The US National Comorbidity Survey Replication (NCS-R): design and field procedures

RONALD C. KESSLER,[1] PATRICIA BERGLUND,[2] WAI TAT CHIU,[1] OLGA DEMLER,[1] STEVEN HEERINGA,[2] EVA HIRIPI,[1] ROBERT JIN,[1] BETH-ELLEN PENNELL,[2] ELLEN E. WALTERS,[1] ALAN ZASLAVSKY,[1] HUI ZHENG[1]

[1] Department of Health Care Policy, Harvard Medical School, Boston MA, USA
[2] Institute for Social Research, University of Michigan, Ann Arbor MI, USA

ABSTRACT *The National Comorbidity Survey Replication (NCS-R) is a survey of the prevalence and correlates of mental disorders in the US that was carried out between February 2001 and April 2003. Interviews were administered face-to-face in the homes of respondents, who were selected from a nationally representative multi-stage clustered area probability sample of households. A total of 9,282 interviews were completed in the main survey and an additional 554 short non-response interviews were completed with initial non-respondents. This paper describes the main features of the NCS-R design and field procedures, including information on fieldwork organization and procedures, sample design, weighting and considerations in the use of design-based versus model-based estimation. Empirical information is presented on non-response bias, design effect, and the trade-off between bias and efficiency in minimizing total mean-squared error of estimates by trimming weights.*

Key words: design effects, epidemiologic research design, sample bias, sample weights, survey design efficiency, survey sampling

The current paper presents an overview of the National Comorbidity Survey Replication (NCS-R) survey design and field procedures. We also discuss weighting and design effects. Although the instrument used in the NCS-R is described in a separate report in this issue (Kessler and Üstün, 2004), there is some discussion of broad outlines of the interview, with implications for the design of the survey.

Survey mode

The NCS-R was carried in the homes of a nationally representative sample of respondents between February 2001 and April 2003. The survey was administered using laptop computer-assisted personal interview (CAPI) methods by professional survey interviewers employed by the Survey Research Center (SRC) of the Institute for Social Research at the University of Michigan. The decision to use face-to-face administration rather than telephone, mail, or Internet administration was based on four main factors, the first three of which come from the literature on survey methodology (Groves, Fowler, Couper, Lepkowski, Singer and Tourangeau, in press) and the fourth of which is based on considerations unique to the NCS-R. First, the coverage properties of an area probability sample are superior to other samples such as those used in telephone, mail, or Internet surveys. Second, the accuracy of screening and household enumeration procedures, which are required to create a probability sample, is greater in face-to-face surveys than in surveys based on these other modes of data collection. Third, response rates are generally much higher in face-to-face surveys than in those based on other modes of data collection. Fourth, the NCS-R interview schedule was quite long and highly complex, making it impossible to use these other modes effectively.

Although the above four considerations were sufficient to convince us that a face-to-face survey mode was needed, there were also additional advantages of this mode that we recognized and used to our advantage. One was related to the issue of length and

complexity of the interview, which can lead to high respondent burden in some cases. The face-to-face survey mode made it possible for interviewers to gauge respondent fatigue and to suggest short breaks if respondents needed time to regain their focus. Along the same lines, interviews with respondents who had complex histories of psychopathology were often broken up into two or more sessions that were spread out over a period of days or even weeks. In a related way, use of the face-to-face mode minimized the problem of the respondent prematurely halting the interview (which is common in telephone administration) or completing only part of the assessment (which is common in mail questionnaires and Internet surveys). Break-offs of this sort are rare in in-person surveys. This is especially true when, as in the NCS-R, interviewers are trained to monitor respondent fatigue and to suggest breaks. Consistent with this thinking, only 107 out of 9,389 initial NCS-R respondents broke off an interview.

The decision to use CAPI rather than paper-and-pencil (PAPI) administration was based on the fact that the interview schedule had many complex skips. These skips create opportunities for interviewer error in PAPI that are avoided in CAPI due to the computer controlling the skip logic. Computer-assisted personal interviews can also be cost-effective when the sample size is large because the investment in the application programming is less than the labour needed to keypunch PAPI responses. An additional appeal of CAPI compared to PAPI is that the interviewer can be prompted for missing or inconsistent responses while the interview is in progress, allowing these problems to be resolved immediately.

Given the fact that the NCS-R interview asked about a number of embarrassing feelings and behaviours, a question could be raised whether the method of audio computer-assisted self-administered interviewing (A-CASI) should have been used instead of more conventional CAPI. The use of A-CASI allows respondents to enter answers to embarrassing questions into a laptop without the interviewer knowing their answers by using digital audio recordings and headsets connected to the laptop to administer the survey questions. There is now impressive evidence suggesting that A-CASI leads to significantly higher reports of some illegal or embarrassing behaviours (Turner, Lessler and Devore, 1982; Tourangeau and Smith, 1998; Turner, Ku, Rogers, Lindberg, Pleck and Sonenstein, 1998). Our decision not to use A-CASI despite this evidence was based on a concern about non-comparability of responses for purposes of trending with the baseline NCS and also with timing considerations. Regarding the latter, the field period for the NCS-R was set back more than 2 years from its original start date because of unplanned complexities in mounting a parallel national survey of adolescent mental health. The use of A-CASI would have added to this delay.

As noted above, it was sometimes necessary to administer an interview over two or more sessions. Most of the follow-up sessions were carried out in person in the homes of respondents. Interviewers were allowed to complete follow-up sessions over the telephone in three kinds of situations: (1) when the respondent requested telephone administration for the remaining questions; (2) when the respondent lived in a remote rural area that required a great deal of travel time for the interviewer to reach; and (3) when the remaining questions were few in number. In each of these three situations, the interviewer left a respondent booklet with the respondent to be used as a visual aid in completing the remainder of the interview by telephone.

The NCS-R interview schedule

As described in more detail by Kessler and Üstün (2004), the NCS-R interview schedule was the version of the World Health Organization (WHO) Composite International Diagnostic Interview (CIDI) that was developed for the WHO World Mental Health (WMH) Survey Initiative. This instrument is referred to as the WMH-CIDI. A small number of supplemental sections were also included that were unique to the US version of the survey. A hard copy of the instrument is posted at www.hcp.med.harvard.edu/ncs. Only one important aspect of the interview is discussed here – its length – as this had important implications for design and field procedures.

The NCS-R interview schedule was quite long. It took a minimum of 90 minutes to complete among respondents who reported no lifetime disorders, an average of approximately 2 hours and 30 minutes among people with a history of disorder, and as long as 5 to 6 hours among respondents with a very complex history of many different disorders. This long length was due to the fact that the scientific

questions guiding instrument development required us to assess a wide variety of topics in the same sample of respondents. Some of the sections could have been administered independently of the main interview in a separate survey without undercutting the investigation of the issues considered in those sections. However, given that major mental health surveys like the NCS-R are only funded once every decade, we were eager to keep all of these sections in the instrument.

The problem of interview length was addressed in three ways. First, we carefully reviewed each question in the interview to make sure it added value. We also carefully evaluated each skip instruction to make sure that we were skipping respondents out of sections as soon as we had the information needed to evaluate the issues under consideration. This was especially important in the diagnostic sections, where it was possible to skip respondents once it became clear that they failed to meet any symptom required for a diagnosis. Given that one of our aims was to explore sub-threshold diagnoses, though, we had to balance the desire to invoke skips with the need to obtain sub-threshold information.

Second, we evaluated the data-analysis goals to determine if they could be achieved by administering the section to a probability subsample of respondents rather than to all respondents. For example, one section replicated questions about psychological distress that were originally developed for the 1957 Americans View Their Mental Health survey (Gurin, Veroff and Feld, 1960) and were later repeated in a 1976 replication of that survey (Veroff, Douvan and Kulka, 1981). The aim was to merge the archival individual-level data files from these earlier surveys with NCS-R data to study trends in self-reported psychological distress over time. However, as the earlier surveys were based on samples of 1,800–2,100 respondents, there would have been little incremental improvement in the statistical power of trend comparisons if we administered these questions to the entire sample. This section was consequently administered to a 30% subsample. Similar subsampling was used to evaluate family burden and to assess a number of disorders that were either included for exploratory purposes (for example, adult attention-deficit/hyperactivity disorder, adult separation anxiety disorder) or that required in-depth assessment (non-affective psychosis, obsessive-compulsive disorder).

Third, the interview schedule was divided into two parts. Part I, administered to all respondents, included all core WMH-CIDI disorders. The administration time of Part I averaged 33.8 minutes and had an inter-quartile range between 22.6 and 39.8 minutes (Table 1). Part II included assessments of risk factors, consequences, services, and other correlates of the core disorders. Part II also included assessments of additional disorders that were either of secondary importance or that were very time-consuming to assess. The administration time of Part II had a mean of 109.4 minutes, a median of 101.7 minutes, and an inter-quartile range between 83.9 and 124.1 minutes. In order to reduce respondent burden, Part II was administered only to 5,692 of the 9,282 NCS-R respondents, over-sampling those with clinically significant psychopathology. All respondents who did not receive Part II were administered a brief demographic battery and were then either terminated or sampled in their appropriate proportions into sub-sampled interview sections that are described below.

Table 1. Administration time (minutes) of the Part I and Part II NCS-R interview schedule

Percentile	Part I only	Among respondents who completed: Part I and Part II	
		Part I	Part II
0	2.4	0.1	0.1
25	22.6	33.5	83.9
50	29.7	49.2	101.7
75	39.8	70.8	124.1
99	102.9	182.4	256.5
Mean	33.8	57.3	109.4

Selection into Part II was controlled by the CAPI program, which divided respondents into three strata based on their Part I responses. The first stratum consisted of respondents who either (1) met lifetime criteria for at least one of the mental disorders assessed in Part I, (2) met subthreshold lifetime criteria for any of these disorders and sought treatment for at least one of them at some time in their life, or (3) ever in their life either made a plan to commit suicide or attempted suicide. All of these people were administered Part II. The second stratum consisted of respondents who, although not meeting criteria for membership in the first stratum, gave responses in Part I indicating that they either (1) ever met subthreshold criteria for any of the Part I disorders, (2) ever sought treatment for any emotional or substance problem, (3) ever had suicidal ideation, or (4) used any psychotropic medications (whether or not under the direct supervision of a physician) in the past 12 months to treat emotional problems. A probability sample of 59% of the respondents in this second stratum was selected to receive Part II. The third stratum consisted of all other respondents, of whom 25% were selected to receive Part II.

