International Journal of Methods in Psychiatric Research
Int. J. Methods Psychiatr. Res. 19(Supplement 1): 36–48 (2010)
Published online in Wiley InterScience
(www.interscience.wiley.com) **DOI**: 10.1002/mpr.311

The National Survey on Drug Use and Health Mental Health Surveillance Study: calibration study design and field procedures

LISA J. COLPE,[1] PEGGY R. BARKER,[1] RHONDA S. KARG,[2] KATHY R. BATTS,[2] KATHERINE B. MORTON,[2] JOSEPH C. GFROERER,[1] STEPHANIE J. STOLZENBERG,[2] DAVID B. CUNNINGHAM,[2] MICHAEL B. FIRST[3] & JEREMY ALDWORTH[2]

1 Substance Abuse and Mental Health Services Administration, Office of Applied Studies, Rockville, MD, USA
2 RTI International, Research Triangle Park, NC, USA
3 Columbia University, College of Physicians and Surgeons and the New York State Psychiatric Institute, Biometrics Research Department, New York City, NY, USA

Key words
calibration, K6, Mental Health Surveillance Study

Correspondence
Lisa J. Colpe, PhD, MPH; National Institute of Mental Health, 6001 Executive Blvd, Bethesda, MD 20892, USA.
Telephone (+1) 301-443-3815
Email: Lisa.Colpe@NIH.gov

Received 8 March 2010;
revised 19 April 2010;
accepted 19 April 2010

Abstract

The Mental Health Surveillance Study (MHSS) is an ongoing initiative by the Substance Abuse and Mental Health Services Administration (SAMHSA) to monitor the prevalence of serious mental illness (SMI) among adults in the USA. In 2008, the MHSS used data from clinical interviews to calibrate mental health data from the National Survey on Drug Use and Health (NSDUH) for estimating the prevalence of SMI based on the full NSDUH sample. The clinical interview used was the Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV; SCID). NSDUH interviews were administered via audio computer-assisted self-interviewing (ACASI) to a nationally representative sample of the population aged 12 years or older. A total of 46 180 NSDUH interviews were completed with adults aged 18 years or older in 2008. The SCID was administered by mental health clinicians to a sub-sample of 1506 adults via telephone. This paper describes the MHSS calibration study procedures, including information on sample selection, instrumentation, follow-up, data quality protocols, and management of distressed respondents.

Introduction

This paper describes the study design and field procedures from the calibration study conducted in 2008 as part of the Substance Abuse and Mental Health Services Administration's (SAMHSA's) National Survey on Drug Use and Health (NSDUH) Mental Health Surveillance Study (MHSS). We also discuss calibration study procedures that were put in place to ensure high-quality data and to manage distressed respondents. While analysis procedures and the resulting estimates of serious mental illness (SMI) in the US adult population are presented in

a separate report in this issue (Aldworth *et al.*, 2010), discussion of sampling and design features relevant to establishing SMI estimation procedures are contained herein.

The MHSS was designed to satisfy the need for SMI prevalence estimates as outlined in Public Law (PL) No. 102-321, the Alcohol, Drug Abuse and Mental Health Administration (ADAMHA) Reorganization Act (Alcohol Drug Abuse and Mental Health Administration, 1992). This law established a block grant for US states to fund community mental health services for adults with SMI. The law requires state-level SMI incidence and prevalence estimates to be included in the applications for block grant funds that states submit annually. This legislation also requires SAMHSA to develop an operational definition of SMI and to establish an advisory group of technical experts to develop an estimation methodology based on this definition for use by states. SMI is defined by SAMHSA as: at least one Diagnostic and Statistical Manual of Mental Disorders (DSM) disorder, other than developmental or substance-use disorder, in the past 12 months that results in serious impairment. SAMHSA subsequently defined the term 'serious impairment' as impairment equivalent to a Global Assessment of Functioning (GAF) score of ≤50 (Substance Abuse and Mental Health Services Administration, 2005).

