Methods Articles

Survey Conditioning in Self-Reported Mental Health Service Use: Randomized Comparison of Alternative Instrument Formats

Naihua Duan, Margarita Alegria, Glorisa Canino, Thomas G. McGuire, and David Takeuchi

Objective. To test the effect of survey conditioning (whether observed survey responses are affected by previous experience in the same survey or similar surveys) in a survey instrument used to assess mental health service use.

Data Sources. Primary data collected in the National Latino and Asian American Study, a cross-sectional household survey of Latinos and Asian Americans residing in the United States.

Study Design. Study participants are randomly assigned to a Traditional Instrument with an interleaved format placing service use questions after detailed questions on disorders, or a Modified Instrument with an ensemble format screening for service use near the beginning of the survey. We hypothesize the ensemble format to be less susceptible to survey conditioning than the interleaved format. We compare self-reported mental health services use measures (overall, aggregate categories, and specific categories) between recipients of the two instruments, using $2 \times 2 \chi^2$ tests and logistic regressions that control for key covariates.

Data Collection. In-person computer-assisted interviews, conducted in respondent's preferred language (English, Spanish, Mandarin Chinese, Tagalog, or Vietnamese).

Principal Findings. Higher service use rates are reported with the Modified Instrument than with the Traditional Instrument for all service use measures; odds ratios range from 1.41 to 3.10, all p -values $<.001$. Results are similar across ethnic groups and insensitive to model specification.

Conclusions. Survey conditioning biases downward reported mental health service use when the instrument follows an interleaved format. An ensemble format should be used when it is feasible for measures that are susceptible to survey conditioning.

Key Words. Attenuation, ensemble format, interleaved format, mental health, panel conditioning, psychiatric disorder, questionnaire, service use, survey instrument

Survey conditioning, when observed survey responses are affected by previous experience in the same or similar surveys, is often observed in survey research. A version of survey conditioning, known as *panel conditioning* in panel or longitudinal studies, occurs when previous exposure to a survey affects responses in later waves (Bailar 1975; Cantor 1989; Silberstein and Jacobs 1989; Pennell and Lepkowski 1992; Duan and Valdez 1999). Similar conditioning can also occur within a single survey, if the responses to survey items placed in the latter segment of a survey instrument are affected by experience gained from earlier segments of the survey (Kessler et al. 1994, 1998, 2000; Knowles et al. 1996; Vega et al. 1998; Lucas et al. 1999; Piacentini et al. 1999; Kessler and Merikangas 2004; Kessler and Ustun 2004). One form of survey conditioning is *attenuation*: the systematic reduction of symptom reports over time, either from one administration to the next, or within the same administration. In particular, respondents might learn to avoid the burden of follow-up questions by responding negatively to stem questions; accordingly, survey responses in the latter sections of a long instrument are biased towards underreporting.

Interleaved and Ensemble Formats for Instrument Design

Attenuation has been tested empirically by randomizing respondents to various versions of the same surveys with questions in different orders. Jensen, Watanabe, and Richters (1999) conducted such a randomized trial, and reported that psychiatric disorder modules placed near the front of a survey were more frequently endorsed; attenuation of symptom reports was observed with items placed in later sections of the survey. The instrument used in their study (the Diagnostic Interview Schedule [DIS]) was based on a set of stem questions for various psychiatric disorders, each followed by a series of branch

Address correspondence to Naihua Duan, Ph.D., UCLA Health Services Research Center, 10920 Wilshire Blvd., Suite 300, Los Angeles, CA 90024. Dr. Duan, is with the Health Services Research Center, Jane & Terry Semel Institute for Neuroscience and Human Behavior, University of California, Los Angeles, CA and also with the Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine, University of California, Los Angeles, CA and the Department of Biostatistics, School of Public Health, University of California, Los Angeles, CA. Margarita Alegria, Ph.D., is with the Center for Multicultural Mental Health Research, Department of Psychiatry, Cambridge Health Alliance, Cambridge, MA and with the Department of Psychiatry, Harvard Medical School, Boston, MA. Glorisa Canino, Ph.D., is with the Behavioral Sciences Research Institute, University of Puerto Rico, School of Medicine, San Juan, Puerto Rico. Thomas G. McGuire, Ph.D., is with the Department of Health Care Policy, Harvard Medical School, Boston, MA. David Takeuchi, Ph.D., is with the School of Social Work, University of Washington, Seattle, WA.

questions for measuring the specific disorder. If the respondent answered affirmatively to the stem question for a disorder, the corresponding branch questions were asked before asking the next stem question. Therefore the stem and branch questions are *interleaved*—we refer to this structure of the survey instrument as the *interleaved format*. A plausible interpretation of Jensen et al.'s finding is that respondents learned the stem-and-branch structure from their experience with the early segments of the survey, and avoided burdensome branch questions in the later segments of the survey by responding negatively to the stem questions.

Attenuation resulting from the interleaved format can be mitigated by placing the stem questions for all disorders near the beginning of the interview, before any branch questions are presented (Kessler et al. 1998, 2000; Kessler and Merikangas 2004; Kessler and Ustun 2004). We refer to this structure of the survey instrument as the *ensemble format*. As all stem questions are asked ahead of the branch questions under the ensemble format, respondents do not experience the branch questions until after all stem questions have been asked.