As virtually all planned analyses of the NCS-R data feature comparisons of cases with non-cases, and as the number of non-cases substantially exceeds the number of cases, the under-sampling of non-cases in the second and third strata has only a small effect on statistical power to study correlates of disorder. This assertion is demonstrated empirically later in the paper. An additional efficiency of the Part II sample is that respondents in the second and third strata were selected with probabilities proportional to their household size. This approach, which is described in more detail below, is opposite of the procedure used in initial sample selection, leading to a partial cancelling out of the variation in the within household probability of selection weight.

The survey population

The survey was designed to be representative of English-speaking adults ages 18 or older living in the non-institutionalized civilian household population of the coterminous US (excluding Alaska and Hawaii) plus students living in campus group housing who have a permanent household address.

Fieldwork organization and procedures

As noted above, the NCS-R fieldwork was carried out by the professional SRC national field interview staff. Over 300 interviewers participated in data collection. The SRC field staff was supervised by a team of 18 experienced regional supervisors. Supervisors in larger regions also had team leaders who worked with them. A study manager located at the central SRC facility in Michigan oversaw the work of the supervisors and their staff.

After sample selection (see below), each interviewer received a folder from his or her supervisor for each household in which an interview was to be obtained. An advance letter and study fact brochure were sent to each of these households at a time designed to arrive a few days before the interviewer made their first contact attempt. The letter explained the purpose of the study and gave a toll-free number for respondents who had additional questions. The study fact brochure contained answers to frequently asked questions. Upon making in-person contact with the household, the interviewer explained the study once again and obtained a household listing. This listing was then used to select a random respondent in the household. The random respondent was approached, the interview was explained, and verbal informed consent was obtained. Respondents were given $50 as a token of appreciation for participating in the survey. It should be noted that verbal rather than written informed consent was obtained because the NCS-R was designed as a trend study replication of the baseline NCS, which used verbal informed consent.

In cases where the interviewer had difficulty contacting the household or in which the household was reluctant to participate, persuasion letters were sent to the household. Sixty days before the end of the field period, a special effort was made to recruit as many unresolved cases as possible by sending a special recruitment letter and offering a higher financial incentive ($100) to complete a truncated version of the interview either in person or over the telephone. Interviewers were allowed to make unlimited in-person contact attempts to complete these final interviews and were given financial incentives for completing these interviews during the closeout period.

Survey Research Center interviewers were paid by the hour rather than by the interview. This is

important in light of evidence that interviewers paid by the interview tend to get low response rates because they focus their efforts on interviewing easy-to-recruit respondents. In the case of the WMH-CIDI, where there is great variation in length of interview, per-interview payment rather than per-hour payment also encourages interviewers to rush through long interviews. We wanted to avoid this and, in fact, we encouraged interviewers during training to work especially hard at long interviews because respondents with long interviews are generally those with complex histories of psychopathology, which are of special interest to the project. We also encouraged interviewers to set up appointments for second and third interview sessions to complete long interviews. It is easy to facilitate such procedures when interviewers are paid by the hour.

The Human Subjects Committees of both Harvard Medical School (HMS) and the University of Michigan approved these recruitment, consent, and field procedures.

Interviewer training

Each professional SRC interviewer must complete a two-day general interviewer training (GIT) course before working on any SRC survey. Moreover, experienced interviewers have to complete GIT refresher courses on a periodic basis. Each interviewer who worked on NCS-R also received 7 days of study-specific training. Each interviewer had to complete an NCS-R certification test that involved administering a series of practice interviews with scripted responses before beginning production work.

Fieldwork quality control

As noted above, sample households were selected centrally to avoid interviewers selectively recruiting respondents from attractive neighbourhoods. The random respondent in the household was selected using a standardized method that minimizes the probability of interviewer cheating to select easy-to-recruit household members. In addition, the CAPI program controlled skip logic and had a built-in clock to record speed of data entry, making it difficult for interviewers to truncate interviews by skipping sections or to fake parts of interviews by filling in the last sections quickly.

Despite these structural disincentives to error and cheating, supervisors checked for all of these types of error. Supervisors also made contact with a random 10% of interviewed households to confirm the household address, household enumeration and random selection procedures. Supervisors also confirmed the length of the interview and repeated a random sample of questions in order to make sure that interviewers administered the full interview to respondents and to make sure responses were recorded accurately

Survey Research Center field procedures called for completed CAPI interviews to be sent electronically by interviewers every night. This allowed supervisors to review the completeness of open-ended responses and to make other quality control checks of the data on a daily basis. In cases where problems were detected, the interviewers were contacted and instructed to re-contact the respondent to obtain missing data.

The SRC uses a computerized survey tracking software system to facilitate field quality control. The number of interviews completed, outstanding, response rate, hours per interview, and so forth, are all recorded on a constantly updated basis by this system at the interviewer level along with benchmark comparisons. Supervisors can, at a glance, call into the SRC computer server to monitor these statistics and go over them with individual interviewers in their weekly phone calls. These same data are available to study staff. The supervisors, SRC, and HMS staff used these systems to pinpoint interviewers with low response rates in the first replicate for remedial re-training. Interviewers who persisted in low performance or who were found to make conscious errors were terminated from the study.

In the case of interviews with stem-branch logic, like the WMH-CIDI, an additional type of potential problem is that interviewers who want to shorten the length of the interview can do so by entering negative responses to diagnostic stem questions even when respondents endorse these questions. As noted earlier, SRC interviewers are paid by the hour in an effort to avoid providing an incentive for this type of cheating, but we have seen clear evidence of it in surveys where interviewers were paid by the interview and supervision was only minimal. Therefore, in the NCS-R, as in the other WMH surveys, interviewer-level data were monitored to look for

evidence of systematic patterns of under-reporting diagnostic stem questions. Consistent with the rationale for paying interviewers by the hour, no such evidence was found in the NCS-R.

The sample design

Household sample selection procedures

Respondents were selected from a four-stage area probability sample of the non-institutionalized civilian population using small area data collected by the US Bureau of the Census from the year 2000 census of the US population to select the first two stages of the sample.

The first stage of sampling selected a probability sample of 62 primary sampling units (PSUs) representative of the population. These PSUs were linked to the original PSUs used in the baseline NCS in order to maximize the efficiency of cross-time comparison. Each PSU consisted of all counties in a census-defined metropolitan statistical area (MSA) or, in the case of counties not in an MSA, of individual counties. Primary sampling units were selected with probabilities proportional to size (PPS) and geographic stratification from all possible segments in the country.

The 62 PSUs include 16 MSAs that entered the sample with certainty, an additional 31 non-certainty MSAs, and 15 non-MSA counties. The certainty selections include, in order of size, New York City, Los Angeles, Chicago, Philadelphia, Detroit, San Francisco, Washington DC, Dallas/Fort Worth, Houston, Boston, Nassau-Suffolk NY, St Louis, Pittsburgh, Baltimore, Minneapolis, and Atlanta. These 16 PSUs are referred to as 'self-representing' PSUs because they were not selected randomly to represent other MSAs, but are so large that they represent themselves. Rigorously speaking, these 16 are not PSUs in the technical sense of that term, but rather population strata. The other 46 PSUs are 'non-self-representing' in the sense that they were selected to be representative of smaller areas in the country. Systematic selection from an ordered list was used to select the non-self-representing PSUs, with the order building in secondary stratification for geographic variation across the country. After selection, each of the three largest self-representing PSUs (New York, Los Angeles, Chicago) was divided into four pseudo-PSUs, while each of the remaining 13 self-representing PSUs was divided into two pseudo-PSUs. When combined with the 46 non-self-representing PSUs, this yielded a sample of 84 PSUs and pseudo-PSUs. We henceforth refer to both PSUs and pseudo-PSUs as PSUs.

The second stage of sampling divided each PSU into segments of between 50 and 100 housing units based on 2,000 small-area census data and selected a probability sample of 12 such segments from each non-self-representing PSU. A larger number of segments were selected from the self-representing PSUs using the ratio of the population size over the systematic sampling interval. Within each PSU, the segments were selected systematically from an ordered list with probabilities of selection proportional to size. The order in the list built in secondary stratification for geographic variation. A total of 1,001 area segments were selected in the entire sample.

Once the sample segments were selected, each segment was either visited by an interviewer to record the addresses of all housing units (HUs) (in the case of segments that were not included in the baseline NCS) or the listing used in the baseline NCS was updated for new construction and demolition (in the case of segments that were included in the baseline NCS). These lists were entered into a centralized computer data file. In order to adjust for discrepancies between expected and observed numbers of HUs, a random sample of HUs was selected that equals $10 \times O/E$, where O is the observed number of households listed in the segment and E is the number of HUs expected in the segment from the Census data files.

This three-stage design creates a sample in which the probability of any individual HU being selected to participate in the survey is equal for every HU in the coterminous US. The fourth stage of selection then obtained a household listing of all residents in the age range 18 and older from a household informant. The informant was also asked whether each adult HU resident spoke English. Once the HU listing was obtained, a probability procedure was used to select one or in some cases (see below) two respondents to be interviewed. As the number of sampled cases in the HU did not increase proportionally with increase in HU size, there is a bias in this approach toward underrepresenting people who live in large HUs. As described below, this bias was

corrected by weighting the data to adjust for the differential probability of selection within HUs.

Procedures for sampling students living away from home
The largest segment of the US population not living in HUs consists of students who live in campus group housing (for example dormitories, fraternities, and sororities). We included these students in the sampling frame if they had a permanent home address (typically the home of their parents) by including them in HU listings in all sample households. If they were selected for interview their contact information at school was obtained and special arrangements were made to interview them. Students living in off-campus private housing rather than campus group housing were eligible for sample selection in their school residences. Students of this latter sort were not included in the HU listings of their permanent residences in order to avoid giving such students two chances to enter the sample.

Within-household sample selection procedures
Recruitment of HUs began by the interviewer mailing an advance letter and study fact brochure to the HU. These materials explained the purposes of the study, described the funding sources and the survey organization that was carrying out the survey, listed the names and affiliations of the senior scientists involved in the research, provided information about the content and length of the interview, described confidentiality procedures, clearly stated that participation was voluntary, and provided a toll-free number for respondents who had additional questions. The advance letter said that the interviewer would make an in-person visit to the HU within the next week in order to answer any remaining questions and to determine whether the HU resident selected to participate would be willing to do so.

