration problems detected by the data editing staff, as well as problems with verification contact information or procedural errors for the interview

Record of Calls (ROC) Time Discrepancy Report: a list of any ROC event which was entered more than an hour off of "real screening device time" (e.g., an interviewer manually enters an ROC time that is at least one hour different from actual time)

Case Data Information Reports: a list of cases with a significantly lower than average amount of time between screenings or between interviews completed based on ROC data entered

When there are serious concerns about the validity of an interviewer's work, a mixed sample of the interviewer's completed cases is selected for in-person field verification. The sample generally includes a variety of screening and interview cases completed by the interviewer. The number and type of cases selected for field verification differs depending on the nature of the concerns. For example, if there is only concern about the interviewer's work in one segment, only cases in that segment may be field verified. In general, each field verification trip will involve verifying at least ten cases and verifying all the eligible screenings and interviews that are in close proximity to one another. If falsification is detected among this initial set of field verified cases, all cases completed by that interviewer that had not been previously verified are then field verified. All cases determined to be falsified are re-worked by another interviewer (usually the interviewer doing the field verification). Those found guilty of falsification are terminated from employment immediately.

In late 2002, it was found that four interviewers had recorded their own or another interviewer's telephone number for verification, a violation of project protocol. In order to investigate the legitimacy of the interviews they conducted, in-person field verifications took place. Through this process it was determined that all four interviewers had committed frequent curbstoning. Of the total of 760 screenings and 464 interviews completed by the four interviewers, 287 of the screening cases (38 percent) and 134 of the interview cases (29 percent) were deemed valid. A total of 473 screenings and 330 interviews were determined to have been falsified and were subsequently reworked.

The interview response and question-level timing data from the falsified cases were examined and compared to data from valid interviews completed in those states. Some significant differences between the falsified and valid cases were detected and this led to the development of a new component to the falsification detection system: the regular review of interview response and question-level timing data.

Methods

Response Deviation Several studies have shown that the frequent review of interview response data can identify potential interviewer falsification (Inciardi, 1981; Kaplowitz & Schlapentokh, 1982). By analyzing interviewer-level response distributions in the 1997-98 Baltimore STD and Behavior Survey (BSBS), Turner et al. (2000) identified differences in answers to questions about sexual and other sensitive behaviors between valid and falsified interviews. They concluded that the distributions of falsified responses varied significantly from valid response distributions because it is difficult for an interviewer to anticipate and replicate the pattern of responses given by the population at large.

Through the analysis of data from known falsifying interviewers in the 2002 NSDUH, an enhanced system for falsification detection was established. The data were well-suited to such an analysis because interviewers can conduct up to several hundred interviews a year, making interviewer-level estimates relatively robust and significant differences easier to detect. Also, much of the NSDUH interview is conducted via ACASI, meaning that interviewers do not have access to the response data and may therefore have trouble replicating "realistic" responses when falsifying.

Table 1 shows the response distributions on several key measures of lifetime substance use for three of the falsifiers in 2002. These three interviewers all worked in the same state and the distributions from their falsified cases are compared to all valid cases from that state in the same year. The response distributions reflect data captured in the ACASI from questions asking whether the respondent has ever used the particular substance (yes or no).

Table 1. Unweighted Rates of Lifetime Substance Use from Valid and Falsified Interviews in State with 3 Falsifiers: 2002

Interviewer / Cases	Number of Interviews	Lifetime Use (%)				
		Cigarettes	Alcohol	Marijuana	Cocaine	Heroin
All valid cases	1,188	56.9	71.3	45.1	15.8	1.8
Falsifier 1's fraudulent cases	92	26.1 [a]	59.8 [a]	35.8	20.6	1.1
Falsifier 2's fraudulent cases	119	48.7	50.4 [a]	24.4 [a]	7.6 [a]	4.2 [a]
Falsifier 3's fraudulent cases	77	29.9 [a]	37.6 [a]	22.1 [a]	1.3 [a]	0.0

[a] rate significantly different from valid case rate at p<.05 level

In most cases, the falsified rates were significantly lower than the valid responses. This suggests that interviewer-level response distributions that are significantly different than what is expected could be indicative of potential falsification. However, in the course of data collection, it is not immediately known which cases are falsified and which are not. Table 2 shows the response distributions for the falsifying interviewers and all others in their state for 2002, assuming (or pretending) that falsification had not yet been detected. All cases (both falsified and valid) are included for the falsifying interviewers. To determine significant differences, each interviewer's responses are compared to those from all other interviewers in the state. Data from interviewers who completed a very small number of interviews are excluded because of insufficient sample size.