Kessler et al. (1998) conducted a randomized trial for the Composite International Diagnostic Instrument (CIDI), comparing the *ensemble format* with the *interleaved format* that had been the traditional way of asking about psychiatric disorders. For most disorders, the ensemble format resulted in higher prevalence rates than the interleaved format, suggesting that the latter was vulnerable to survey conditioning that lead to systematic attenuation of symptom reports.

Survey conditioning and attenuation might interfere with comparisons across studies that used different formats of an instrument. One of the first psychiatric epidemiology studies in the United States, the Epidemiologic Catchment Area (ECA) study, used the DIS that followed an interleaved format (Robins et al. 1981) to assess psychiatric disorders. Ten years later, the National Comorbidity Survey (NCS, <http://www.hcp.med.harvard.edu/ncs/>) adopted the ensemble format, and found higher rates of psychiatric disorders (Kessler et al. 1994, 1998, 2000). Because of the likelihood of survey conditioning, it is difficult to tell whether the higher rates reported in NCS are due to an increase in the prevalence of disorders over time, or due to differences in the format of the survey.

Collaborative Psychiatric Epidemiology Studies

Lucas et al. (1999) recommend using the ensemble format, with stem questions for disorders placed up front, to reduce the effect of survey conditioning. This

recommendation has been followed with respect to the placement of stem questions for psychiatric disorders in the National Comorbidity Survey Replication (NCS-R, Kessler and Merikangas 2004), the National Survey of American Life (NSAL, Jackson et al. 2004), and the National Latino and Asian American Study (NLAAS, Alegria et al. 2004), three parallel surveys of psychiatric epidemiology conducted recently, and known collectively as the Collaborative Psychiatric Epidemiology Studies (CPES; Colpe et al. 2004). In addition to assessing the prevalence of psychiatric disorders, the CPES also measure the rates of mental health service use. By assessing both disease and service use, the CPES can assess “unmet need” for psychiatric care, occurring when a respondent has psychiatric illness but does not receive services.

While the CPES apply the lessons from research on attenuation to the stem questions for psychiatric disorders, the stem questions for service use still follow the traditional interleaved format and are placed in the middle of the CPES survey, raising the possibility that service use might be underreported. If psychiatric disorders are reported accurately, but service use is underreported due to survey conditioning, estimates of “unmet need” from these surveys would be biased upward.

Randomized Comparison in NLAAS

We conducted a randomized trial, as part of the NLAAS, to assess the presence and magnitude of attenuation in the reported service use. Following Kessler et al. (1998), we randomized respondents to two versions of the survey instrument, allowing us to compare responses to service use questions presented in an interleaved format to responses to an ensemble format with the stem questions for service use placed near the beginning of the survey.

The nonexperimental version of the survey instrument used in the NLAAS is essentially the same as the instruments used in NCS-R and NSAL, with only minor modifications in some questions that do not affect the service use questions directly. While this version of the instrument follows the ensemble format for questions about disorders, questions about service use are interleaved. We refer to this version of the NLAAS instrument as the *Traditional Instrument*. We refer to the experimental version of the NLAAS instrument, with stem questions for service use moved up and arranged in an ensemble format, as the *Modified Instrument*. Both versions of the instrument follow the ensemble format for questions about disorders.

METHODS

The NLAAS

In 2002–2003, the NLAAS surveyed a nationally representative sample of Latinos and Asian Americans, aged 18 and above, residing in households in the United States. The main objectives of the NLAAS were to estimate the prevalence of psychiatric disorders and the rates of mental health service use; to investigate the relationship between social position, environmental context, and psychosocial factors and disease prevalence and service use; and to compare disease prevalence and service use with other race/ethnic groups in the CPES. The sample design and survey methods of the NLAAS have been described in detail elsewhere (Alegria et al. 2004; Heeringa 2004; Heeringa et al. 2004). Briefly, a four-stage area probability sample was implemented to sample (1) U.S. Metropolitan Statistical Areas (MSAs) and counties, (2) area segments, (3) housing units, and (4) respondents. The survey was conducted during May 2002–November 2003 in English, Spanish, Mandarin Chinese, Tagalog, and Vietnamese. The final sample of 4,649 respondents consists of 2,554 Latinos that include four major subethnic groups (Mexicans, Puerto Ricans, Cubans, and Other Latinos), and 2,095 Asian Americans that include four major subethnic groups (Chinese, Vietnamese, Filipinos, and Other Asians). The overall response rate was 75.5 percent for Latinos and 65.6 percent for Asians.

The NLAAS shares with the CPES common core sections, including the World Mental Health Survey Initiative version of the World Health Organization Composite International Diagnostic Interview (WMH-CIDI, Kessler and Ustun 2004), the Thirty-Day Functioning, the Service Use Battery, and sociodemographic variables. Although the studies include many common elements, they are not identical: each study also has study-specific questions in non-Core sections.

Traditional Instrument

The structure of the Traditional Instrument is described schematically in the first column of Figure 1. The instrument begins with the Symptoms Screener section comprised of a series of stem questions probing essential symptoms for various psychiatric disorders, such as major depression, dysthymia, panic disorder, social phobia, and generalized anxiety disorder. Those stem questions were asked in an ensemble format (shown in Figure 1 as “Symptoms Screener”). After the completion of the Symptoms Screener section, the Traditional Instrument asks a series of follow-up questions in the diagnostic

Figure 1: Schematic Diagram for NLAAS Instruments.