In cases where it was difficult to find a HU resident at home, at least 15 in-person call attempts were made to each sample HU to make an initial contact with a HU member. Additional contact attempts were made by telephone using a reverse directory and by leaving notes at the HU with the study toll-free number and asking someone in the HU to call the study office to schedule an appointment. Additional in-person contact attempts were made based on the discretion of the supervisor.

Once a HU resident was contacted, a listing of all eligible HU residents was made that recorded the age and sex of each such person, confirmed that each could speak English, and ensured that permanent residents who were students living away from home in campus group housing were included. The Kish table selection method was used to select one eligible person as the primary predesignated respondent. In addition, in a probability sample of households in which there were at least two eligible residents, a second predesignated respondent was selected in order to study within-household aggregation of mental disorders and to reduce variation across HUs in within-household probability of selection into the sample. No attempt was made to select or to interview a secondary respondent unless the primary respondent was interviewed first. This decision was made to avoid compromising the response rate for primary respondents

The sampling fraction for the second predesignated respondent was lower in HUs with only two eligible respondents than those with three or more eligible respondents in order to reduce variance in the within-household probability of selection weight. No more than two interviews were taken per HU even when the number of HU residents was large because we were concerned that taking more than two interviews would adversely affect the response rate by increasing household burden. Moreover, we limited selection of a second respondent to 25% of HUs because we were concerned that using this procedure in a larger proportion of HUs in a sample with a target of 10,000 interviews would adversely affect the geographic dispersion of the sample by reducing the total number of HUs in the sample.

Interviewing hard-to-recruit cases
Hard-to-recruit cases included both HUs that were difficult to contact because the residents were usually away from home and respondents who were reluctant to participate once they were reached. The first of these problems was addressed by being persistent in making contact attempts at all times of day and all days of the week. We also left notes at the HUs where we were consistently unable to make contacts that included toll free numbers for residents to contact us to schedule interview appointments. The second problem was addressed by offering a $50 financial incentive to all respondents and by using a

variety of standard survey persuasion techniques to explain the purpose and importance of the survey and to emphasize the confidentiality of responses.

Main data collection continued until all non-contact HUs had their full complement of call attempts and all reluctant predesignated respondents had a number of recruitment attempts. It was clear at this point that the remaining hard-to-recruit cases were busy people who were unlikely to participate in a survey that could reasonably be expected to last between 2 and 3 hours. We therefore developed a short-form version of the interview that could be administered in less than 1 hour and we made one last attempt to obtain an interview with hard-to-recruit cases using this shortened instrument. Remaining hard-to-reach cases were sent a letter informing them that the study was about to end and that we were offering an honorarium of $100 for completing a shortened version of the interview in the next 30 days that would last no more than one hour. Interviewers were also given a bonus for each short-form interview they completed in this 30-day close-out period, at which point fieldwork ended.

Sample disposition

The sample disposition as of the end of the main phase of data collection is shown in the first four columns of Table 2. The first two columns describe primary predesignated respondents in the 10,843 final sample HUs, while the third and fourth columns describe secondary predesignated respondents in the 1976 HUs that were selected to obtain second interviews. Ineligible HUs are excluded from the table. The response rate at this point in the data collection was 70.9% among primary and 80.4% among secondary predesignated respondents, with a total of 9,282 completed interviews. Non-respondents included 7.3% of primary and 6.3% of secondary cases who refused to participate, 17.7% of primary and 11.6% of secondary respondents who were reluctant to participate (who told the interviewer that they were too busy at the time of contact, but did not refuse), 2.0% of primary and 1.7% of secondary respondents who could not be interviewed because of either a permanent condition (for example, mental retardation) or a long-term situation (such as an overseas work assignment), and HUs that were never contacted (2.0%).

We subsequently attempted to administer short-form interviews to the 2,143 reluctant and no-contact primary predesignated respondents and the 230 reluctant secondary predesignated respondents described in the first two columns of Table 2. In the course of completing the short-form interviews with primary respondents, an additional 104 secondary pre-designated respondents were generated, resulting in a total of 334 secondary pre-designated respondents who we attempted to

Table 2. The NCS-R sample disposition

	Main interview				Short-form interview			
	Primary		Secondary		Primary		Secondary	
	%	(n)	%	(n)	%	(n)	%	(n)
Interview	70.9	(7693)	80.4	(1589)	18.6	(399)	46.4	(155)
Refusal	7.3	(791)[1]	6.3	(124)[1]	75.1	(1609)	44.3	(148)
Reluctant	17.7	(1922)	11.6	(230)	–[2]	–[2]	–[2]	–[2]
Circumstantial	2.0	(216)	1.7	(33)	0.6	(14)	9.3	(31)
No contact	2.0	(221)	–[3]	–[3]	5.6	(121)	–[3]	–[3]
Total		(10,843)		(1976)		(2143)		(334)

[1] Refusals among primary respondents include informants who refused to provide a HU listing, respondents who refused to be interviewed, and respondents who began the interview but broke off before completing Part I. Refusals among secondary respondents include the last two of these three categories.

[2] All non-respondents in the short-form interview phase of the survey were classified as refusals rather than reluctant unless they had circumstantial reasons for not being interviewed.

[3] No contact is defined at the HU-level as never making any contact with a resident. As a result, none of the secondary predesignated respondents was included in this category even if no contact was ever made with them. In the latter case, they were classified as reluctant in the main interview and as refusals in the short-form interview.

interview. The sample disposition is shown in the last four columns of Table 2. The conditional response rate was 18.6% among primary and 46.4% among secondary pre-designated respondents, with a total of 554 completed short-form interviews.

The 9,282 main interviews combined with the 554 short-form interviews total to 9,836, with 8,092 among primary and 1,744 among secondary respondents. Focusing on primary respondents, the overall response rate is 74.6%, the overall cooperation rate (excluding from the denominator HUs that were never contacted) is 76.1%, and the cooperation rate among pre-designated respondents without circumstantial constraints on their participation is 77.8%. Among secondary respondents, the conditional overall response rate is 83.8% and the cooperation rate among pre-designated respondents without circumstantial constraints is 86.5%.

Weighting

Two approaches were taken to weighting the NCS-R data. The first, which we refer to as the non-response (NR) adjustment approach, treats the 9,282 main interviews as the achieved sample. The short-form interviews are treated as non-respondent interviews, which are used to develop a non-response adjustment weight applied to the 9,282 main interviews. The second approach, which we refer to as the multiple-imputation (MI) approach, combines the 9,282 main interviews with the 554 short-form interviews to create a combined sample of 9,836 cases in which the 554 short-form interviews are weighted to treat them as representative of all initial non-respondents. The method of MI (Rubin, 1987) is used to adjust for the fact that the short-form interviews did not include all the questions in the main interviews. Additional weights are applied in both approaches to adjust for differences in within-household probability of selection and to post-stratify the final sample to approximate the distribution of the 2000 Census on a range of socio-demographic variables.

The NR approach is the main one used in analysis of the NCS-R data. The MI approach is more exploratory because MI can lead to downward bias in estimates of associations if the models used to make the imputations do not capture the full complexity of the associations in the main cases. On the other hand, by using explicit models to impute missing values, the MI approach allows more flexibility than the NR approach in combining observed variables for short-form cases with patterns detected in the complete data. In addition, mild modelling assumptions can be used in the MI approach to improve efficiency compared to the NR approach. In cases of this sort, when the parameter estimates are similar in the two approaches, the SE will generally be lower in the MI approach. As a result of these potential advantages of the MI approach, we plan to use MI to replicate marginally significant substantive patterns as a sensitivity analysis, bearing in mind that shrinkage of the associations might occur due to lack of precision in the imputations. Based on this hierarchy between the two approaches in the NCS-R analyses, the focus in this section of the paper is largely on the NR approach.

The non-response adjustment approach

Five weights are applied to the data in the NR approach: a locked building subsampling weight (WT1.1), a within-household probability of selection weight (WT1.2), a non-response adjustment weight (WT1.3), a post-stratification weight (WT1.4), and a Part II selection weight (WT1.5). The joint product of the first four of these weights is the consolidated weight used to analyse data from the Part I sample in the NR approach (n = 9,282), while the joint product of all five weights is the consolidated weight used to analyse data from the Part II sample (n = 5,692). When the data to be analysed include some variables assessed in Part I and other variables assessed in Part II, the Part II sample and weight are used. This section of the paper reviews each of these five weights.

The locked building subsampling weight (WT1.1) gives a double weight to 60 sample HUs from an original 120 apartments in locked apartment buildings where we could not make contact with a superintendent or other official to gain entry. A random 50% of the predesignated units in these buildings were selected for especially intense recruitment effort. These 60 were double weighted. It should be noted that residents of a number of other locked apartment buildings were also included and interviewed, based on the interviewers contacting the building owner, superintendent, or official committee of residents and obtaining permission to carry out interviews in the building. In addition, we were officially refused permission to carry out interviews in five large

locked buildings in New York City, each representing a complete sample segment. Non-probability replacements of other locked buildings in the same neighbourhoods were made in order to avoid complete non-response in these segments.

The within-household probability of selection weight (WT1.2) adjusts for the fact that the probability of selection of respondents within HUs varies inversely with the number of people in the HU. This is true because, as noted earlier, only one or two respondents were selected for interview in each HU. WT1.2 was created by generating a separate observational record for each of the weighted (by WT1.1) 14,025 eligible HU residents in the 7,693 participating HUs. The three variables in the household listing were included in the individual-level records: respondent age and sex and number of eligible residents in the HU. The weighted distributions of these three variables were then calculated for the 9,282 respondents, the 4,733 eligible HU residents who were not interviewed, and for all 14,015 eligible HU residents weighted to adjust for WT1.1. The sum of weights was 14,025 because of this adjustment.

As shown in Table 3, these distributions differ meaningfully, with the greatest difference being for size of household. This is because the individual-level probability of selection within HUs was inversely proportional to HU size, leading to a distortion in the age and sex distributions among respondents because these variables vary with size of HU. WT1.2 was constructed to correct this bias. The weighted three-way cross-classification of age, sex, and household size was computed separately for the 9,282 respondents (weighted to 9,287 because of WT1.1) and for all 14,015 residents of the participating HUs (weighted to 14,025). The ratio of the number of cases in the latter to the former was then calculated separately for each cell and these ratios were applied to each respondent. The sum of the weights across respondents was normed to sum to 14,025. These normed ratios define WT1.2.