Table 2. Interviewer-Level Comparison of Lifetime Substance Use Rates

Interviewer	Number of Interviews	Lifetime Use (%)				
		Cigarettes	Alcohol	Marijuana	Cocaine	Heroin
Falsifier 1	104	28.9 [a]	62.5 [a]	37.5	20.2	1.9
(all others)	*858*	*60.3*	*72.4*	*46.0*	*15.3*	*1.8*
Falsifier 2	154	49.4 [a]	52.6 [a]	28.6 [a]	9.7 [a]	3.3
(all others)	*808*	*58.3*	*74.9*	*48.3*	*17.0*	*1.5*
Falsifier 3	126	45.2 [a]	55.6 [a]	32.5 [a]	7.1 [a]	0.0
(all others)	*836*	*58.6*	*73.7*	*47.0*	*17.1*	*2.0*
Non-falsifier 1	86	65.1	77.9	50.0	23.3 [a]	1.1
(all others)	*876*	*56.1*	*70.7*	*44.6*	*15.1*	*1.8*
Non-falsifier 2	62	69.4 [a]	75.8	53.2	14.5	0.0
(all others)	*900*	*56.0*	*71.0*	*44.6*	*15.9*	*1.9*
Non-falsifier 3	101	70.3 [a]	81.2 [a]	61.4 [a]	20.8	5.0 [a]
(all others)	*861*	*55.3*	*70.2*	*43.2*	*15.2*	*1.4*
Non-falsifier 4	77	63.6	77.9	55.8 [a]	19.5	0.5
(all others)	*885*	*56.3*	*70.7*	*44.2*	*15.5*	*1.3*
Non-falsifier 5	53	66.0	81.1	34.0	17.0	1.9
(all others)	*909*	*56.3*	*70.7*	*45.8*	*15.7*	*1.8*
Non-falsifier 6	149	68.5 [a]	86.6 [a]	59.1 [a]	18.1	1.3
(all others)	*813*	*54.7*	*68.5*	*42.6*	*15.4*	*1.9*
Non-falsifier 7	38	57.9	86.8 [a]	47.4	13.2	1.8
(all others)	*924*	*56.8*	*70.7*	*45.0*	*15.9*	*2.8*

[a] significantly different from all other interviewers at p<.05 level

While the falsifying interviewers had more significant differences on average, their falsified data were "watered down" by their valid interviews, making it difficult to identify them as potential falsifiers. Another disadvantage to this type of comparison for falsification detection is that it does not account for differences in interviewers' caseloads. Interviewers working in different areas of a state may be assigned cases that vary from those of other interviewers on respondent demographics such as gender, age, and ethnicity, all of which are known correlates of substance use (SAMHSA, 2003). Because each interviewer works in a different area with different demographic characteristics, it is not accurate to assume that their data will all reflect the same estimates. Another problem with using the distributions in Table 2 for falsification detection is the fact that interviewers might have had knowledge of the national estimates for lifetime substance use and were thus able to replicate them in their falsified interviews relatively well. Interviewers are provided with a summary of findings report from the previous year's survey for use in gaining respondent interest and cooperation. Falsifiers may use these reports to obtain information on what valid responses would look like.

Although these falsifying interviewers were "successful" in approximating likely response distributions compared to national totals, it was hypothesized that they did not accurately predict the distributions by the significant correlates of age, gender, and ethnicity. By stratifying the comparison by respondent demographics, differences in interviewers' caseloads could also be better controlled for. The data show that the falsifiers' response distributions did not match the expected values when compared by demographics. An example from one of the falsifiers is found with the response distribution for marijuana use for respondents aged 12 to 17 compared to the rate for respondents 18 and older. The rate reflected in the falsifier's interview data for respondents aged 12 to 17 (valid and falsified combined) was 26.1 percent. This is not significantly different from the rate for 12 to 17 year-olds from all other interviewers in the state (29.1 percent). However, the falsifier's rate for respondents 18 and older was significantly lower at 29.0 percent than the rate from all other interviewers (57.0 percent). Another example is the rate of lifetime alcohol use for another falsifier by Hispanicity. This interviewer's rate for Hispanics was not significantly different than the rate from all other interviewers in the state (53.5 percent and 69.4 percent, respectively). But the falsifier's rate for Non-Hispanics was significantly lower than the rate from all others (57.3 percent and 80.1 percent, respectively). It appeared that analyzing interviewer-level response distributions on the lifetime use variables could accurately and efficiently predict potential falsification for these cases.