* Additional questions were only administered if respondent endorsed stem questions.

battery about each symptom for which the respondent screened positive. The in-depth questions are used to assess the presence of psychiatric disorders, both lifetime and for the last 12 months.

Upon completion of the diagnostic battery (which includes the screener questions and the diagnostic assessments for all diagnoses that the respondent screened positive), the Traditional Instrument asks about service use, in an interleaved format. Stem questions for services follow the complete diagnostic battery; furthermore, a positive response on a stem question for service use brings the respondent to a detailed assessment of the particular category of service, before presenting the stem question for the next category of service use.

The service use section begins with the stem question whether the respondent was ever admitted for an overnight stay in a hospital or other facility to receive help for problems with their emotions, nerves, mental health, or use of alcohol or drugs. If the respondent answers positively to this stem question, he/she is queried immediately about details of inpatient services. Next, the respondent is asked whether he/she ever used an Internet support group or chat room. If the respondent answers affirmatively to this stem question, he/she is queried immediately about details related to these services. The same approach is used for the next series of probes: whether the respondent used a self-help group; a hotline; and a psychological counseling or therapy that lasted 30 minutes or longer with any type of professionals. For each of these questions, if the respondent answers affirmatively, he/she is immediately asked a series of branch questions.

Next, the Traditional Instrument shifts to a partial ensemble format for categories of professional services. The respondent is asked ten stem questions whether he/she ever went to see any of the following professionals to receive help for problems with their emotions, nerves, mental health, or use of alcohol or drugs: (1) a psychiatrist; (2) a general practitioner or family doctor; (3) any other medical doctor, like cardiologist, gynecologist, or urologist; (4) a psychologist; (5) a social worker; (6) a counselor; (7) any other mental health professional, such as a psychotherapist or mental health nurse; (8) a nurse, occupational therapist, or other health professional; (9) a religious or spiritual advisor like a minister, priest, pastor, or rabbi; and (10) any other healer, like a herbalist, or doctor of oriental medicine, or chiropractor, or spiritualist. After the completion of the 10 stem questions, the respondent is presented with a series of branch questions for each provider category he/she reported ever using, including the age he/she was the first time he/she visited the provider, the last time he/she saw the provider, and the number of visits in the last 12 months.

Service use is vulnerable to survey conditioning and attenuation in the Traditional Instrument in at least two ways. First, the three groups of service use stem questions described above are interleaved rather than presented in the ensemble format, which can result in underreporting for the service types encountered later in the instrument. Partial use of the ensemble format for professional services may be of some help. However, these stem questions can still be attenuated by experience with earlier service use questions about inpatient services, Internet support group, etc. Second, all of the service use stem questions are placed after diagnostic assessments for psychiatric disorders; the experience gained from the diagnostic assessments provides further

opportunities for survey conditioning and attenuation. Irrespective of the ordering of the service use questions among themselves, the respondent might have already learned to attenuate his/her response from the experience gained earlier from the diagnostic batteries.

Modified Instrument

The Modified Instrument follows the recommendation in Lucas et al. (1999) to place all stem questions up front in an ensemble format. The Modified Instrument places all service use stem questions immediately after the symptom screener, as shown schematically in the third column of Figure 1. As the stem questions for service use are presented before the diagnostic batteries for psychiatric disorders, respondents are not exposed to the consequences of positive endorsement, minimizing the potential for survey conditioning and attenuation.

For both the Traditional and Modified Instrument, in addition to the sequential inquiries described above, additional service use inquiries are also given to respondents who fulfill criteria for the syndrome for a specific psychiatric disorder (meeting the symptom criteria, not necessarily meeting other diagnostic criteria, such as criteria related to time and severity). At the end of each diagnostic battery, these respondents are asked whether they have ever talked to a medical doctor or other professional provider for the symptoms of the assessed disorder. If the respondent answers affirmatively, he/she is given the list of professional providers to identify those he/she has talked to. A respondent who positively endorses using certain providers is then entered directly to the services battery.

Randomization Procedure

In designing a randomized trial to test for survey conditioning in service use measures, the main methodological challenge was to avoid unduly compromising the original purpose of the NLAAS—to assess rates of disorders and service use among Latino and Asian American populations and to compare those rates to the national sample (NCS-R) and to the African American sample (NSAL). Allocating part of a limited sample to the Modified Instrument to assess survey conditioning might reduce the comparability between the NLAAS and the other CPES surveys (NCS-R and NSAL). Based on pre-data collection precision analyses, we chose to allocate 25 percent of the sample to the Modified instrument leaving 75 percent of the sample with the Traditional Instrument. With 25 percent allocated to the Modified Instrument,

we estimated having sufficient power to detect meaningful differences in reported service use rates, but still maintain comparability with the CPES sister studies.