Overview

The main objective of the MHSS is to establish an ongoing national and state-level system of mental health surveillance using measures embedded in the NSDUH as indicators of mental health status. The current MHSS calibration study was preceded by a smaller study conducted in 2001 that tested the ability of a variety of brief mental health indicators to detect SMI. Three SMI screening scales were developed for the NSDUH: abbreviated versions of the Composite International Diagnostic Interview Short-Form (CIDI-SF) scale (Kessler *et al.*, 1998), the World Health Organization Disability Assessment Schedule (WHODAS) (Rehm *et al.*, 1999), and the K10/K6 nonspecific distress scales (Furukawa *et al.*, 2003). All screening scales were administered to an enriched convenience sample of 155 respondents followed by the 12-month Structured Clinical Interview for DSM-IV (SCID) (First *et al.*, 2002) and the GAF scale (American Psychiatric Association, 1994). At that time, SMI was defined as any 12-month *DSM-IV* disorder, other than a substance-use or developmental disorder, with a GAF score of <60 (a higher threshold than the current GAF cutoff score of 50). The results of the pilot study indicated that although all screening scales were significantly related to SMI, neither the CIDI-SF nor the WHODAS improved prediction significantly over the K10 or K6 scales (Kessler *et al.*, 2003). As a result, the K6 alone was used to estimate SMI in the 2001–2003 NSDUH surveys. The six questions in the K6 scale ask respondents how often they experienced certain symptoms of psychological distress during the 1 month in the past 12 months when they were at their worst emotionally. Responses for each question are scored from 0 ('none of the time') to 4 ('all of the time'), so the maximum score is 24. The cutoff point for SMI was determined to be a score of ≥13 (Kessler *et al.*, 2004). However, once certain changes in the NSDUH instrument were put into place, the prevalence estimate of SMI (derived from the K6 indicator) significantly increased from 9.9% to 12.2%. Recognizing that a more comprehensive study would be necessary to confidently measure and monitor SMI (Colpe *et al.*, 2009), the K6 score of 13+ was conceptualized as capturing Serious Psychological Distress rather than SMI. Serious Psuchological Distress has been reported in the NSDUH annual national findings report since 2004 (Substance Abuse and Mental Health Services Administration, 2005). In 2006, a meeting of experts who were convened to discuss options for SMI surveillance recommended that a full-scale calibration study using a population-based sample be embedded in NSDUH for (1) initial calibration of an SMI indicator and (2) continuous monitoring of the performance of the SMI indicator over time. The experts also recommended using a measure of impairment in addition to the K6 scale.

An initial feasibility study to test procedures associated with re-contacting and administering a clinical interview to respondents who agreed to participate in a follow-up study was carried out in June 2007. The design of the full-scale study conducted in 2008 (described in this paper) was based on findings from the initial feasibility study, both of which were carried out under contract by RTI International.

Methods

NSDUH survey

The NSDUH is an annual nationwide survey of the civilian, non-institutionalized population aged 12 years or older within the 50 states and the District of Columbia. Designed to provide national and state-level substance-use and mental health data, the NSDUH questionnaire is administered using computer-assisted interviewing methods. Computer-assisted interviewing encompasses both audio computer-assisted self-interviewing (ACASI) and computer-assisted personal interviewing (CAPI)

methods. For the NSDUH interview, the more sensitive content (on substance use and mental health) is administered using ACASI to facilitate privacy and confidentiality. The respondent reads along and listens through headphones as the computer plays an audio-recording of each question and then enters answers directly into the computer. The remainder of the interview is administered by the interviewer using CAPI. In 2008, a total of 46 180 interviews were conducted with adults aged 18 years or older. The 2008 NSDUH instrument was modified to include an expanded K6 scale (to include past 30-day and past 12-month reference periods) and two mental health impairment scales; WHODAS and the Sheehan Disability Scale (SDS). Using a split sample design, approximately equal numbers of adult respondents received an abbreviated WHODAS (see Novak *et al.*, 2010, this issue) or the SDS (Leon *et al.*, 1997). All adult respondents received the expanded K6 scale.

Sampling plan

The MHSS calibration study was designed to yield 1500 clinical follow-up interviews during 2008. A sub-sample of respondents aged 18 or older was selected from approximately 45 000 NSDUH main study adult interviews conducted throughout the year. A probability sampling algorithm, with stratification based on K6 scores, was programmed in the ACASI instrument so that field interviewers (FIs) could recruit selected respondents for the subsequent clinical psychiatric interview.

To optimize the MHSS sample allocation within seven scoring bands, assumed SMI rates were estimated using raw K6 scores and clinical case data from the National Comorbidity Survey Replication (NCS-R) clinical calibration study. Population percentages were obtained from the 2006 NSDUH. Using Neyman's optimal allocation (Lohr, 1999), a solution that minimized the design effect for prevalence of SMI was computed. Sampling rates for the MHSS were substantially lower for K6 scores in the 0 to 7 range under the assumption that fewer clinical positives would be identified in that scoring range when the K6 data were used in combination with impairment data to estimate SMI. Table 1 shows the sample distribution for the planned 1500 clinical follow-up interviews, as well as the expected design effect, effective sample size, and standard error (SE) and relative standard error (RSE) of the estimate of SMI. [The design effect is the product of the usual design effect for adults in the main survey (about 3.0) and the design effect for the two-phase sample stratified by K6 scores (about 0.2).]