The non-response adjustment weight (WT1.3) was then developed and applied to the weighted (by the product of WT1.1 and WT1.2) dataset of 9,282 respondents to adjust for the fact that non-respondents differ from respondents. The participants in the short-form interviews from initial non-respondent HUs were treated as representative of all non-respondents in order to develop this weight. As described in the next subsection, a two-

Table 3. Household listing information for respondents and other eligible residents of the households that participated in the main survey[1]

	Respondents		Others		Total	
	%	(SE)	%	(SE)	%	(SE)
Age						
18–29	22.7	(0.4)	26.9	(0.6)	24.1	(0.4)
30–44	31.7	(0.5)	29.8	(0.7)	31.1	(0.4)
45–59	24.6	(0.4)	26.6	(0.6)	25.3	(0.4)
60+	21.0	(0.4)	16.6	(0.5)	19.5	(0.3)
Sex						
Male	44.6	(0.5)	52.0	(0.7)	47.1	(0.4)
Female	55.4	(0.5)	48.0	(0.7)	52.9	(0.4)
Number of eligible HU residents						
1	29.3	(0.5)	0.0	(0.0)	19.4	(0.3)
2	53.9	(0.5)	60.6	(0.7)	56.2	(0.4)
3+	16.8	(0.4)	39.4	(0.7)	24.4	(0.4)
Total (n)	(9,287)		(4,738)		(14,025)	

[1] A total of 9,282 respondents from 7,693 HUs participated in the main survey. An additional 4,733 eligible people lived in the 7,693 HUs. The locked building weight (WT1.1) was used in calculating the distributions in the table, converting the 9,282 respondents into a weighted total of 9,287 and the 4,733 other eligible HU residents into a weighted total of 4,738, for a total of 14,025.

part weight was applied as a preliminary step to these short-form respondents aimed at making them as representative as possible of all residents of non-respondent HUs. As the short-form respondents completed the disorder screening section as well as the demographic sections, it was possible to use diagnostic stem questions in addition to individual-level demographic variables as predictors in comparing the main survey respondents with the initial non-respondents. Prior to carrying out the stepwise logistic regression analysis, the main survey respondents were weighted to a sum of weights of 14,025 to represent all eligible residents of the respondent HUs, while the initial non-respondents were weighted to a sum of weights of 6,302 to represent all eligible residents of non-respondent HUs. In both cases, these sums were obtained from HU listings.

WT1.3 is a very important weight because the logistic regression equation on which it is based evaluates whether there is a significant bias in the prevalence of mental disorders in the main sample. We consequently consider the construction of WT1.3 in some detail in this part of the paper. Four stages of model fitting were used to create the logistic regression equation on which WT1.3 is based. These stages sequentially evaluated the effects of geographic variables, individual-level demographic variables, census small area aggregate variables, and screening measures of mental disorders. Results of the first three stages are presented in Table 4. The first stage (Model 1) examined segment-level differences between respondent and non-respondent HUs in Census region and Census urbanicity. Results in the first two columns of Table 4 show that respondent HUs differ meaningfully from non-respondents HUs in urbanicity ($\chi^2_7 = 106.0$, p = 0.00) and region ($\chi^2_3 = 6.6$, p = 0.09). The odds-ratios (ORs) for urbanicity and region show that the sample over-represents non-metropolitan counties, the Midwest, and the South.

The second stage (Model 2) added basic individual-level demographic variables to the equation – age, sex, race-ethnicity, marital status, education, employment status, and household size. As shown in Table 4, household size ($\chi^2_3 = 60.4$, p = 0.00) and marital status ($\chi^2_2 = 9.2$, p = 0.01) were the only significant predictors in this set, with respondents significantly less likely than non-respondents to live in houses with one to three eligible residents and to be never married or previously married.

The third stage (Model 3) was based on a forward stepwise logistic regression analysis that searched for 2000 census block group (BG) aggregate measures that significantly discriminate between respondents and non-respondents. A wide range of variables was used to describe BG characteristics, including percentage variables (for example, the percentage of adults in the BG living in poverty, the percentage living alone) and mean variables (for example, the average family income of HUs in the BG, the average number of adults per HU). In cases where a sample segment crossed BG boundaries, weighted averages across these units were calculated.

Five aggregate variables were found to be significant predictors in the third stage of analysis. All were discretized in elaborations of the basic logistic regression equations in order to study their functional form in distinguishing respondents from non-respondents (0.05 level, two-sided tests). Dichotomous classification was found to be appropriate to characterize the functional form of four predictors. A trichotomous specification was required for the fifth predictor. Results, presented in Table 4, show that respondents were significantly more likely than non-respondents to live in areas with a low proportion of people over the age of 65, a low proportion of foreign-speakers, a high proportion of non-Hispanic whites, a high proportion of never-married people, and high proportions of people not in the labour force.

In the fourth stage of estimating the non-response bias equation, positive responses to diagnostic stem questions in the screening section of the interview were added to the predictors in Model 3. Stem questions were included for mood disturbance (dysphoria, euphoria, irritability), anxiety (persistent worry, panic, agoraphobia, specific fear, social fear, childhood and adult separation anxiety), substance problems (alcohol, drugs, nicotine), and impulse control problems (oppositional-defiant, conduct disorder, intermittent explosive, attentional, and hyperactive). As shown in Table 5, none of these variables alone was either a statistically or substantively significant predictor in the fourth stage of the analysis. These variables are significant as a group ($\chi^2_{20} = 38.1$, p = 0.010), though, even though none of the predictors in the equation is individually significant. We also considered more complex

Table 4. Geographic and socio-demographic predictors of response versus non-response based on the comparison of main survey respondents (n = 9,282) versus short-form survey respondents (n = 475)[1]

	Model 1			Model 2			Model 3			
	OR	(95% CI)	χ^2	OR	(95% CI)	χ^2	OR	(95% CI)	χ^2	d.f.
Region										
Midwest	1.7	(1.1–2.6)	6.6	1.6	(1.0–2.5)	5.3	1.4	(0.9–2.1)	2.8	3
South	1.8	(1.1–2.9)		1.6	(1.0–2.6)		1.3	(0.8–2.2)		
West	1.3	(0.7–2.5)		1.4	(0.7–2.6)		1.3	(0.7–2.4)		
Northeast	1.0	–								
County urbanicity[2]										
Central counties of metro areas of 1+million	1.0	–		1.0	–		1.0	–		
Fringe counties of metro a reas 1+ million	1.0	(0.5–2.1)	106.0*	1.2	(0.5–2.4)	87.5*	0.8	(0.4–1.5)	23.2*	7
Central and fringe counties of metro areas of 250,000 –1 million	1.5	(0.9–2.5)		1.6	(1.0–2.7)		1.5	(1.0–2.5)		
Central and fringe counties of metro areas of less than 250,000	1.5	(0.8–2.9)		1.6	(0.9–2.9)		1.3	(0.7–2.5)		
Non-metro counties of 20,000 or more, adjacent to a metro area	1.9	(0.9–4.2)		2.1	(1.0–4.5)		2.0	(1.1–4.0)		
Non-metro counties of 20,000 or more, not adjacent to a metro area	6.9	(4.2–11.1)		6.9	(4.1–11.8)		2.8	(1.4–5.4)		
Non-metro counties of 2,500 –19,999	1.7	(0.8–3.7)		1.9	(0.9–4.0)		1.6	(0.8–3.5)		
Non-metro counties of less than 2,500	3.1	(1.8–5.5)		3.4	(2.0–5.8)		2.2	(1.2–4.3)		
Age										
18–29				1.0	–		1.0	–		
30–44				0.9	(0.6–1.4)	2.3	1.0	(0.7–1.5)	2.9	3
45–59				0.8	(0.6–1.2)		0.9	(0.6–1.4)		
60+				1.1	(0.6–1.9)		1.3	(0.8–2.3)		
Sex										
Female				1.0	(0.9–1.3)	–	1.0	(0.9–1.3)	–	
Male				1.0	–		1.0	–		
Race										
Non-Hispanic White				1.0	–		1.0	–		
Non-Hispanic Black				1.5	(1.0–2.2)	3.9	1.7	(1.1–2.6)	7.6	3
Hispanic				1.1	(0.7–1.7)		1.4	(0.9–2.0)		
Other				0.9	(0.5–1.5)		0.9	(0.5–1.5)		
Marital status										
Married[3]				1.0	–		1.0	–		
Never married				1.6	(1.1–2.4)	9.2*	1.7	(1.1–2.5)	9.1*	2
Separated/widowed/divorced				1.8	(1.2–2.8)		1.8	(1.2–2.7)		
Education										
0–11				1.0	–		1.0	–		
12				0.9	(0.6–1.4)	1.8	0.9	(0.6–1.4)	3.2	3
13–15				1.0	(0.7–1.4)	1.0		(0.7–1.4)		
16+				1.1	(0.8–1.6)	1.2		(0.8–1.7)		

Table 4. contd.

	Model 1		Model 2		Model 3		
	OR (95% CI)	χ^2	OR (95% CI)	χ^2	OR (95% CI)	χ^2	d.f.
Employment status							
Employed			1.0 –		1.0 –		
Homemaker			1.3 (0.8–2.0)	2.0	1.4 (0.9–2.1)	2.8	4
Retired			1.1 (0.6–1.8)		1.1 (0.7–1.8)		
Student			0.9 (0.4–1.8)		0.9 (0.4–1.8)		
Other			1.1 (0.6–1.9)		1.1 (0.6–2.0)		
Number of eligible household residents							
1			0.7 (0.4–1.1)	60.4*	0.6 (0.4–1.0)	72.0*	3
2			1.9 (1.1–3.5)		1.8 (1.0–3.3)		
3			2.7 (1.4–5.2)		2.6 (1.3–5.0)		
4+			1.0 –		1.0 –		
Block group-level socio-demographics[4]							
% Unable to speak English D1-6					1.9 (1.4–2.6)		
% Never married D5-10					1.5 (1.1–2.2)		
% Not in labor force D1-3					0.6 (0.4–0.9)		
% Age 65+ D1-3					2.7 (1.6–4.5)		
% Age 65+ D4-9					1.5 (0.9–2.3)		
% Non-Hispanic White D1					0.4 (0.2–0.6)		

* Significant at the 0.05 level, two-sided test

[1] Based on logistic regression equations in which main survey respondents were coded 1 and short-form survey respondents (excluding secondary respondents in HUs where the primary respondent participated in the main survey) were coded 0 on a dichotomous dependent variable. Short-form respondents, in addition, were weighted to be representative of all non-respondents using methods described in the text.