Randomization of the survey to the Traditional Instrument versus the Modified Instrument was programmed into the computerized survey instrument, resulting in 3,499 interviews (1,909 Latinos, 1,590 Asian Americans) conducted using the Traditional Instrument, and the remaining 1,150 interviews (645 Latinos, 505 Asians) conducted using the Modified Instrument. In order to ascertain that the randomization was implemented appropriately, we compare the sociodemographic characteristics between the NLAAS subsamples receiving the two versions of the instrument, using χ^2 tests for the association between each sociodemographic variable and instrument version (traditional versus modified).

Measures

The dependent variables are indicators of mental health service utilization: (1) overall use of any services, (2) use of aggregate service categories: specialists, generalists, human services, and alternative services, and (3) use of specific service categories: psychiatrists, psychologists, other mental health professionals (defined as specialists); general practitioners, other medical doctors, other health professionals (defined as generalists); social workers, counselors, religious or spiritual advisors (defined as human services); and other healers (defined as alternative services). The specific service categories follow the ones provided in the service battery; the aggregate service categories are combined across specific service categories. We do not include the other service types, such as inpatient, as dependent variables because of the small number of respondents endorsing these services.

The primary predictor variable is Instrument Version (Traditional versus Modified). We also compare Traditional and Modified Instruments stratified by ethnicity to assess the impact of Instrument Version within each ethnic group, and compare the impact between the two ethnic groups to assess the interaction between ethnicity and Instrument Version. Finally, we also employ the following covariates to examine the sensitivity of the results to the inclusion of covariates: gender, age (18–34, 35–49, 50–64, 65+), annual household income (<\$15,000, \$15,000–\$34,999, \$35,000–\$74,999, \$75,000+), education (<12 years, 12 years, 13–15 years, 16+ years), interview language (English, Spanish, Asian [combining Mandarin Chinese, Tagalog, and Vietnamese]), and the number of lifetime disorders (0, 1, 2, 3+).

Hypothesis and Data Analyses

We hypothesize that survey conditioning manifests as follows:

- H1:** The reported rate of service use is higher under the Modified Instrument than under the Traditional Instrument.

For each service use measure, we compare the rate of service use reported under the two versions of the instrument, using the χ^2 test for the 2×2 cross-tabulation of Instrument Version by service use status. We also repeat these comparisons using logistic regression models that control for additional covariates, to make sure the results were not due to imbalance in the randomization of the respondents to the two versions of the instrument. The analyses are conducted for the entire NLAAS sample and for the Latino and Asian subsamples separately. All statistical tests are conducted at the significance level $\alpha = 0.05$. As the primary focus of this analysis is on the internal validity of the randomized comparison across Instrument Versions, we do not incorporate the complex sample design features (weighting and clustering) into the analyses.

RESULTS

As shown in the “Interview Timing” column in Figure 1, an average of approximately 25 minutes elapsed between Symptoms Screener and service use stem questions for the Traditional Instrument, giving respondents interviewed with this instrument plenty of exposure to the consequences of endorsing stem questions, for survey conditioning to take place.

Table 1 presents the comparisons of the sociodemographic characteristics between the recipients of the traditional instrument and recipients of the modified instrument. The first panel of this table presents the comparisons for the overall sample, under the heading “Total.” The second panel presents the comparisons among Latinos; the third panel presents the comparisons among Asian Americans. No statistically significant differences are found between the two subsamples in terms of gender, age, education, interview language, and the number of disorders, both for the overall sample and for each ethnic group (p -values ranging from .180 to 1.000), except for a moderate difference among Latinos: household income is moderately different ($p = .026$) between those administered the traditional instrument versus those given the modified instrument. In light of the large number of tests conducted simultaneously (18

Table 1: Comparison for Demographics between Traditional and Modified Versions of NLAAS Instrument

	Total			Latino			Asian		
	Traditional (N = 3,499)	Modified (N = 1,150)	p-Value*	Traditional (N = 1,909)	Modified (N = 645)	p-Value*	Traditional (N = 1,590)	Modified (N = 505)	p-Value*
Gender: male	45.21	47.22	.236	43.37	46.36	.187	47.42	48.32	.726
Age (years)			.494			.180			1.000
18-34	40.50	39.13		42.48	39.84		38.11	38.22	
35-49	32.58	32.78		31.22	31.78		34.21	34.06	
50-64	18.26	20.09		16.92	20.31		19.87	19.80	
65 or more	8.66	8.00		9.38	8.06		7.80	7.92	
Imputed household income			.416			.026			.253
<\$15,000	23.09	24.52		27.34	29.30		17.99	18.42	
\$15,000-\$34,999	21.72	21.04		26.77	27.91		15.66	12.28	
\$35,000-\$74,999	27.64	25.57		28.71	22.79		26.35	29.11	
\$75,000+	27.55	28.87		17.18	20.00		40.00	40.20	
Education (years)			.420			.225			.531
0-11	27.62	29.85		38.26	40.78		14.84	15.87	
12	22.04	20.28		25.31	23.26		18.11	16.47	
13-15	38.28	37.95		30.19	28.06		47.99	50.60	
16+	12.06	11.92		6.24	7.91		19.06	17.06	
Interview language			.368			.980			.291
English	51.73	51.30		41.38	41.24		64.15	64.16	
Spanish	32.07	32.96		58.62	58.76		0.19	0.00	
Chinese	6.32	6.61		0.00	0.00		13.90	15.05	
Taglog	1.17	1.74		0.00	0.00		2.58	3.96	
Vietnamese	8.72	7.39		0.00	0.00		19.18	16.83	
Number of disorders			.569			.575			.942
0	76.88	76.26		71.45	71.01		83.40	82.97	
1	12.43	13.57		13.93	15.19		10.63	11.49	
2	6.29	5.48		8.54	7.13		3.58	3.37	
3	4.40	4.70		6.08	6.67		2.39	2.18	

*The p-values compare recipients of traditional instrument versus recipients of modified instrument.

tests in Table 1), this moderate difference might be attributable to chance in multiple comparisons.