An achieved sample of 1506 clinical follow-up interviews was distributed across the four calendar quarters with a slightly larger sample in the first quarter (468 follow-up interviews; see Table 2) and the remaining sample divided approximately equally among quarters 2 through 4. The larger sample in quarter 1 provided some cushion should clinical interview response rates have been lower than expected and permitted preliminary analyses to be conducted to ensure that the sampling and selection parameters were performing as anticipated. A consequence of the sample design was that respondents with low K6 total scores typically had relatively large weights. Three records with unusually large weights that had the effect of unduly influencing the receiver operating characteristic models, were removed from the dataset. One record with missing data on all K6 items was also removed, leaving 1502 records in the data file used for the calibration analysis. Based on the 1506 respondent records

Table 1 Mental Health Surveillance Study sample allocation ($N = 1500$)[1]

K6 score	Percent of population[2]	Assumed SMI rate (%)	Expected sample size
0 to 3	48.04	0.03	96
4 to 5	13.98	0.30	88
6 to 7	11.16	0.30	110
8 to 9	6.95	10.00	200
10 to 11	5.53	13.00	214
12 to 15	8.00	40.00	450
16 or higher	6.34	67.00	343
Total	100.00	8.95	1501

SMI, serious mental illness.
[1] Overall design effect: 0.6363; effective sample size: 2357; projected standard error (%): 0.59; projected relative standard error: 6.57.
[2] Source: 2006 National Survey on Drug Use and Health.

Table 2 2008 Mental Health Surveillance Study, summary of quarters 1 to 4

Design parameter	Quarter 1	Quarter 2	Quarter 3	Quarter 4	Total
NSDUH interview respondents aged 18 or older	10798	12931	11499	10952	46180
Unweighted K-6 distribution (%)					
Score 0 to 3	0.46	0.45	0.45	0.44	0.45
Score 4 to 5	0.15	0.14	0.14	0.14	0.14
Score 6 to 7	0.10	0.10	0.10	0.10	0.10
Score 8 to 9	0.07	0.07	0.07	0.07	0.07
Score 10 to 11	0.05	0.06	0.06	0.06	0.06
Score 12 to 15	0.09	0.09	0.09	0.09	0.09
Score 16 or higher	0.08	0.09	0.09	0.09	0.09
Eligible for MHSS	10324	12268	10994	10519	44105
Eligibility rate	0.96	0.95	0.96	0.96	0.95
Selected for clinical follow-up	697	531	487	576	2291
Agreed to clinical follow-up	587	462	418	510	1977
Percentage agreeing to clinical follow-up-unweighted	0.84	0.87	0.86	0.88	0.86
Percentage agreeing to clinical follow-up weighted	0.63	0.86	0.74	0.84	0.76
Completed clinical interviews	470	361	319	356	1506
Clinical interview completion rate unweighted	0.80	0.78	0.76	0.70	0.76
Clinical interview completion rate weighted	0.821	0.83	0.74	0.70	0.77

collected, an unweighted 86.3% agreement rate for the clinical follow-up interview and an unweighted 76.2% clinical interview completion rate were achieved; weighted agreement and completion rates were 76.5% and 77.1%, respectively.

Instrumentation

Clinical interview

The clinical interview measure used in the MHSS calibration study was the Structured Clinical Interview for DSM-IV-TR Axis I Disorders Non-patient Edition (SCID-I/NP) (First *et al.*, 2002). The SCID-I/NP is a semi-structured interview that has been widely used in clinical calibration studies such as the NCS-R (Kessler *et al.*, 2004), the National Survey of American Life (Jackson *et al.*, 2004), and the NSDUH substance-use disorders reappraisal study (Jordan *et al.*, 2008). It has demonstrated good reliability (Segal *et al.*, 1995; Zanarini and Frankenburg, 2001; Zanarini *et al.*, 2000) and validity (Fennig *et al.*, 1994; Kranzler *et al.*, 1996, 1995; Ramirez Basco *et al.*, 2000; Shear *et al.*, 2000; Steiner *et al.*, 1995). The interview was modified to assess past 12-month mental health disorders and functioning via telephone interview by a trained clinical interviewer (CI).