[2] Metropolitan counties are defined as either central or fringe counties of census metropolitan statistical areas. Non-metropolitan counties are defined residually as all counties that are not in census metropolitan statistical areas. For more details on the Census Bureau definition of metropolitan statistical areas, see http://www.census.gov/population/www/estimates/metrodef.html.

[3] Includes co-habitors.

[4] The percentages in each county were converted to percentiles based on weighted (by population size) county-level distributions and converted into deciles (D) of the weighted percentile distributions. The coefficients from preliminary regression analyses included nine dummies for the deciles of each substantive variable. These coefficients were examined in order to choose the optimal way to collapse the deciles. A dichotomous coding was optimal for four of the five substantive variables and a trichotomous coding for the fifth.

specifications that included disorder clusters and counts, but none yielded any more compelling evidence than in Table 5 for significant differences between respondents and initial non-respondents in the prevalence of these diagnostic stem questions.

Based on these results, the final non-response adjustment weight was based on Model 3. Each of the 9,282 main study respondents was assigned a predicted probability of response based on this equation. The non-response adjustment weight was defined as the multiplicative inverse of that predicted probability. This weight was then normed to sum to 9,282. This normed weight defines WT1.3.

The 9,282 cases were then weighted by the joint product of WT1.1, WT1.2 and WT1.3. A fourth weight (WT1.4) was then created to adjust for variation between the joint distribution of several socio-demographic variables in this weighted sample compared to the March 2002 Current Population Survey (CPS) data. This post-stratification weight

was based on comparisons of age, sex, race-ethnicity, education, marital status, region, and urbanicity in the weighted (by the joint product of WT1.1, WT1.2, and WT1.3) sample and the 2002 CPS sample for persons ages 18+ in the continental US. As the full cross-classification of these seven variables, even using coarse categories, has more cells than there are respondents in the sample, a smoothing method was used to fit only the two-way marginals by estimating a logistic regression equation that combined the weighted 9,282 respondents (coded 1 on the dichotomous dependent variable) with a synthetic dataset representative of the individual-level 2002 CPS data for the population ages 18+ (coded 0 on the dependent variable). The seven socio-demographic variables and their two-way interactions were the predictors. The predicted probability of being in the NCS-R sample rather than in the synthetic CPS population sample was defined for each respondent (p_{Nn}) based on this equation. The ratio $p_{Nn} / (1 - p_{Nn})$ was then calculated for each respondent. The sum of these ratios across the 9,282 respondents was normed to sum to 9,282. These normed ratios define WT1.4.

The four-way product of weights WT1.1 through WT1.4 was then created and applied to the 9,282 respondent cases. This defines the consolidated Part I weight based on the NR approach. An additional weight (WT1.5) is needed, though, to analyse data in the Part II sample. This is because, as noted earlier, the Part I sample was divided into three strata that differed in their probabilities of selection into Part II (100% in stratum I, 50% in stratum II, and 25% in stratum III). The proportions who actually completed Part II differ from these targets: 99.2% of the 4,088 respondents in stratum I (n = 4,054), 53.4% of the 1,555 respondents in stratum II (n = 830), and 22.2% of the 3,639 respondents in stratum III (n = 808), for a total Part II sample of 5,692 respondents. These differences from the target proportions occurred because some respondents refused to complete Part II (approximately 1% of designated cases in each stratum) and because the random selection procedure for respondents implemented in the CAPI program had some random error. The latter accounts for the fact that the number of Part II respondents in stratum II is larger than the target proportion.

We could have generated WT1.5 as the simple inverse of the weighted (by the consolidated Part I weight) proportion of Part I respondents in the stratum who completed Part II. However, in order to check for the possibility of systematic bias we estimated separate stepwise logistic regression equations in each stratum to predict participation in Part II from Part I responses. Only a handful of variables, all of them demographic, were found to be significant predictors in these equations. Each Part II respondent was assigned the inverse of his or her predicted probability of participation in Part II from the final within-stratum equation. These values were then normed to have a sum equal to the sum of the Part I weights in the full Part I sample in the stratum. These normed values were then summed across the entire Part II sample of 5,692 cases and renormed to have a sum of weights of 5,692. These renormed values define WT1.5. WT1.5 was then multiplied by the consolidated Part I weight to create the consolidated Part II weight.

Comparison of the weighted and unweighted distributions of the Part I and Part II samples with 2002 CPS population distributions provides information on the effects of weighting. As shown in Table 6, the unweighted Part I samples overrepresented racial minorities, females, residents of the Midwest, people with 13+ years of education, and residents of metropolitan areas. All of these biases were corrected with the consolidated Part I weight. Biases in the unweighted Part II sample were similar, although more extreme biases were found than in the Part I sample with regard to the over-representation of females, young adults (ages 18–34), and residents of Metropolitan areas. All of these biases were corrected with the consolidated Part II weight.

Weighting short-form respondents to be representative of all non-respondents

We noted in the last subsection that WT1.3 relies on a two-part weighting scheme that was applied to the short-form respondents before they were compared with the main survey respondents. This was done in order to increase the extent to which the short-form respondents represent all main survey non-respondents. The first part of this weighting scheme compared the 399 HUs in which short-form interviews were completed with all other non-respondent HUs on the same aggregate 2000 census block group (BG) data as those used in the third stage (Model 3)

Table 5. Diagnostic stem question predictors of response versus non-response based on the comparison of main survey respondents (n = 9,282) versus short-form survey respondents in non-respondent households (n = 475)[1]

	Bivariate		Multivariate	
	OR	(95% CI)	OR	(95% CI)
I. Mood disturbance				
Dysphoria	0.9	(0.7–1.2)	1.0	(0.8–1.4)
Euphoria	0.8	(0.6–1.1)	0.9	(0.7–1.2)
Irritability	0.8	(0.6–1.0)	0.8	(0.6–1.1)
Extreme irritability	0.8	(0.6–1.2)	0.9	(0.7–1.3)
II. Anxiety				
Persistent worry	0.9	(0.7–1.1)	0.9	(0.7–1.2)
Panic	0.8	(0.6–1.1)	0.8	(0.6–1.1)
Agoraphobic fear	0.9	(0.7–1.2)	0.9	(0.7–1.2)
Specific fear	1.0	(0.7–1.3)	1.0	(0.8–1.3)
Social fear	1.0	(0.7–1.3)	1.0	(0.7–1.4)
Separation anxiety				
– Childhood only	1.2	(0.8–1.8)	1.3	(0.9–2.0)
– Adult only	1.1	(0.7–1.7)	1.3	(0.8–2.0)
– Both	1.3	(0.7–2.4)	1.6	(0.8–2.9)
III. Substance problems				
Nicotine	1.2	(0.9–1.5)	1.2	(0.9–1.6)
Alcohol-drugs	1.1	(0.8–1.5)	1.1	(0.8–1.4)
IV. Impulse-control problems				
Oppositional-defiant	1.0	(0.8–1.3)	1.0	(0.8–1.4)
Conduct	0.9	(0.7–1.1)	0.9	(0.6–1.1)
Intermittent explosive	1.1	(0.9–1.4)	1.2	(1.0–1.6)
Attention-hyperactive problems				
– Attention only	0.9	(0.6–1.3)	0.9	(0.7–1.3)
– Hyperactive only	1.1	(0.7–1.9)	1.1	(0.6–2.0)
– Both	1.0	(0.6–0.6)	1.0	(0.6–1.6)

[1] Based on logistic regression equations in which respondents in the main survey were coded 1 and short-form survey respondents (excluding secondary respondents in HUs where the primary respondent participated in the main survey) were coded 0 on a dichotomous dependent variable. Short-form respondents, in addition, were weighted to be representative of the residents of all non-respondent HUs using methods described in the text. All equations controlled for the predictors in Model 3 in Table 4.

of the non-response adjustment model (Table 4). Stepwise logistic regression was used to select significant predictors of HU-level participation versus non-participation. The final set of significant predictors included region, urbanicity and household size. The participating HUs were then weighted by the inverse of their predicted probability of participation to adjust for segment-level variation in response. This weight was then normed to have a sum of weights equal to 3,258, the weighted total number of non-respondent HUs in the sample.

With this first weight applied separately to each of the 773 residents of the 399 participating short-form HUs, a comparison of the short-form respondents in these 399 HUs with the remaining eligible residents of the same HUs was made based on the cross-classification of listing information (age and sex of each eligible HU resident and number of eligible residents in each HU). A second weight was then developed that was equivalent to WT1.2 in adjusting the short-form respondents to be representative of all 773 eligible residents of these HUs. As the 399 HUs

were weighted to represent all 3,258 non-respondent HUs in the entire sample, the sum of weights across the 773 eligible residents of the 399 participating HUs was equal to 6,302. The latter is our best estimate of the number of eligible residents of all non-respondent HUs. The two weights were then multiplied together and the sum of the product normed to equal 6,302.

This weighting scheme was then used to compare the 9,282 main sample respondents (with a sum of weights of 14,025) with the short-form respondents who lived in HUs where the primary pre-designated respondent did not complete an interview in the main survey (n = 475, with a sum of weights of 6,302) to create a non-response adjustment weight. Note that no weighting of the 9,282 cases based on aggregate Census small area data was made before comparison with the short-form respondents. The reason is that the 9,282 respondents represented only their own HUs in this comparison, whereas the short-form respondents were made to represent all non-respondents rather than only non-respondents in their 399 HUs. Note that the secondary short-form respondents from HUs where the primary respondent completed an interview in the main survey were excluded from this exercise.

The multiple-imputation approach

Five weights were used in the MI approach: a locked building subsampling weight (WT2.1), a weight that adjusts for geographic variation in response rate (WT2.2), a within-household probability of selection weight (WT2.3), a post-stratification weight (WT2.4), and a Part II selection weight (WT2.5). The joint product of the first four weights is the consolidated weight used to analyse data from the Part I sample in the MI approach (n = 9,836), while the joint product of all five weights is the consolidated weight used to analyse data from the Part II sample (n = 6,279). When the data to be analysed include some variables assessed in Part I and other variables assessed in Part II, the Part II sample is used. This section of the paper briefly reviews each of these five weights.