The results of the χ^2 tests for survey conditioning are shown in Table 2. For the overall sample combining Latinos and Asians, survey conditioning is statistically significant for overall use of services (odds ratio [OR] = 1.54, 95 percent CI 1.33–1.78, $p < .001$), and for all aggregate and specific service use categories (ORs ranging from 1.41 to 3.10, all p -values $< .001$). Reported rate of service use increases substantially across all categories when the assessment is made with the Modified Instrument instead of the Traditional Instrument.

Similar results hold for the Latino and Asian American subsamples. Among Latinos, survey conditioning is statistically significant for overall use of services (OR = 1.50, 95 percent CI 1.25–1.81, $p < .001$), and for all categories (ORs ranging from 1.47 to 3.26, p -values ranging from .001 to $< .001$). Among Asian Americans, survey conditioning is statistically significant for overall use of services (OR = 1.61, 95 percent CI 1.27–2.03, $p < .001$), and for most categories (ORs ranging from 1.76 to 3.30, p -values ranging from .042 to $< .001$), except for the use of specialists (OR = 1.19, 95 percent CI 0.86–1.64, $p = .298$), psychiatrists (OR = 1.39, 95 percent CI 0.94–2.06, $p = .094$), psychologists (OR = 1.37, 95 percent CI 0.90–2.09, $p = .139$), other health professionals (OR = 1.86, 95 percent CI 0.96–3.62, $p = .068$), and counselors (OR = 1.45, 95 percent CI 0.98–2.14, $p = .065$).

We also compare the effect of survey conditioning between Latinos and Asians to test the moderating effect of ethnicity on the effect of survey conditioning. The moderating effect is statistically insignificant for overall service use and for all service use categories, with OR's ranging from 0.77 to 1.35, and p -values ranging from .164 to .959.

Two sets of logistic regression analysis, one adjusting for gender and annual household income (Model 1) and one adjusting in addition for age, education, and the number of lifetime disorders (Model 2) were fitted for the overall sample, and separately for Latinos and Asians (results not shown but available upon request). The ORs based on the adjusted logistic regression models are very similar to the ORs based on the cross-tabulation analysis without adjusting for covariates, with the ratios between adjusted ORs and unadjusted ORs ranging from 0.98 to 1.04 for Model 1, and ranging from 0.99 to 1.18 for Model 2, indicating that the findings presented in Table 2 (more service use is reported using the Modified Instrument than the Traditional Instrument) remained when adjusted for covariates, and might be slightly stronger when adjusted for the covariates in Model 2.

Table 2: Comparison of Service Use for NLAAS Subsamples Receiving Traditional and Modified Versions of NLAAS Instrument: Odds Ratios (ORs), 95% Confidence Intervals (CIs), and *p*-Values