Diagnostic modules contained in the MHSS version of the SCID are listed in Table 3. The module for lifetime manic episode was included to provide context for understanding whether a past 12-month major depressive episode was experienced as part of a unipolar mood disorder or as a component of a bipolar disorder (regardless of whether a manic episode was also experienced in the past year). The module for lifetime major depressive episode was included for a separate NSDUH analysis. The module to assess intermittent explosive disorder was obtained from the (optional) impulse control disorders section of the SCID.

In addition to the diagnostic modules, the MHSS SCID included four other modules:

1 an open-ended overview module, designed to elicit information about the respondent's diagnostic and treatment history and current status in a way that establishes some level of rapport between the interviewer and the respondent
2 a screener module containing questions for several of the anxiety disorders and eating disorders
3 a module containing the DSM-IV Axis V GAF Scale (a CI rating of the respondent's period of worst psychological, social, and occupational functioning during the past year)

Table 3 Diagnostic modules contained in the Mental Health Surveillance Study structured clinical interview

Mood disorders	Past year eating disorders
Past year major depressive episode	Anorexia nervosa
Lifetime major depressive episode	Bulimia nervosa
Past year manic episode	
Lifetime manic episode	Past year impulse control disorders
Dysthymic disorder	Intermittent explosive disorder
Past year psychotic disorders	Past year substance use disorders
Psychotic screen	Alcohol abuse
	Alcohol dependence
Past year anxiety disorders	Non-alcohol substance abuse
Post-traumatic stress disorder	Non-alcohol substance dependence
Panic disorder with and without agoraphobia	
Agoraphobia without history of panic disorder	Past year adjustment disorders
Social phobia	Adjustment disorder
Specific phobia	
Obsessive compulsive disorder	
Generalized anxiety disorder	

4 a module for documenting the CI's impressions of the interview situation, including ratings of the respondent's level of privacy, co-operation, and comprehension, as well as the overall validity of the interview data (any interview deemed by the CI or clinical supervision team to be of questionable validity was discarded).

CI recruitment and training

CI recruitment

Applicants for the CI position were recruited from graduate programs accredited by the American Psychological Association (APA) in clinical and counseling psychology, professional psychology internships, and postdoctoral training (APA Web directory; American Psychology Postdoctoral and Internship Centers directory) and key professional organizations in psychology, psychiatry, social work, and counseling that emphasize research, as well as states' psychological associations. Necessary CI credentials for this study included:

1 having a master's or doctoral degree in clinical or counseling psychology, a medical degree with a specialty in psychiatry, or an advanced degree in a related field such as clinical social work
2 a willingness to attend a 4-day training
3 a willingness to meet specific scheduling requirements for the position.

Other key skills recommended for CIs included experience with semi-structured diagnostic interviews in a research setting (preferably experience with the SCID); strong conceptual skills; good attention to detail; the ability to accurately administer a complex interview protocol; the ability to develop and maintain strong rapport with respondents, including the ability to adjust interview style to competencies and the personality of the respondent; and the flexibility and capacity to work as part of a team and accept constructive feedback. A total of 268 Web-based applications were received between 1 September and 1 October 2007. The 84 applicants who met the criteria noted above were invited for telephone interviews, which included administering parts of the SCID during a mock interview; 79 applicants completed this stage, and 30 were selected. To ensure adequate coverage during peak times of interview requests, the 30 CIs hired for the study were distributed across the Eastern ($N = 11$), Central ($N = 12$), Mountain ($N = 1$), and Pacific ($N = 6$) standard time zones. Based on experience conducting similar studies, between 20% and 30% of the CIs hired were not expected to pass certification; therefore 30% over-hiring was carried out. A pool of back-up applicants who completed the telephone interviews but who did not make the first round of hiring was also created. The applicants on this alternate list ($N = 26$) agreed to be contacted if additional opportunities for hiring on this study arose.

CI training – non-clinical

The initial 4-day training session was attended by 30 CI candidates. Before the training session, candidates

received an MHSS handbook outlining data collection procedures as well as the use of audio-recording equipment and study materials.