The locked building weight (WT2.1) is identical to WT1.1. The non-response adjustment weight (WT2.2), in comparison, differs from the similar weight in the NR approach because the short-form cases, which were treated as non-respondents in the NR approach, are treated as respondents in the MI approach. This means that non-response adjustment in the MI approach must be based entirely on comparisons of small area census data between respondent HUs and non-respondent HUs. As a result, the MI non-response adjustment weight (WT2.1) adjusts for segment-level variation in the household response rate. The simple way of doing this would have been to weight each participating HU by the inverse of the response rate in its segment. However, this would have created a problem for the small number of segments in which no interviews were obtained and it would also have introduced an unnecessarily large amount of variation in the weight because of the small level of aggregation. As an alternative, then, the same approach was used as in developing the non-response adjustment weight (Table 4). Specifically, stepwise logistic regression analysis was used to calculate the predicted probability of participation of each HU that actually participated in the survey based on a comparison of BG demographic data from the 2000 census. The inverse of this predicted probability was then used as the non-response adjustment weight.

The product of WT2.1 and WT2.2 was then computed for each participating HU and this product was normed so that it summed to 8,092, the number of participating HUs in the entire survey. A third weight (WT2.3) was then created at the respondent level to adjust for variation in within-household probability of selection. As in the NR approach, this was done by creating a separate weighted (by the product of WT2.1 and WT2.2) observational record for each of the 14,788 eligible HU residents in the 8,092 participating HUs, with each case assigned his or her HU weight. The three variables in the household listing (respondent age and sex and size of household) were included in the individual-level records. The weighted distributions of these three variables were then calculated separately for the 9,836 respondents and the remaining 4,952 eligible residents in the same HUs. As with the NR approach, these distributions were found to differ meaningfully, with the greatest difference being for size of household. As in the non-response adjustment approach, the weighted three-way cross-classifications of age, sex, and household size was computed separately for the 9,836 respondents and for all 14,788 residents of the participating HUs, the

Table 6. Comparison of unweighted and weighted Part I and Part II NCS-R respondents with socio-demographic information from the March 2002 Current Population Survey (CPS)

	NCS-R Part I				NCS-R Part II				CPS
	Unweighted		Weighted		Unweighted		Weighted		
	%	(SE)	%	(SE)	%	(SE)	%	(SE)	%
Race									
Non-Hispanic White	72.1	(1.8)	73.2	(1.9)	73.4	(1.7)	72.8	(1.8)	73.0
Non-Hispanic Black	13.3	(1.1)	11.6	(1.1)	12.6	(1.0)	12.4	(1.0)	11.6
Hispanic	9.5	(1.0)	10.8	(1.0)	9.3	(1.0)	11.1	(1.2)	11.0
Other	5.1	(0.7)	4.4	(0.4)	4.7	(0.7)	3.8	(0.4)	4.4
Sex									
Male	44.6	(0.5)	47.9	(0.5)	41.8	(0.6)	47.0	(1.0)	48.0
Female	55.4	(0.5)	52.1	(0.5)	58.2	(0.6)	53.0	(1.0)	52.0
Region									
Northeast	18.4	(1.8)	19.3	(3.3)	18.3	(1.8)	18.8	(3.0)	19.2
Midwest	26.7	(1.7)	23.2	(1.8)	27.5	(1.7)	23.5	(1.8)	22.9
South	34.5	(1.0)	35.8	(2.0)	32.5	(1.2)	35.6	(1.9)	35.8
West	20.5	(0.6)	21.7	(2.2)	21.7	(1.0)	22.1	(1.9)	22.1
Age									
18–34	32.7	(1.0)	31.5	(1.2)	34.0	(1.1)	31.5	(1.2)	31.2
35–49	30.9	(0.6)	31.5	(0.8)	32.2	(0.7)	30.9	(1.0)	31.7
50–64	20.7	(0.5)	21.1	(0.6)	21.3	(0.7)	20.9	(1.0)	21.1
65+	15.7	(0.5)	16.0	(0.5)	12.5	(0.6)	16.7	(1.0)	16.1
Education									
< 12 Years	44.9	(1.3)	48.4	(1.7)	45.0	(1.4)	49.3	(1.5)	48.7
13+ Years	55.1	(1.3)	51.6	(1.7)	55.0	(1.4)	50.7	(1.5)	51.3
Marital status									
Married	57.3	(0.9)	55.8	(1.1)	56.9	(1.0)	55.9	(1.2)	56.0
Not married	42.7	(0.9)	44.2	(1.1)	43.1	(1.0)	44.1	(1.2)	44.0
Urbanicity status									
Metropolitan	75.6	(3.0)	67.5	(5.4)	76.9	(3.1)	68.2	(5.0)	67.3
Non-metro	24.4	(3.0)	32.5	(5.4)	23.1	(3.1)	31.8	(5.0)	32.7

ratio of the number of cases in the latter versus the former was calculated within cells, and these ratios were applied to each of the 9836 respondents. The sum of the ratios was then normed to sum to 9836. These normed ratios define WT2.3.

The three-way product of WT2.1, WT2.2, and WT2.3 was computed for each respondent and this product was normed so that it summed to 9,836. A fourth weight (WT2.4) was then created to adjust for variation between the joint distribution of several socio-demographic variables in this weighted sample compared to the 2000 Census. This post-stratification weight was constructed in exactly the same way as in the NR approach (WT1.4). As a result, a description of this method will not be repeated here.

The four-way product of WT2.1-WT2.4 was then computed to define the consolidated Part I weight based on the MI approach. An additional weight

(WT2.5) was then constructed to analyse data included in Part II of the interview (n = 6,279). WT2.5 was constructed using the same three strata and the same methods as WT1.5 in the NR approach. The product of the Part I MI weight and WT2.5 was computed for each Part I respondent and this product was normed so that it summed to 6,279. This normed product defines the consolidated Part II weight based on the MI approach.

Item-level imputation

The amount of item-missing data is much smaller in NCS-R than in a paper-and-pencil survey of comparable complexity because the CAPI program eliminated interviewer skip errors. However, respondents occasionally refused to answer some questions and sometimes reported that they did not know the answers to other questions. As in many surveys, the highest item-level non-response was for the questions about earnings and family income. A regression-based multiple imputation approach was used to impute missing values for these income data using information about age, sex, education, employment status, and occupation of household residents. Conservative rational imputation was used for other items that have item-missing cases. For example, in the life events section, the small numbers of missing values were recoded as negative responses. In the case of missing items in a psychometric scale, respondents were assigned a total scale score based on partial values using mean standardized scores on the remaining items in the scale. The MI approach could have been used to take these item-level imputations into account in evaluating statistical significance, but we did not do so because the amount of item-missing data was quite small.

Design-based estimation

Weighting and clustering introduce imprecision into descriptive statistics. Conventional methods of estimating significance, which assume a simple random sample, do not take this imprecision into consideration. As a result, special design-based methods of estimating SEs and significance tests are being used in the analysis of the NCS-R data. The Taylor series linearization method is the main approach used here (Wolter, 1985), although we also use the more computationally intensive method of jackknife repeated replications (JRR) for some applications (Kish and Frankel, 1974). JRR is used for applications where a convenient software application using the Taylor series method is not readily available and for highly non-linear estimation problems in which the linearization of the Taylor series method might be problematic.

As noted earlier in the paper, the NCS-R is based on 84 PSUs and pseudo-PSUs. These 84 were divided into 42 matched pairs for purposes of estimating design effects. Design-based estimation was then carried out using 42 sampling strata and two sampling error calculation units (SECUs) per stratum. Although the decision to work with only two SECUs per stratum was arbitrary from the perspective of the Taylor series estimation method, it was necessary for using JRR.

Although the effects of weighting on clustering can be described in a number of ways, a particularly convenient approach is to calculate a statistic known as the design effect (DE) (Kish and Kish, 1965) for a number of variables of interest. The DE is the square of the ratio of the design-based SE of a descriptive statistic divided by the simple random sample SE. The DE can be interpreted as the approximate proportional increase in the sample size that would be required to increase the precision of the design-based estimate to the precision of an estimate based on a simple random sample of the same size.

Design effects

Design effects due to clustering are usually a good deal larger in estimating means and other first-order statistics than more complex statistics. The reason for this is that the number of respondents having the same characteristics in the same SECU of a single stratum becomes smaller and smaller as the statistics become more complex. This leads to a reduction in the effects of clustering in the estimation of DE. Design effects due to weighting are also usually somewhat smaller for multivariate than bivariate descriptive statistics because DEs are due not only to the variance in the weights but also to the strength of the association between the weights and the substantive variables under consideration. Means typically have higher DEs than other statistics, so evaluations of DEs typically focus on the estimation of means. We do the same here, considering prevalence estimates for a number of the mental disorders assessed in the NCS-R. Five

dichotomous measures of lifetime prevalence of DSM-IV disorders included in the Part I sample were included in the evaluation of Part I DEs. These include major depressive disorder (MDD), bipolar disorder (BPD), generalized anxiety disorder (GAD), panic disorder (PD), and intermittent explosive disorder (IED). The DEs for those estimates are 1.6 for MDD, 1.4 for BPD, 1.6 for GAD, 0.9 for PD, and 1.9 for IED.

As described earlier, only 5,692 of the 9,282 Part I respondents were administered Part II. If Part II respondents were a simple random sample of the Part I sample, the expected design effect of the former compared to the latter would be approximately 1.6 (9,282/5,692). However, we expected the design effects for most prevalence estimates to be considerably lower than this due to the fact that Part II respondents include all Part I respondents who met criteria for any of the DSM-IV disorders assessed in Part I plus other respondents selected proportional to household size. In order to evaluate whether this expectation is borne out in the data, we calculated the DEs for the same five prevalence estimates as in the last paragraph for the Part II sample and then computed the ratios of the DEs based on the Part II versus Part I samples. The results are as follows: 1.56 for MDD, 0.92 for BPD, 1.12 for GAD, 0.62 for PD, and 0.80 for IED.

The fact that these estimates, with the exception of MDD, are all considerably smaller than the 1.6 we would have expected based on using simple random sampling to select respondents into Part II demonstrates that the disproportional case-control sampling approach used to select the Part II sample increased the efficiency of the Part II sample. Indeed, the fact that the design effect is close to 1.0 in a number of cases means that this sub-sampling approach was able to retain the vast majority of the precision in the full Part I sample with only slightly more than 60% of the Part I sample. The exception is MDD, by far the most prevalent of the disorders considered here, with a ratio of controls to cases in the Part II sample of less than 4:1. As a result of this comparatively low ratio, a meaningful amount of precision was lost by subsampling controls into Part II. However, this is a self-limiting problem in that the high prevalence of MDD means that we have greater precision for studying this condition than the less common conditions.