To test this hypothesis, response data were analyzed for all interviewers – those who falsified and those who did not – from the 2002 survey. It was expected that response distributions for the falsifying interviewers would be significantly different from others in their state when stratified by age, gender, and ethnicity. Even though it was known in advance which cases were falsified, *all* cases worked by the falsifying interviewers (valid and falsified) were included in the model. This was done to simulate the appropriateness of the method in the situation where falsification would not be known in advance.

A "response deviation score" was computed for each interviewer from comparisons at each level of stratification. We assigned one "point" to an interviewer if their rate on a substance was significantly different than the average for all other cases in the state at the $p<.05$ level. We assigned two points if their rate on a substance was significantly different than the average for the state at the $p<.01$ level. The sum of all points for each interviewer across all substances represents the deviation score. Higher scores indicate a consistent pattern of significantly different rates of substance use for the cases generated by that interviewer, with 0 as the minimum and 60 as the maximum. The equation used to calculate the response deviation score follows:

Response deviation score =

$$
\begin{aligned}
&(r_{cg,m} + r_{al,m} + r_{mj,m} + r_{cc,m} + r_{he,m}) + \\
&(r_{cg,f} + r_{al,f} + r_{mj,f} + r_{cc,f} + r_{he,f}) + \\
&(r_{cg,y} + r_{al,y} + r_{mj,y} + r_{cc,y} + r_{he,y}) + \\
&(r_{cg,o} + r_{al,o} + r_{mj,o} + r_{cc,o} + r_{he,o}) + \\
&(r_{cg,h} + r_{al,h} + r_{mj,h} + r_{cc,h} + r_{he,h}) + \\
&(r_{cg,n} + r_{al,n} + r_{mj,n} + r_{cc,n} + r_{he,n})
\end{aligned}
$$

where: r = 0 if difference between interviewer and all other interviewers was insignificant,1 if difference was significant at $.01<=p<.05$, and 2 if difference was significant at $p<.01$,

cg = cigarettes,
al = alcohol,
mj = marijuana,
cc = cocaine,
he = heroin,
m = male respondents,
f = female respondents,
y = respondents age 12 to 17,
o = respondents age 18 or older,
h = Hispanic respondents,
n = non-Hispanic respondents

Table 3 presents the response deviation scores for the interviewers in the state with three falsifiers in 2002.

Table 3. Response Deviation Scores for Interviewers in the State with Three Falsifiers: 2002

Interviewer	Response Deviation Score
Falsifier 1	20
Falsifier 2	23
Falsifier 3	25
Non-falsifier 1	8
Non-falsifier 2	2
Non-falsifier 3	21
Non-falsifier 4	5
Non-falsifier 5	2
Non-falsifier 6	18
Non-falsifier 7	2

Since the scores for the falsifiers were among the highest (three of the top four), the response deviation score accurately predicted which interviewers had committed falsification in this case. Based on the success of this test, the response deviation score was implemented in the enhanced falsification detection system beginning in 2004. A discussion of the use of this score in the verification process is described later in this paper.

Rare Response Combinations

Another hypothesis about these falsifiers was that they did not realize rare combinations of responses in the NSDUH and may have included a higher-than-average number of these rare combinations in their falsified cases. For instance, it is rare that a respondent will report that he or she has used marijuana, cocaine, or heroin and report having never tried cigarettes or alcohol. This is consistent with the general notion that use of cigarettes or alcohol usually precedes the use of illicit drugs like marijuana, cocaine and heroin in one's life.

The response distributions in the state with three falsifiers in 2002 were analyzed and a "rare combination score" was computed for each interviewer. The rare combination score is computed by assigning one "point" to each interviewer completing an interview with the following response pattern:

- ever used marijuana, cocaine, or heroin, and
- never used cigarettes, and
- never used alcohol

Table 4 presents the rare combination scores for interviewers in the state with three falsifiers in 2002. The percent of cases with a rare response combination for each interviewer is also presented.