Training was split into two sections: a non-clinical portion detailing administrative tasks and a clinical portion focusing on SCID administration. The non-clinical portion lasted 1 day and was led by two data collection managers. This training included: (1) an overview of the NSDUH and the MHSS; (2) respondents' rights and confidentiality; (3) proper informed consent procedures; (4) procedures for contacting respondents; (5) use of the Web-based case management system; (6) procedures for audio-recording interviews and uploading audio files for quality control review; and (7) proper shipment of interview materials. A virtual laboratory was set up each evening, with trainers on hand to answer mock interviews from CIs. The CIs gained additional practice in and outside the class setting up and using customized audio-recording software and a Web-based case management system, answering respondent questions, documenting cases and overcoming non-response.

CI training – clinical assessment

The second, third, and fourth days of the training session were dedicated to mastering the clinical interview. This portion of the training was led by four clinical supervisors (CSs) – experts in the DSM and the SCID – and was overseen by the lead author of the SCID. The main themes of this portion of the training were: (1) understanding the dynamics of a semi-structured interview; (2) learning the specific format/conventions of the SCID; (3) acquiring proper probing techniques to elicit codable information; (4) learning the nuances of assessing the various DSM-IV diagnostic criteria included in the SCID; (5) managing challenging respondents; and (6) properly assigning a GAF score. The trainers used a variety of active learning techniques such as large and small group round robin practice sessions, paired mock interviews, and homework to be completed after the training session had concluded for the day.

Other issues were addressed during training. For example, instruction was given on how to compensate for the absence of visual cues in a telephone interview. Trainees were also provided with training on sensitivity to people of different cultures and varying socio-economic levels. Finally, because the interview involved questions about past 12-month feelings and behaviors, instruction was provided about properly dealing with potentially distressed, suicidal, and/or homicidal respondents.

All trainees who successfully completed training were required to conduct at least two certification interviews with real respondents to demonstrate proficiency with the study protocol and the clinical interview. Data from the certification interviews were not included in the final dataset. The clinical supervision team reviewed the audio tapes of the certification interviews and provided prompt feedback to the trainees after each interview. Trainees who demonstrated an acceptable level of proficiency in their certification interviews were hired as interviewers for the study. Trainees who did not demonstrate mastery of the interview over the course of three certification interview attempts were not hired.

Field interviewer (FI) training

All NSDUH FIs ($N = 698$) were required to review an MHSS handbook, complete an MHSS electronic training course, and attend a 1-hour classroom or teleconference training session. This training program provided details on the FI-specific protocols and procedures for recruiting respondents for the follow-up clinical interviews. These procedures included following the computer-assisted interviewing scripts to recruit respondents, providing the follow-up study description for informed consent, paying and documenting payment to respondents and collecting adequate respondent contact information for CI follow-up.

Data collection

When a respondent was selected for the follow-up clinical interview, a series of recruitment screens were automatically displayed on the FI's computer at the end of the main study's NSDUH interview. The FI read these screens aloud to the respondent to determine whether he or she would be willing to participate in an additional study that would gather more information about his/her recent mental health history. Respondents were then presented with an MHSS description that provided information necessary for informed consent. Respondents agreeing to take part in the MHSS were then given $30 in cash and a payment receipt that included the toll-free number for the National Suicide Prevention Lifeline. The FI then collected contact information from the respondent, including the respondent's first name, telephone number, and the best time for a CI to call.

This respondent contact information was transmitted, typically later that day, to the MHSS research site by the FI, via the normal transmission of NSDUH interview data from the FI's laptop. Upon receipt, this contact information was processed and forwarded to the data collection

managers, who were responsible for assigning cases and handling all administrative functions for the CIs (e.g. approving time sheets, tracking production and costs). Individual cases were assigned by the data collection managers to a specific CI, taking time zone considerations and availability into account. The CI then completed the follow-up clinical interview as soon as feasible after the NSDUH interview was completed, with contact attempts beginning within 24 hours of receiving a case. Each follow-up interview had to be completed within 4 weeks of the date of the NSDUH interview to ensure comparability of the two datasets. During each MHSS interview, the CI completed the SCID on paper and audio-recorded the interview (with permission). Within 48 hours after completion of the interview, the CI uploaded the audio file to the Web-based case management system, and edited and shipped the paper SCID to the research site for proper handling, keying, and analysis.

To maximize interview completion rates, several strategies were set in place. CIs were instructed to schedule an interview in spite of not being personally available because cases could easily be assigned to another CI. Respondents who were difficult to reach were contacted on different days of the week and at different times of the day. 'Unable-to-contact' letters were sent to respondents who were difficult to reach after multiple attempts. These letters explained the importance of the study, reminded respondents of the pre-payment they had received for participating, and provided RTI's toll-free number so the respondent could contact a data collection manager with any questions and to schedule an appointment.