Trimming weights to reduce design effects

Estimates of DE can be sensitive to extreme weights. Weight trimming of various sorts is often used to reduce this sensitivity. A small amount of trimming was built into the construction of two of the weighting steps described above. First, the within probability of selection weights (WT1.2, WT2.3) were trimmed by combining respondents in households with three or more eligible residents into a single weighting stratum. This means that no attempt was made, for example, to assign a weight of 8.0 to the rare respondent who lived in a household with eight eligible residents. Instead, respondents living in HUs with three or more eligible residents were combined and the weight to adjust for their under-sampling was distributed equally among all of them. Second, trimming was used in the non-response adjustment weights (WT1.3, WT2.2) to distribute the weights at the tails of these distributions (the upper and lower 2% of each distribution) equally across all cases at these tails.

We also empirically investigated the implications of trimming the final consolidated Part I weight in the non-response adjustment approach. Although weight trimming usually reduces the variance of weights, and in this way improves the precision of estimates and the statistical power of tests, trimming can also lead to bias in the estimates. If the reduction in variance created due to added efficiency exceeds the increase in variance due to bias, the trimming is helpful overall. Weighting is unhelpful, in comparison, if the opposite occurs. It is possible to study this trade-off between bias and efficiency empirically in order to evaluate alternative weight trimming schemes by making use of the equality:

$$MSE_{Yp} = B_{Yp}^2 + Var(Y_p), \quad (1a)$$

$$= (B_{Yp})^2 - Var(B_{Yp}), + Var(Y_p), \quad (1b)$$

where $MSE_{Y}p$ is the estimated mean squared error of the prevalence of outcome variable Y at trimming point p, $B_{Y}p$ is the estimated bias of that prevalence estimate, $Var(B_{Y}p)$ is the estimated variance of $B_{Y}p$, and $Var(Y_p)$ is the estimated variance of Y_p.

Each of the three elements in equation (1b) can be estimated empirically for any value of p, making it possible to calculate MSE across a range of trimming points and to determine in this way the

trimming point that minimizes MSE. The first element, $(B_{Yp})^2$, was estimated directly as $(Yp - Y0)^2$, where Y_0 represents the weighted prevalence estimate of Y based on the untrimmed weight. The other two elements in equation (1b) were estimated using a pseudo-replicate method in which 84 separate estimates were generated for Y_p at each value of p (Zaslavsky, Schenker and Belin, 2001). The number 84 is based on the fact that the NCS-R sample design has 42 strata, each with two SECUs, for a total of 84 stratum-SECU combinations. The separate estimates were obtained by sequentially modifying the sample and then generating an estimate based on that modified sample. The modification consisted of removing all cases from one SECU and then weighting the cases in the remaining SECU in the same stratum to have a sum of weights equal to the original sum of weights in that stratum. If we define Y_p as the weighted estimate of Y at trimming point p in the total sample and we define $Y_{p(sn)}$ as the weighted estimate at the same trimming point in the sample that deletes SECU n (n = 1,2) of stratum s (s = 1–42), then $Var(Y_p)$ can be estimated as

$$Var(Y_p) = SUM_S[(Y_{p(s1)} - Y_p)^2 + (Y_{p(s2)} - Yp)^2]/2. \quad (2)$$

$Var(B_{Yp})$ was estimated in the same fashion by replacing $Y_{p(sn)}$ in Eq. (2) with $B_{Yp(sn)} = Y_{p(sn)} - Y_{0(sn)}$ and replacing Y_p with $B_{Yp}(sn) = Y_p - Y_0$.

This analysis compared the design-based MSE of lifetime prevalence estimates for the same five outcomes considered in the last subsection plus five comparable outcomes for 12-month prevalence using the consolidated Part I weight and 10 successively more severely trimmed versions of these weights in which between 1% and 10% of cases were trimmed at each tail of the distribution. Trimming consisted of distributing the weights at each of these tails equally across all cases in that tail. As results were very similar across the outcomes, averages of MSE_{Yp}, B_Y^2, and $Var(Y_p)$ were calculated at each trimming point. The mean of MSE_{Y0} across the outcomes was then arbitrarily set at 100 and all other values were defined in relation to that mean for ease of interpretation.

Summary results are presented in Table 8. Three patterns are immediately apparent. First, the equality in Eq. (1a)–(1b) does not hold in the table because the results are based on means across a number of equations that were calculated on raw data. Second, the decrease in the variance of the prevalence is very modest with successively higher percentages of trimming (approximately 2.5%) and only has effects at the highest levels of trimming (9–10%). Third, the variance due to bias increases from a low of approximately 0.5% at 1% trimming to approximately 14% at 7% trimming and remains fairly constant in the range 7–10%. Because this increase is greater than the decrease in the variance of Y due to trimming, MSE increases as a result of trimming. Based on this result, we did not trim the consolidated weight.

Model-based versus design-based estimation

Although analysis of the NCS-R data will largely involve design-based methods, we will also explore the possibility of using model-based estimation of risk factors. This approach uses weights as predictor variables in unweighted analyses. It is also possible to carry out model-based analyses of the effects of clustering by including dummy variables for segments or SECUs as predictor variables. In cases where the weights or clusters significantly interact with substantive predictors, it is necessary to include these interactions in prediction equations and to recognize that the effects of the substantive predictors vary depending on the determinants of the weights and/or on geography. Insights into interactions that involve weights can be obtained by substituting the substantive variables on which the weights are based for the weights themselves in expanded prediction equations. Similar insights into interactions that involve clusters can be obtained by including small area census data in expanded prediction equations using random effects models to interpret the modifying effects of the clusters.

In cases where the weight or cluster variables are found to be significant predictors but not to have significant interactions with substantive predictors, the coefficients associated with substantive predictors can be interpreted as they would if the analyses were based on weighted data. The reason for doing the extra work involved in the model-based estimation in such a situation is that the SE of the substantive predictors are likely to be smaller, perhaps substantially so, than in analyses based on

weighted data. Finally, in cases where the weights are neither significant predictors nor significant modifiers of the effects of the substantive predictors, the weights can be ignored entirely, leading to a similar reduction in the estimated SE of the coefficients associated with substantive predictors.

As model-based analysis is very labour intensive, our use of this approach in the NCS-R will be reserved for the most important areas of investigation in which concerns exist about the statistical power and sensitivity of parameter estimates. A very preliminary exploration of likely complexities in implementing such an approach, based loosely on the approach proposed by DuMouchel and Duncan (1983) was carried out by examining the associations of weights WT1.2 through WT1.4 in predicting lifetime prevalence of the same five disorders considered in the last section. In addition to evaluating the main effects of the weights, we also evaluated the statistical significance of interactions between the weights and several socio-demographic variables (age, sex, education, race-ethnicity) in predicting the outcomes.

Summary results are presented in Table 8. The χ^2 values of main effects are shown in Part I for clusters (83 dummy predictor variables) and weights (three separate continuous variables) and in Part II for selected socio-demographic variables (age, sex, education, race-ethnicity). In Part I, none of the main effects of either clusters or WT1.3 is significant at the 0.05 level, while WT1.2 is significant in four equations and WT1.4 in one equation. Inspection of the coefficients associated with the significant weights shows that the effects of WT1.2 are due to people who live alone (who are overrepresented in the unweighted sample) having higher prevalence than other respondents, while the effect of WT1.4 is due to women (who are overrepresented in the unweighted sample) having a higher prevalence of MDD than men. In Part II, age is significant in all five equations (due to high prevalence in the late middle-aged age group), sex is significant in four of the five (due to women having higher prevalence than men), education in one (due to high prevalence of BPD among people with the highest education), and race-ethnicity is significant in three of the five (due to non-Hispanic Whites having higher prevalence than non-Hispanic Blacks or Hispanics).

Parts III-V show χ^2 values of interactions between the socio-demographic variables and the weights. Using 0.05-level tests, 10% of the interactions involving WT1.2, 10% involving WT1.3, and 30% involving WT1.4 are statistically significant. Of the 10 significant interactions, six involve age and four involve education. Inspection of the coefficients shows that the age interactions are due to greater

Table 7. The effects of trimming the Part I weight in the range 1–10% on the mean squared error of prevalence estimates[1]

% of trimming at each tail	Bias (BYp2)	Inefficiency [Var(Yp)]	Mean squared error
0	0.0	100.0	100.0
1	0.5	100.6	101.3
2	0.8	102.5	102.9
3	2.1	100.6	102.2
4	3.2	99.1	101.6
5	6.0	98.7	103.8
6	8.5	100.3	108.4
7	14.3	100.3	114.4
8	13.2	99.7	111.7
9	12.9	97.5	108.7
10	12.8	98.1	109.0

[1] Lifetime and 12-month prevalence estimates for positive screens of broad categories of DSM-IV disorders were included in the evaluation of design effects. These included any lifetime anxiety disorder, mood disorder, impulse-control disorder, substance use disorder, and any of the four kinds of disorder. The Taylor series method was used to estimate each prevalence with the untrimmed Part I weight (Trimming = 0%) and with trimming in the range 1–10% at each tail. Trimming consisted of equally distributing the sum of weights at the tail across all cases at the tail. As results were quite similar for all ten screening measures, results are averaged here. Note that the equality BYp2 + Var(Yp) = mean squared error does not hold here as the averaging was done at the level of absolute bias, SE, and root mean squared error.

effects of age among people who live alone (WT1.2), among people who have a low probability of participating in the survey (WT1.3), and among people who are underrepresented in the sample after adjusting for non-response bias (WT1.4). The education interactions are due to greater effects of education among people who are under-represented in the sample after adjustment for non-response bias (WT1.4).

Three broad conclusions can be drawn from this exercise. The first is that the effects of weights cannot be ignored even in studying very simple associations of the sort considered in Table 8. Some way of dealing with the weights is needed, using either design-based or model-based methods. The second conclusion is that the effects of many substantively important predictors are not modified by weighting. This means that it will often be useful to carry out model-based analysis to investigate whether risk factors of special importance are unaffected by weights. The third conclusion is that the significant modifying effects of weights can often be decomposed to yield substantively meaningful interpretations in unweighted analyses. This third conclusion supports the use of model-based methods to study important risk factors even in cases where weight modification is found to exist.