Service Use Categories	Latinos and Asians Combined (Total)						Latinos						Asians					
	Traditional (<i>N</i> = 3,499)			Modified (<i>N</i> = 1,150)			Traditional (<i>N</i> = 1,909)			Modified (<i>N</i> = 645)			Traditional (<i>N</i> = 1,590)			Modified (<i>N</i> = 505)		
	(%)			(%)			(%)			(%)			(%)			(%)		
Overall use	25.1	34.1	1.54 (1.33–1.78)	<.001	30.7	40.0	1.50 (1.25–1.81)	<.001	18.4	26.5	1.61 (1.27–2.03)	<.001	18.4	26.5	1.61 (1.27–2.03)	<.001		
Specialists	13.7	18.3	1.41 (1.18–1.68)	<.001	17.0	23.7	1.52 (1.22–1.88)	<.001	9.7	11.3	1.19 (0.86–1.64)	.298	9.7	11.3	1.19 (0.86–1.64)	.298		
Psychiatrists	8.6	12.9	1.57 (1.27–1.93)	<.001	11.1	16.9	1.64 (1.27–2.10)	<.001	5.7	7.7	1.39 (0.94–2.06)	.094	5.7	7.7	1.39 (0.94–2.06)	.094		
Psychologists	7.4	11.7	1.64 (1.32–2.05)	<.001	9.6	15.7	1.75 (1.35–2.27)	<.001	4.8	6.5	1.37 (0.90–2.09)	.139	4.8	6.5	1.37 (0.90–2.09)	.139		
Other mental health professionals	2.3	5.0	2.17 (1.54–3.07)	<.001	3.1	6.2	2.07 (1.37–3.13)	.001	1.4	3.4	2.37 (1.26–4.48)	.008	1.4	3.4	2.37 (1.26–4.48)	.008		
Generalists	13.0	19.4	1.61 (1.35–1.92)	<.001	16.8	22.8	1.47 (1.18–1.83)	.001	8.5	15.0	1.91 (1.41–2.58)	<.001	8.5	15.0	1.91 (1.41–2.58)	<.001		
General practitioners	11.3	17.0	1.62 (1.34–1.95)	<.001	14.4	19.8	1.47 (1.17–1.86)	.001	7.5	13.5	1.92 (1.40–2.64)	<.001	7.5	13.5	1.92 (1.40–2.64)	<.001		
Other medical doctors	3.4	9.2	2.86 (2.18–3.75)	<.001	4.5	11.2	2.66 (1.92–3.69)	<.001	2.1	6.7	3.30 (2.03–5.37)	<.001	2.1	6.7	3.30 (2.03–5.37)	<.001		
Other health professionals	2.0	4.3	2.21 (1.52–3.21)	<.001	2.4	5.4	2.38 (1.51–3.73)	<.001	1.5	2.8	1.86 (0.96–3.62)	.068	1.5	2.8	1.86 (0.96–3.62)	.068		
Human services	11.3	20.2	1.99 (1.66–2.37)	<.001	13.6	25.0	2.11 (1.69–2.63)	<.001	8.5	14.1	1.76 (1.30–2.40)	<.001	8.5	14.1	1.76 (1.30–2.40)	<.001		
Social workers	3.5	6.3	1.86 (1.38–2.52)	<.001	4.8	8.5	1.84 (1.30–2.60)	.001	1.8	3.4	1.88 (1.02–3.44)	.042	1.8	3.4	1.88 (1.02–3.44)	.042		
Counselors	7.2	12.2	1.78 (1.43–2.21)	<.001	8.7	15.7	1.95 (1.50–2.54)	<.001	5.5	7.7	1.45 (0.98–2.14)	.065	5.5	7.7	1.45 (0.98–2.14)	.065		
Religious/spiritual advisors	5.1	12.2	2.60 (2.06–3.28)	<.001	6.1	15.3	2.78 (2.09–3.69)	<.001	3.8	8.1	2.25 (1.49–3.40)	<.001	3.8	8.1	2.25 (1.49–3.40)	<.001		
Alternative services (other healers)	1.7	5.2	3.10 (2.16–4.46)	<.001	2.0	6.2	3.26 (2.07–5.12)	<.001	1.4	4.0	2.81 (1.53–5.16)	.001	1.4	4.0	2.81 (1.53–5.16)	.001		

DISCUSSION

Lucas et al.'s (1999) recommendation to present stem questions in the ensemble format (placing the stem questions together up front) was adopted for psychiatric symptom reports to reduce attenuation in the CPES, but except for the 25 percent "experimental sample" in the NLAAS component of the CPES, the same strategy was not applied to assessments of mental health service use. Instead, the service use assessments in the rest of the CPES were presented in the interleaved format.

The randomized trial conducted as part of the NLAAS and reported here finds strong evidence for survey conditioning, with substantially more service use in all categories reported under the Modified Instrument than under the Traditional Instrument. This finding holds for Latinos and Asians for the overall use of services and for most of the subcategories of mental health services, and is insensitive to inclusion of covariates. The difference is substantial for most service use categories examined. For overall service use, the OR is approximately 1.5 for both Latinos and Asians, and for the two groups combined. For Latinos and Asians combined, the rate of overall service use is underreported by approximately nine percentage points out of a total of 34 percentage points when measured with the Traditional Instrument instead of the Modified Instrument, i.e., approximately a quarter of overall service use is missed. This is not only highly significant statistically, but also highly significant in terms of health care policy.

It is therefore reasonable to conclude that the Traditional Instrument, based on the interleaved format for service use assessments, underestimates mental health service use. Future surveys should consider using the Modified Instrument, based on the ensemble format, in order to avoid underreporting of service use. More generally, we recommend that an ensemble format be used whenever it is feasible, if the measures being assessed may be susceptible to survey conditioning.

Accurate reporting of rates of service utilization is important for a number of reasons, including addressing one of the primary purposes of the CPES, assessing the extent of unmet need for psychiatric care. Many discussions of mental health policy are motivated by statements similar to that contained in the Surgeon General's Report on Mental Health (USDHHS 1999):

Studies reveal that less than one-third of adults with a diagnosable mental disorder, and even a smaller proportion of children, receive any mental health services in a given year (p. 65).

The primary recommendation of the report is to support policies that increase the number of people who use services (p. 13). While underuse of mental health services among some populations is certain to remain an important policy issue, our findings imply that overall service use could be undercounted in existing surveys substantially, e.g., approximately a quarter of the overall service use might be missed. Our results imply that the extent of unmet need may have been overestimated in past research that used an interleaved format in assessing use of services. We have not attempted to explore in detail the characteristics of those respondents that seem to be increasing their reports after the survey modification, but this is obviously a critical issue to pursue in future research.

A desire to maintain comparability with other surveys (even though possibly flawed) could discourage researchers from modifying existing instruments for improvements. Continuing to use an existing instrument makes it easy to contrast new results with that of previous studies. On the other hand, continuing to use an existing instrument with recognized problems can inhibit advances in the field and enshrine misleading findings. From the perspective of "continuous quality improvement," the randomized trial conducted in the NLAAS provides a way to improve upon an existing instrument while also maintaining the comparability of data between new and previous studies.