Data management

The CI administered the SCID over the telephone from a private location in his/her home or office. When calling to conduct the SCID, the CI accessed the respondent's name and telephone number from the secure RTI-hosted case management system, and called the respondent while this information was displayed on-screen. Interviewers kept respondent names, telephone numbers, and any other piece of contact information confidential at all times.

The CIs noted the respondent's answers directly in the SCID booklet and kept the booklet secure until sending it via Federal Express to a designated individual at RTI. Once received at RTI, the SCID booklets remained in a secure location accessible only by authorized project personnel. The only identifying information on the SCID booklet was a randomly generated seven-digit number to link the SCID data to the NSDUH data, which could only be linked by RTI's researchers who had completed mandatory confidentiality training and had signed confidentiality pledges. All CIs were issued project-owned laptop computers, pre-configured with encryption software and custom software to automate the secure upload of audio files using the encryption facilities of the HTTPS (Hypertext Transfer Protocol Secure) protocol. Audio files were strongly encrypted both at rest and also during transmission.

Quality assurance

Ongoing supervision

The CSs reviewed all SCID booklets item-by-item, comparing the notes provided by the CI and the diagnostic rating, and listening to the accompanying audio files, as needed, to ensure confidence in the data. The CSs also reviewed the full audio recordings for a randomly selected 10% of the clinical interviews (N = 150). Full audio recordings were also conducted for an additional sub-set of interviews that were complex or otherwise challenging (N = 177; 13%). If the CS reviewing the data was not confident in the rating (based on the CI notes and audio files), the data were discussed with the other CSs and the CI who conducted the interview, and modified as needed. The SCID data were also checked electronically for internal coding consistency. Documenting the CIs' performance and providing feedback to the CIs was an integral part of the clinical supervision process. Written supervision rating forms were completed for each clinical interview conducted, and hard copies of these completed rating forms were on file for each of the CIs. Feedback was provided to CIs for each full review conducted. Interviews that were complex or problematic were discussed with the other CSs and the CIs who had administered the interviews.

Clinical interviewer agreement

To further ensure the quality of the data being collected in the clinical interviews, inter-rater agreement (IRA) exercises were conducted at the end of each calendar quarter. The purpose of the exercises was to assess the level of agreement between each CI and a reference with respect to the DSM-IV diagnoses covered in the interview. Based on the IRA results, further retraining was provided for the CIs to increase agreement in the future. Four CSs supervised data collection of 21 CIs who participated in the IRA exercises. The reference was determined by a consensus among the four CSs. Stimulus interviews used for the IRA/calibration exercises were selected from the MHSS SCID interviews completed

during the year. During the first half of the year, the focus of these end-of-quarter exercises was evaluation and enriched training using more complex cases to ensure that CI's were approaching difficult cases appropriately. During the second half of the year, the focus of the end-of quarter exercises shifted from retraining CIs to obtaining an IRA estimate for assessments that were more straightforward and characteristic of the MHSS than the more complex cases used in the first half of the year; therefore, more typical cases for the stimulus interviews (e.g. full assessment of a small number of disorders, the presence of one or more disorders not related to substance use, and a relatively straightforward clinical history and presentation of symptoms) were used during the IRA exercises conducted during the second half of the year.

Agreement between the DSM-IV diagnoses rated by the reference (consensus ratings among four CSs) and each of the CIs was quantified using Cohen's kappa (Cohen, 1960), a reliability statistic that corrects for chance agreement. A kappa of 0.61 or higher is considered substantial reliability (Landis and Koch, 1977). For each CI, an agreement ratio (kappa coefficient) was computed across the SCID rating categories of present (1) or absent (0) across disorders. For each CI we also calculated the total percentage of agreement between his or her ratings and the reference ratings across all symptoms and disorders. The kappa coefficients for the diagnoses assessed ranged from 0.49 (SE = 0.17) to perfect agreement (kappa = 1.00, SE = 0.00). The kappa coefficient was statistically significant ($P < 0.05$) for 98% of the CI/reference agreement analyses; that is, for these cases the null hypothesis of independence between the CI and reference ratings was rejected. Among the CI kappa coefficients, 83% were characterized as demonstrating 'substantial reliability' (0.61 or higher); 9% of the kappa coefficients were in the moderate reliability range (0.41–0.60), and 8% of the kappa coefficients were in the fair reliability range (0.21–0.40) (Landis and Koch, 1977).