Table 4. Rare Response Combination Scores for Interviewers in State with Three Falsifiers: 2002

Interviewer	Rare Combination Score	Percent of Interviews with a Rare Response Combination
Falsifier 1	2	1.9
Falsifier 2	4	2.6
Falsifier 3	4	3.2
Non-falsifier 1	0	0.0
Non-falsifier 2	0	0.0
Non-falsifier 3	2	2.0
Non-falsifier 4	1	1.3
Non-falsifier 5	0	0.0
Non-falsifier 6	0	0.0
Non-falsifier 7	0	0.0

Because the rare combination scores and percentages were highest for the falsifiers (three of the top four), it is believed that this score could be useful in predicting future falsification. The rare combination score was also added to the falsification detection system in 2004 and a discussion of the use of this score in the verification process is described later in this paper.

Timing and Item Nonresponse Data

As described earlier in this paper, the NSDUH data quality procedures include a review of interview lengths. Interviews that are excessively short or long compared to expectation are subject to increased verification efforts to determine whether any protocol violations were committed by the interviewer. This technique can be useful in the detection of potential falsification when the falsifying interviewer spends less time (shortcutting) or more time completing the interview than the typical respondent. However, the interview length measure may be too broad to capture behaviors that could indicate falsification. It was hypothesized that interviewer-level average timings for falsifiers will be different, and probably shorter, than timings for non-falsifying interviewers overall, at the questionnaire module level, and on specific interview items that require interaction with the respondent or out-loud reading of text. It was also hypothesized that interviewers would make less use of the Don't Know and Refuse response options in order to make their data appear more "valid."

To test these hypotheses, module and item-level timing data from 2002 were compiled for the falsified and valid interviews from the questionnaire audit trail. Table 5 presents a summary of significance tests on the module and item times and seconds per question rate for each module and item. The module names indicate the content of each module and a description of the items follows:

CALENDAR: Interviewer explains use of calendar dates
INTROACA: Interviewer gives instructions on ACASI and turns over computer to respondent
ENDAUDIO: ACASI section ends and respondent turns the computer back to interviewer
VERIFID: Interviewer enters the verification identification number to continue
CASEID: Interviewer enters the case identification number
TOALLR3I: Interviewer gives form to respondent for household verification information
INCENT01: Interviewer gives incentive receipt form to respondent
FIDBFINTR: FI starts debriefing section

In a manner similar to the response deviation score, 1 "point" is assigned if the p-value is between .01 and .001. Two points are assigned if the p-value is less than .001. The score is then assigned a positive (longer) or negative (shorter) value depending on the direction of the difference.

Table 5. Summary of Significant T-tests for Differences Between Module Times, and Seconds per Question Rates

	Falsifier 1				Falsifier 2				Falsifier 3			
	Total time		Seconds/Q		Total time		Seconds/Q		Total time		Seconds/Q	
Module / Item	+	-	+	-	+	-	+	-	+	-	+	-
Total interview	2		2		2		2			-2		
ACASI Total	1		2				2					
FI administered	2		2		2		2			-2		-2
ACASI Core	2		2				2					
ACASI Non-Core												
Demographics Part 1	2		2			-2				-2		-2
ACASI Tutorial	2		2							-2		-2
Tobacco Use											1	
Alcohol Use												
Marijuana Use			2							-2		-2
Cocaine Use	1		2									-2
Crack Use												
Heroin Use	1		1									
Hallucinogen Use	1		2									
Inhalant Use	2		2		2		2					
Pain Reliever Use												
Tranquilizer Use	2		2									
Stimulant Use										-2		-2
Sedative Use												
Prescription Drugs	1		1				1					
Risk Perceptions												
Social Environment												
Demographics Part 2					2		2					
FI Debriefing	2		2		1					-2	--	--
CALENDAR		-2	--	--			--	--		-2	--	--
INTROACA		-2	--	--		-2	--	--		-2	--	--
ENDAUDIO	2		--	--	2		--	--			--	--
VERIFID	2		--	--	2		--	--			--	--
CASEID	2		--	--	2		--	--	2		--	--
TOALLR3I		-2	--	--		-2	--	--		-2	--	--
INCENT01	2		--	--	2		--	--		-2	--	--
FIDBFINT			--	--			--	--			--	--
Total Score	**24**		**26**		**11**		**13**		**-20**		**-11**	

Based on these scores, it appeared that only Falsifier 3 committed shortcutting. The cases for this interviewer showed lower module and item timings and lower seconds per question rates than all other interviewers in the state. For the other two interviewers, the results are quite different. Each have high positive scores indicating that they tended to take more time in a module, and spend more time on each question compared to other interviewers in the state. Based on these results, it was determined that both short and long module and item timings could be indicative of falsification.