When two versions of an instrument are found to be discrepant in the reported rates of service use, as we found in the NLAAS randomized trial, the service use status that would be reported by a respondent under the traditional instrument is missing for respondents assigned to the modified instrument, and vice versa. Multiple imputation (Rubin 1987; Schafer 1997) can be used to impute these missing values, using other variables predictive of the missing service use status. This strategy will provide more efficient comparisons between the current study (the NLAAS) and previous studies conducted using the existing instrument, as well as comparisons between the current study and future studies that might be conducted entirely using the modified instrument. Therefore, the payoff for the randomized trial conducted in the NLAAS might continue beyond the NLAAS and CPES, in terms of more accurate measurements in future studies.

A number of limitations apply to the present study. No patient chart or insurance information is available to evaluate the accuracy of the self-reported service use data. The survey conditioning study did not assess the accuracy with which respondents with limited experience with the health care system identify the type of providers they saw. For example, if a respondent saw a therapist but did not know whether the therapist was a psychiatrist,

psychologist, or social worker, the respondent might misclassify the type of provider seen. Future studies are required to evaluate the concordance of reports of interviewees with other objective data sources. Furthermore, our interpretation of the attenuation effect in survey response in terms of the respondents learning the stem-and-branch structure and avoiding positive responses to stem questions might not be the only interpretation; other interpretations such as improved comprehension of the survey questions by the respondents might also be plausible interpretations. Nevertheless, what seems to be beyond doubt is that the format of the questions can have major effects on respondent reports in psychiatric and service studies, suggesting more attention should be paid to this issue in future studies.

ACKNOWLEDGMENTS

The NLAAS data used in this analysis were provided by the Center for Multicultural Mental Health Research at the Cambridge Health Alliance. The project was supported by NIH U01 MH62207 and U01 MH62209 funded by the National Institute of Mental Health as well as SAMHSA/CMHS and OBSSR. Naihua Duan also received support from NIMH P30 MH068639, NIMH P30 MH58107, and NCMHD P20 MD000148. We acknowledge helpful discussions with Drs. Bonnie Ghosh-Dastidar and Xiaoli Meng; helpful programming assistance from Dr. Zhun Cao and Chihnan Chen; helpful editorial and administrative assistance from Danielle Seiden and Maria Torres; and helpful comments on an earlier version of this paper by Dr. Patrick Shrout, the *HSR* editor, and two anonymous reviewers.

REFERENCES

- Alegria, M., D. Takeuchi, G. Canino, N. Duan, P. Shrout, X. L. Meng, W. Vega, N. Zane, D. Vila, M. Woo, M. Vera, P. Guarnaccia, S. Aguilar-Gaxiola, S. Sue, J. Escobar, K. M. Lin, and F. Gong. 2004. "Considering Context, Place and Culture: The National Latino and Asian American Study." *International Journal of Methods in Psychiatric Research* 13 (4): 208–20.
- Armknrecht, P. A., W. F. Lane, and K. J. Stewart. 1997. "New Products and the US Consumer Price Index." In *The Economics of New Goods*, edited by T. F. Bresnahan and R. J. Gordon, pp. 375–96. Chicago: University of Chicago Press.
- Bailar, B. A. 1975. "The Effects of Rotation Group Bias on Estimates from Panel Surveys." *Journal of the American Statistical Association* 70 (349): 23–30.

- . 1989. "Information Needs, Surveys, and Measurement Errors." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 1–25. New York: John Wiley.
- Cantor, D. 1989. "Substantive Implications of Longitudinal Design Features: The National Crime Survey as a Case Study." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 25–51. New York: John Wiley.
- Colpe, L., K. R. Merikangas, B. Cuthbert, and K. Bourdon. 2004. "Guest Editorial." *International Journal of Methods in Psychiatric Research* 13 (4): 193–4.
- Corder, L. S., and D. G. Horvitz. 1989. "Panel Effects in the National Medical Care Utilization and Expenditure Survey." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 304–19. New York: John Wiley.
- Duan, N., and R. O. Valdez. 1999. "Design Issues in the Gulf War Veterans Health Study." In *Gulf War Veterans: Measuring Health*, edited by L. M. Hernandez, J. S. Durch, D. G. Blazer II, and I. V. Hoverman, pp. 96–126. Washington, DC: National Academy Press.
- Heeringa, S. G. 2004. *2002–2003 National Latino and Asian American Study (NLAAS): Technical Sample Design Documentation*. Ann Arbor, MI: University of Michigan, Institute for Social Research, Survey Research Center.
- Heeringa, S. G., J. Wagner, M. Torres, N. Duan, T. Adams, and P. Berglund. 2004. "Sample Designs and Sampling Methods for the Collaborative Psychiatric Epidemiology Studies (CPES)." *International Journal of Methods in Psychiatric Research* 13 (4): 221–40.
- Holt, D. 1989. "Panel Conditioning; Discussion." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 340–7. New York: John Wiley.
- Jackson, J. S., M. Torres, C. H. Caldwell, H. W. Neighbors, R. M. Nesse, R. J. Taylor, S. J. Trierweiler, and D. R. Williams. 2004. "The National Survey of American Life: A Study of Racial, Ethnic and Cultural Influences on Mental Disorders and Mental Health." *International Journal of Methods in Psychiatric Research* 13 (4): 196–207.
- Jensen, P., M. Roper, P. Fisher, J. Piacentini, G. Canino, J. Richters, M. Rubio-Stipec, M. Dulcan, S. Goodman, and M. Davies. 1995. "Test–Retest reliability of the Diagnostic Interview Schedule for Children (DISC2.1): Parent, Child and Combined Algorithms." *Archives of General Psychiatry* 52: 61–71.
- Jensen, P., H. K. Watanabe, and J. E. Richters. 1999. "Who's Up First? Testing the Order Effects in Structured Interviews Using a Counterbalanced Experimental Design." *Journal of Abnormal Child Psychology* 27 (6): 439–45.
- Kessler, R. C., K. A. McGonagle, S. Zhao, C. B. Nelson, M. Hughes, S. Eshleman, H. U. Wittchen, and K. S. Kendler. 1994. "Lifetime and 12-Month Prevalence of DSM-III-R Psychiatric Disorders in the United States: Results from the National Comorbidity Survey." *Archives of General Psychiatry* 51 (1): 8–19.
- Kessler, R. C., and K. R. Merikangas. 2004. "The National Comorbidity Survey Replication (NCS-R): Background and Aims." *International Journal of Methods in Psychiatric Research* 13 (2): 60–8.
- Kessler, R. C., and T. B. Ustun. 2004. "The World Mental Health (WMH) Survey Initiative Version of the World Health Organization (WHO) Composite Inter-