In addition to the agreement analyses using kappa statistics, simple agreement analyses were conducted by comparing the overall percentage of agreement across disorders among all CIs and the reference. There were high rates of agreement (90–100%) for 88% of the disorders, and 100% agreement between the reference and CI ratings in the presence of one or more mental disorders unrelated to substance use in the past 12 months. Discordance between the CS and CI assessments most often occurred in cases where the CSs rated a disorder as 'insufficient data' whereas the CIs reported that the disorder was ruled out as 'absent' (false negatives) or that the diagnosis was sufficiently supported (false positives). Error occurred for disorders in which CIs overestimated the severity of the symptoms (false positives) or overestimated the clinical significance or the utility of the data collected (false negatives). The lowest values for agreement were found for dysthymic disorder (false positives and false negatives), the anxiety disorders (false positives), and alcohol-use disorders (false negatives). Evaluations of these disorders were therefore paid special attention during the editing processes and in selecting interviews for IRA/calibration exercises.

We also compared continuous GAF scores between the CIs and the reference, and we examined agreement between CIs' ratings of the presence of any mental disorder (unrelated to substance use), and the presence of SMI as defined as any mental disorder (unrelated to substance use) and a GAF of 50 or below compared with the reference ratings. For the GAF, 83% of the CI ratings were in the same decile as the reference rating; 13% of the ratings were within 1 decile of the reference rating, and 4% were within 2 deciles of the reference rating. Using a GAF rating of ≤50 and the presence of one or more mental disorders unrelated to substance use to define the presence of a SMI, there was 100% agreement between the reference and CI ratings for SMI.

Managing distressed respondents

A number of measures were taken to ensure the safety of potentially distressed respondents. First, we provided explicit protocols for CIs to follow if they encountered respondents with suicidal or homicidal thoughts in the past 2 weeks, including passive or active suicidal or homicidal thoughts. Training and supervision were provided for managing respondents who expressed sadness, agitation, frustration, or any other strong emotion during the course of the clinical interview. A detailed distressed respondent protocol, on which CIs were intensively trained, was employed for this study. The distressed respondent protocol provided definitions and examples of five types of distressed respondents, along the continuum of no risk of harm (i.e. respondent is agitated or upset) to imminent danger (e.g. respondent reports active suicidal thoughts, a plan, and a means to carry out that plan). The distressed respondent protocol then gives step-by-step instructions for handling each of the five types of distressed respondents. For example, respondents who admitted to passive suicidal thoughts in the past 2 weeks were encouraged by the CI to be connected in a three-way call with a crisis counselor at a national suicide prevention hotline and, if appropriate, receive information about mental health services available in his or her community. In cases of clear imminent danger for the

respondent or an identifiable victim, CIs were instructed to call 911 to report this incident, along with the respondent's name, home address, and telephone number, all of which was electronically accessible to CIs.

A second measure for ensuring the safety of all participants was hiring CIs who had strong mental health backgrounds, many of whom were also seasoned clinicians with experience assessing risk and providing direct care for distressed individuals. With this advantage, our training focused on the study's distressed respondent protocol, the nuances of assessing risk and providing care for distressed respondents by telephone, and practicing the application of the distressed respondent protocol with case scenarios. Third, the overall supervisor of the clinical interviewing effort was a licensed clinical psychologist and certified health-care provider. This supervisor was integrally involved in supervising the CIs and other CSs, and was on-call any time that a distressed respondent might be encountered so that level of risk could be verified, and consultation and debriefing could be provided. After each encounter with a distressed respondent, the CI immediately contacted the supervisor to review the details of the incident, the assessment of risk, and the application of the distressed respondent protocol. If unusual circumstances arose, the supervisor contacted the study director and Institutional Review Board.

Finally, the study included an automated reporting system that electronically alerted project management, the clinical supervisors, and the data collection managers when a CI had encountered a distressed respondent and provided the details of that event in the form of an electronic incident report. Whenever the distressed respondent protocol was used, the CI received telephone supervision and then immediately documented the details of the incident in the case management system. Submitting the incident report in the case management system occasioned the automated delivery of the report to the study's management team, CSs, and data collection managers. These methods were effective and allowed us to properly handle 32 incidents of distressed respondents in 2008, all of which included recent suicidal ($N = 31$) or homicidal ($N = 1$) thoughts in the absence of plans or intentions to do harm. There were no cases of imminent danger; therefore, it was not necessary to breach confidentiality and contact 911 for any MHSS cases.