Overview

This paper has presented an overview of the NCS-R survey design and field procedures. The use of face-to-face interviewing as the data collection mode is consistent with most other major national government-sponsored household surveys. Our ability to manage the complexity of the interview was increased greatly by the use of CAPI rather than PAPI administration. The use of CAPI was the one major change in the fundamental survey conditions introduced into the NCS-R compared to the baseline NCS. As described in the body of the paper, we stopped short of using A-CASI because of concerns about trending disorder prevalence estimates and timing. However, A-CASI is likely to be the most important innovation introduced into the next round of the NCS, which we tentatively plan to field in 2010.

The NCS-R multi-stage area probability sample design was very similar to the NCS design. This was done to facilitate trend comparisons. However, three changes were made in the within-HU selection procedures. First, we interviewed people in the age range 18+ rather than in the NCS age range of 15–54. The exclusion of the 15–17 age range was dictated by the fact that we are carrying out a separate NCS Adolescent survey of 10,000 respondents in the age range 13–17. The inclusion of the age range 55+ was based on the desire to study the entire adult age range. Second, we sampled two respondents in a probability subsample of NCS-R HUs. The NCS, in comparison, had only one respondent in each HU. The implications of this design change can be evaluated by analysing the NCS-R data separately in the subsample of primary respondents, which is identical to the NCS within-HU sample design. Third, we used a much more efficient way of selecting Part I non-cases for participation in Part II of the interview in the NCS-R than in the NCS. While simple random sampling was used to select Part II respondents in the NCS, differential sampling proportional to HU size was used in the NCS-R. Because of this change, the Part II NCS-R sample of 5,692 cases is considerably more efficient than the Part II NCS sample of 5,877 cases.

The 70.9% response rate of primary respondents in the NCS-R is considerably lower than the 80.2% response rate in the NCS a decade earlier. This is part of a consistent trend in survey response rates becoming much lower over the past decade than in the past. Interestingly, though, while the NCS non-response survey, which was very similar in design to the NCS-R short-form survey, found statistically significant underestimation of disorder prevalence among NCS respondents versus non-respondents, no evidence for downward bias was found in the NCS-R short-form survey. This result suggests that the NCS-R might be as representative of the population, or perhaps even more so, with respect to psychopathology as the baseline NCS despite the lower response rate.

The Part I NCS-R interview was very similar in length to Part I of the NCS interview, while the Part II NCS-R interview was longer by an average of about 30 minutes than NCS Part II. This led to a higher proportion of NCS-R than NCS interviews that had to be completed over multiple interview sessions. This difference may have implications for trending. We were mindful of this possibility in designing the NCS-R interview schedule and we

consequently included comparable questions largely in the first half of the interview schedule in order not to change the time burden across the two surveys at the time the comparable questions were asked. The goal here was to remove any potential order bias that might lead to differences in response.

The design-based estimation approach briefly described in this paper is identical to the approach used to analyse the NCS data. Importantly, as the NCS-R PSUs are linked to the NCS PSUs, it will be possible to merge the two surveys and blend strata for purposes of increasing statistical power in carrying out trend comparisons. Although model-based estimation was not used in the NCS, this will be an important feature of the NCS-R analyses as well as of NCS versus NCS-R trend analyses based on the greater focus than in the NCS on risk factor analyses and on concerns about the extent to which risk factor results generalize to segments of the population that might differ in representation across sample weights and clusters.

Finally, it is important to recognize that a rich multi-purpose survey like the NCS-R contains so much information that it will take many years and

Table 8. χ^2 values of the main effects and interactions of design weights with socio-demographic variables in predicting life time Major Depressive Disorder (MDD), Bipolar Disorder I or II (BPD), Generalized Anxiety Disorder (GAD), Panic Disorder (PD), and Intermittent Explosive Disorder (IED) in the NCS-R Part 2 Sample (n = 5,692)**

	df	MDD	BPD	GAD	PD	IED
I. Main effects of clusters and weights						
Strata/SECU variables	83	100.9	79.7	62.7	51.6	72.6
HU selection weight (WT1.2)	1	9.8*	0.03	13.5*	9.7*	6.3*
Non-response weight (WT1.3)	1	0.03	1.4	1.1	0.0	0.0
Post-stratification weight (WT1.4)	1	8.2*	2.7	1.6	0.2	0.1
II. Main effects of demographics						
Age	3	66.3*	55.8*	19.8*	66.6*	24.0*
Education	3	5.2	13.8*	5.1	5.8	7.0
Sex	1	40.7*	0.03	13.6*	40.6*	11.0*
Race	3	14.0*	2.7	8.9*	13.9*	4.5
III. Interactions involved with WT1.2						
HU selection*age	3	9.5*	5.0	5.0	9.6*	2.8
HU selection*education	3	2.7	1.8	6.2	2.8	0.1
HU selection*sex	1	2.3	1.9	7.1	2.3	0.9
HU selection*race	3	1.9	1.4	0.8	1.8	1.3
IV. Interactions involved with WT1.3						
Non-response *age	3	18.3*	2.1	0.9	18.4*	5.7
Non-response *education	3	5.0	3.2	7.1	5.0	1.4
Non-response *sex	1	1.3	0.8	0.7	1.3	0.5
Nonresponse *race	3	7.0	0.8	3.1	1.8	0.8
V. Interactions involved with WT1.4						
Post-stratification *age	3	5.0	6.7	5.7	4.9	8.4*
Post-stratification *education	3	10.8*	4.5	8.0*	10.8*	12.8*
Post-stratification *sex	1	3.4	0.9	2.6	3.3	0.0
Post-stratification *race	3	4.2	10.1*	0.5	4.1	1.9

* Significant at the 0.05 level, two-sided test, using estimation methods that assume the sample is a simple random sample.

**All disorders were defined with diagnostic hierarchy rules. The Taylor series method was used to estimate design effects. Standard errors based on simple random sampling were assumed to have an expectation of $(pq/n)^{1/2}$.

many workers to learn all the survey has to tell us. This is a much greater undertaking than any one research team can hope to achieve. As a result, a public use NCS-R data file is being released for unrestricted use. We hope that this data file will be the source of many dissertations and secondary analyses that expand on the primary analyses being carried out by our research group. The NCS Web site will post information about timing and procedures for obtaining the public data file. We will also offer a series of training courses in the use of the public data file. Information about these courses will be posted on the Web site as the schedule is set. Finally, we plan to offer help in letting users of the public data file know about each other and coordinate their investigations by creating a separate page on the NCS Web site for users to describe the lines of research they are carrying out with the data.

Acknowledgements

The National Comorbidity Survey Replication (NCS-R) is supported by the National Institute of Mental Health (NIMH; U01-MH60220) with supplemental support from the National Institute of Drug Abuse (NIDA), the Substance Abuse and Mental Health Services Administration (SAMHSA), the Robert Wood Johnson Foundation (RWJF; Grant 044708), and the John W. Alden Trust. Collaborating investigators include Ronald C. Kessler (Principal Investigator, Harvard Medical School), Kathleen Merikangas (Co-Principal Investigator, NIMH), James Anthony (Michigan State University), William Eaton (The Johns Hopkins University), Meyer Glantz (NIDA), Doreen Koretz (Harvard University), Jane McLeod (Indiana University), Mark Olfson (Columbia University College of Physicians and Surgeons), Harold Pincus (University of Pittsburgh), Greg Simon (Group Health Cooperative), Michael Von Korff (Group Health Cooperative), Philip Wang (Harvard Medical School), Kenneth Wells (UCLA), Elaine Wethington (Cornell University) and Hans-Ulrich Wittchen (Institute of Clinical Psychology, Technical University Dresden and Max Planck Institute of Psychiatry). The authors appreciate the helpful comments on earlier drafts of Jim Anthony, Doreen Koretz, Kathleen Merikangas, Bedirhan Üstün, Michael von Korff, and Philip Wang. A complete list of NCS publications and the full text of all NCS-R instruments can be found at http://www.hcp.med.harvard.edu/ncs. Send correspondence to NCS@hcp.med.harvard.edu.

References

DuMouchel WH, Duncan GJ. Using sample survey weights in multiple regression analyses of stratified samples. J Am Stat Assoc 1983; 78: 535–43.

Groves R, Fowler F, Couper M, Lepkowski J, Singer E, Tourangeau R. Survey Methodology. New York: Wiley, in press.

Gurin G, Veroff J, Feld SC. Americans View Their Mental Health: A Nationwide Interview. New York: Basic Books Inc., 1960.

Kessler RC, Üstün TB. The World Mental Health (WMH) survey initiative version of the World Health Organization Composite International Diagnostic Interview (CIDI). International Journal of Methods in Psychiatric Research 2004; (this issue).

Kish L, Frankel MR. Inferences from complex samples. Journal of the Royal Statistical Society 1974; 36 (Series B): 1–37.

Kish L, Kish JL. Survey Sampling. New York: John Wiley & Sons, 1965.

Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, 1987.

Tourangeau R, Smith TW. Collecting sensitive information with different modes of data collection. In M Couper, RP Baker, J Bethlehem, CZF Clark, J Martin, WL Nicholos II, JM O'Reilly eds. Computer Assisted Survey Information Collection. New York: Wiley, 1998, 431–54.

Turner CF, Ku L, Rogers SM, Lindberg LD, Pleck JH, Sonenstein FL. Adolescent sexual behavior, drug use, and violence: increased reporting with computer survey technology. Science 1998; 280 (5365): 867–73.

Turner CF, Lessler JT, Devore J. Effects of mode of administration and wording on reporting of drug use. In CF Turner, JT Lessler and J Gfroerer eds. Survey Measurement of Drug Use: Methodological Studies. Rockville MD: National Institute on Drug Abuse, 1982.

Veroff J, Douvan E, Kulka R. The Inner American: A Self Portrait from 1957 to 1976. New York: Basic Books, 1981.

Wolter KM. Introduction to Variance Estimation. New York: Springer-Verlag, 1985.

Zaslavsky AM, Schenker N, Belin TR. Downweighting influential clusters in surveys: application to the 1990 Post Enumeration Survey. Am J Stat Assoc 2001; 96 (455): 858–69.

Correspondence: RC Kessler, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston MA, USA 02115.
Telephone (+1) 617-432-3587.
Fax (+1) 617-432-3588.
Email: kessler@hcp.med.harvard.edu.