To determine whether falsifiers were less likely to use the Don't Know and Refuse response options than non-falsifiers, similar tallies were run for the ACASI and interviewer-administered sections. Again, 1 "point" is assigned if the p-value is between .01 and .001. Two points are assigned if the p-value is less than .001. The score is then assigned a positive value (more Don't Know/Refuse responses) or negative value (fewer Don't Know/Refuse responses).

Falsifier 1 used the Don't Know and Refuse keys more often than the other interviewers. Falsifier 2 showed no difference compared to the other interviewers in the use of the Don't Know and Refuse keys.

Falsifier 3 used the Don't Know and Refuse keys less often than the other interviewers. Based on the lack of a consistent pattern, this rate has not been included in the enhanced system for falsification detection. Further analysis on these data or data gathered from other falsifiers may shed new light on the applicability of the rate of Don't Know and Refuse responses to falsification detection.

Variance in Module Times

Inciardi (1981) suggests that differences between valid and falsified means may not always be detected, but that the *variance* of falsified data may be lower than for valid data. To examine this using the NSDUH data, the variance of module times for falsifiers and non-falsifiers were compared. It was expected that there would be more variation in the ACASI section of the instrument for non-falsifiers due to the fact that respondents have no prior experience with the instrument before the interviewer. It

was also expected that there would be more variation due to questions and sections being more or less difficult to answer among non-falsifiers. Given the evidence that some falsifiers may actually take longer to answer the instrument (even while controlling for number of items answered), they may be less likely to show the same variation across modules that real respondents produce.

Table 6 summarizes the differences in variances between the falsifiers and non-falsifiers for the 24 survey module seconds per question measures. The values in the Higher Variance column show the number of significant differences in variance where the falsifier had a higher variance on a module compared to other interviewers in the state and values in the Lower Variance column show the number of significant differences in variance where the falsifier had a lower variance on a module than the other interviewers in the state. While there are many differences in variance, most of the differences are not in the direction expected. Falsifiers 2 and 3 had higher variances on the whole than the other interviewers and Falsifier 1 had the same number of higher and lower variances. Based on these results, there was not sufficient evidence of variance differences between falsifiers and non-falsifiers to inform the falsification detection system.

Table 6. Summary of Differences in Variances in Module Seconds per Question

Interviewer	Significant Differences	Higher Variance	Lower Variance
Falsifier 1	20	10	10
Falsifier 2	21	17	4
Falsifier 3	20	13	7

Implementation

Based on the results of the analyses in this paper, a number of items have been added to the NSDUH Data Quality Monitoring System for review by Data Quality and Field Management staff. Interviewers that meet minimal criteria on these indicators are placed on increased verification:

Response Deviation Score: an indicator of the lifetime substance use prevalence rates reflected in each interviewer's cases compared to all other interviewers. Verification is increased if an interviewer has a response deviation score at least five times higher than the average response deviation score.

Rare Response Combinations: an indicator of cases where a respondent reports ever using marijuana, cocaine, or heroin, but never having used cigarettes or alcohol. Verification is increased if an interviewer has at least two instances of rare response combinations, which represent at least 5% of the work completed by the interviewer in the current calendar quarter

Total Interview Seconds per Question (Shorter than Average Times and Longer than Average Times): an indicator of the difference between an interviewers average seconds per question and the average across all interviewers. Verification is increased if an interviewer's average seconds per question in the overall interview is at least two standard deviations below or two standard deviations above the overall average seconds per question.

ACASI Seconds per Question (Shorter than Average Times and Longer than Average Times): an indicator of the difference between an interviewer's average seconds per question for the ACASI section and the average across all interviewers. Verification is increased if an interviewer's average seconds per question in the ACASI portion of the interview is at least two standard deviations below or two standard deviations above the national average seconds per question.