- national Diagnostic Interview (CIDI)." *International Journal of Methods in Psychiatric Research* 13 (2): 93–121.
- Kessler, R. C., H.-U. Wittchen, J. M. Abelson, K. McGonagle, N. Schwarz, K. S. Kendler, B. Knäuper, and S. Zhao. 1998. "Methodological Studies of the Composite International Diagnostic Interview (CIDI) in the US National Comorbidity Survey." *International Journal of Methods in Psychiatric Research* 7: 33–55.
- Kessler, R. C., H.-U. Wittchen, J. Abelson, and S. Zhao. 2000. "Methodological Issues in Assessing Psychiatric Disorder with Self-Reports." In *The Science of Self-Report: Implications for Research and Practice*, edited by A. A. Stone, J. S. Turrkkan, C. A. Bachrah, J. B. Jobe, H. S. Kurtzman, and V. S. Cain, pp. 229–55. Mahwah, NJ: Erlbaum.
- Knowles, E. S., M. C. Coker, R. A. Scott, D. A. Cook, and J. W. Neville. 1996. "Measurement-Induced Improvement in Anxiety: Mean Shifts with Repeated Assessment." *Journal of Personality and Social Psychology* 71 (2): 352–63.
- Lucas, C. P., P. Fisher, J. Piacentini, H. Zhang, P. S. Jensen, D. Shaffer, M. Dulcan, M. Schwab-Stone, D. Regier, and G. Canino. 1999. "Features of Interview Questions Associated with Attenuation of Symptom Reports." *Journal of Abnormal Psychology* 27 (6): 429–37.
- Pennell, S. G., and J. M. Lepkowski. 1992. *Panel Conditioning Effects in the Survey of Income and Program Participation. Proceedings of the Survey Research Methods Section*. Washington, DC: American Statistical Association.
- Piacentini, J., M. Roper, P. Jensen, C. Lucas, P. Fisher, H. Bird, K. Bourdon, M. Schwab-Stone, M. Rubio-Stipec, M. Davies, and M. Dulcan. 1999. "Informant Based Determinants of Symptom Attenuation in Structured Child Psychiatric Interviews." *Journal of Abnormal Child Psychology* 27 (6): 417–28.
- Robins, L. N., J. E. Helzer, J. L. Croughan, and K. S. Ratcliff. 1981. "National Institute of Mental Health Diagnostic Interview Schedule: Its History, Characteristics and Validity." *Archives of General Psychiatry* 38: 328–66.
- Robins, L. N., and D. A. Regier. 1991. *Psychiatric Disorders in America*. New York: The Free Press.
- Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Schafer, J. L. 1997. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Silberstein, A. R., and C. A. Jacobs. 1989. "Symptoms of Repeated Interview Effects in the Consumer Expenditure Interview Survey." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 289–303. New York: John Wiley.
- U.S. Department of Health and Human Services (USDHHS). 1999. *Mental Health: A Report of the Surgeon General*. Rockville, MD: USDHHS.
- Vega, W. A., B. Kolody, S. Aguilar-Gaxiola, E. Alderete, R. Catalano, and J. Caraveo-Anduaga. 1998. "Lifetime Prevalence of DSM-III-R Psychiatric Disorders among Urban and Rural Mexican Americans in California." *Archives of General Psychiatry* 55 (9): 771–8.
- Waterton, J., and D. Lievesley. 1989. "Evidence of Conditioning Effects in the British Social Attitudes Panel." In *Panel Surveys*, edited by D. Kasprzyk, G. Duncan, G. Kalton, and M. P. Singh, pp. 319–39. New York: John Wiley.