Interviewer-administered Seconds per Question (Shorter than Average Times and Longer than Average Times): an indicator of the difference between an interviewer's average seconds per question for interviewer-administered questions and the average across all interviewers. Verification is increased if an interviewer's average seconds per question in the FI-administered portion of the interview is at least two standard deviations below or two standard deviations above the national average seconds per question**.**

In addition to increased verification, other data quality measures are examined for interviewers who meet any of the criteria above, including the measures listed earlier in this paper. When serious concerns about the validity of an interviewer's work emerge from these examinations, an in-person field verification of the interviewer's work is completed.

Conclusion and Future Directions

The analyses in this paper show that falsification detection can be improved through the systematic review of response data and metadata such as module and item timings. Through early detection and remediation, the threat of falsification to survey bias and increased costs can be reduced. While the NSDUH system has improved based on these recent enhancements, there are still many more enhancements that could be incorporated. In particular, the following techniques will be assessed for possible adoption:

Analysis of screening data: Interview data analyses could be replicated using roster data for screenings.

Analysis of record of calls data and other metadata: It can be determined whether interviewers tend to falsify pending unable-to-contact or refusal status cases near the end of the data collection period.

Data mining approach: It may be possible to predict falsified interviews based on combinations of multiple responses and metadata.

Statistical Process Control (SPC): The basic idea behind SPC is that all processes have variance associated with them (Reed & Reed, 1997). This variance can be divided into two main components, a random or uncontrollable component and an assignable or controllable component. The process is charted over time with two parallel graphs, one displaying the mean value for the process, the other chart showing the range. Both graphs are charted horizontally over time with upper control limits and lower control limits. When the points in the charts are inside the control limits the process is said to be in control, when it is outside the limits it is said to

be out of control. When the process is out of control, the process is investigated to determine from where the problems are coming. To apply this to NSDUH data, the seconds per question length of modules could be tracked. Interviewers that produce a high proportion of interviews that are outside of the control limits could be scheduled for a higher rate of field or telephone verifications. The problem could be a training issue and not actually falsification. If new interviewers had higher rates than more experienced interviewers, additional training or feedback could improve their rates. Another application could be to analyze a proportion such as percent of lifetime marijuana users. Here, if an interviewer is showing higher or lower rates compared to other interviewers, their work could be more closely tracked by field and telephone verifications.

Benford's Law: Benford's Law is a simple tool that can be used to help identify possibly fraudulent or error-prone data in sample surveys (Swanson, et al., 2003). The method identifies unique patterns in "real world" numbers that may not be evident to falsifiers. This method could be applied to screening or interviewing data.

Assessment of these techniques must be carried out keeping in mind that, while the costs of undetected falsification can be quite large, both in terms of biased data and production losses, the costs of detection and follow-up should not exceed what is practical. As noted in the introduction, the vast majority of professional field interviewers take their work seriously and would not consider the idea of falsifying data. To spend more resources on policing them than on providing them with the proper incentives for performing good, honest work is likely to be counterproductive.

References

Inciardi, J.A. (1981). Fictitious Data in Drug Abuse Research. *The International Journal of Addictions*, 16 (2), 377-380.

Kaplowitz, S.A., & Shlapentokh, V. (1982). Possible Falsification of Survey Data: An Analysis of a Mail Survey in the Soviet Union. Public Opinion Quarterly, 46(1):1-23.

Reed, S., & Reed JH (1997). The Use of Statistical Quality Control Charts in Monitoring Interviewers. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Schraepler, J.P. and Wagner, G.G. (2004). Characteristics and Impact of Faked Interviews in Surveys - An Analysis by Means of Genuine Fakes in the Raw Data of SOEP. DIW Research Note, Berlin (forthcoming).

Substance Abuse and Mental Health Services Administration (SAMHSA). (2003). Results from the 2002 National Survey on Drug Use and Health: National findings (Office of Applied Studies, NHSDA Series H-22, DHHS Publication No. SMA 03-3836). Rockville, MD.

Swanson, D., Cho, M.J., & Eltinge, J. (2003). Detecting Possibly Fraudulent or Error-Prone Survey Data Using Benford's Law. Proceedings of the American Statistical Association, Section on Survey Research Methods.

Turner, C.F., J.N. Gribble, A.A. Al-Tayyib, & J.R. Chromy (2000). "Falsification in Epidemiologic Surveys: Detection and Remediation." Technical Papers on Health and Behavior Measurement, No. 53. Washington, DC: Research Triangle Institute.