Results

Clinical interviews

Of the 1977 respondents who agreed to the clinical follow-up interview, 1502 usable interviews were completed. At the end of the 4-week response window, an unweighted 86% agreement rate for the clinical follow-up interview and an unweighted 76% completion rate were achieved. The weighted agreement and completion rates were 76% and 77%, respectively.

The most common reason for reduction in response rates was inability to contact respondents by telephone after repeated attempts, at 15%, followed by wrong or disconnected telephone numbers, at 3%. Of those respondents who started the interview, 3% did not complete the interview and timed-out of the follow-up window before completing the interview. Direct refusals were the least likely to affect response rates at 1%. Methods used by the CIs to overcome response barriers included attempting contact during different days of the week and at different times of the day. Also, 551 letters were mailed to respondents who were difficult to reach; this method yielded 209 completed interviews. Other barriers to participation included respondents' concerns about confidentiality, objection to being audio recorded, and time and availability constraints. Respondents were informed of their right to both privacy and confidentiality, which permitted them to waive the audio recording and decline to answer any questions.

Sample characteristics

Initial descriptive analyses and statistical tests were conducted in the MHSS to check for imbalances in key demographic characteristics between the two half-samples assigned to either of the two impairment scales. Key demographic characteristics included gender, age, race/ethnicity, and education. Unweighted descriptive statistics of the demographic variables are shown in Table 4 and weighted versions of those descriptive statistics are shown in Table 5. Included in the descriptive statistics are frequencies and percentages of the entire 12-month NSDUH sample, the sub-set of respondents selected for the SCID, and those who completed the SCID. Chi-square tests were conducted to compare the completed SCID cases between the two half-samples.

Table 4 indicates that for the unweighted sample, the percentage of females for the SCID was high when compared with the 12-month NSDUH sample (likely because of the high K6 score sampling approach used). However, the two half-samples did not differ substantially on any of the demographic measures. For the weighted data, Table 5 shows some discrepancy in the completed cases in two half-samples with respect to gender, but the difference is not statistically significant.

Table 4 Unweighted descriptive statistics of demographic characteristics

	2008 12-month NSDUH respondents[1]		Selected SCID cases				Completed SCID cases							
			Sample A (WHODAS)		Sample B (SDS)		Sample A (WHODAS)		Sample B (SDS)		Total			
Variable	N	%	N	%	N	%	N	%	N	%	N	%	χ^2 (df)	*P*-value
Gender														
Male	21 596	46.8	463	39.4	429	38.4	278	36.5	269	36.3	547	36.4	0.01 (1)	0.92
Female	24 584	53.2	711	60.6	688	61.6	483	63.5	472	63.7	955	63.6		
Race/ethnicity														
White, NH	30 583	66.2	815	69.4	756	67.7	546	71.7	529	71.4	1075	71.6		
Black, NH	5 369	11.6	154	13.1	131	11.7	92	12.1	71	9.6	163	10.9	1.57 (3)	0.20
Other	3 459	7.5	80	6.8	112	10.0	53	7.0	59	8.0	112	7.5		
Hispanic	6 769	14.7	125	10.6	118	10.6	70	9.2	82	11.1	152	10.1		
Age														
18–25	23 198	50.2	678	57.8	630	56.4	454	59.7	428	57.8	882	58.7	0.39 (2)	0.68
26–49	16 376	35.5	390	33.2	375	33.6	246	32.3	246	33.2	492	32.8		
50+	6 606	14.3	106	9.0	112	10.0	61	8.0	67	9.0	128	8.5		
Education														
<High School	7 590	16.4	200	17.0	190	17.0	105	13.8	110	14.8	215	14.3		
High School graduate	15 322	33.2	364	31.0	365	32.7	220	28.9	220	29.7	440	29.3	0.43 (3)	0.73
Some college	13 356	28.9	379	32.3	325	29.1	265	34.8	236	31.8	501	33.4		
College graduate	9 912	21.5	231	19.7	237	21.2	171	22.5	175	23.6	346	23.0		

df = degrees of freedom, *N* = frequency, NH = non-Hispanic, SCID = Structural Clinical Interview for DSM-IV, SDS = Sheehan Disability Scale, WHODAS = eight-item World Health Organization Disability Scale.

[1] This includes all cases for persons aged 18 years or older.

Int. J. Methods Psychiatr. Res. 19(Supplement 1): 36–48 (2010). DOI: 10.1002/mpr

Table 5 Weighted[1] descriptive statistics of demographic characteristics(numbers in thousands)

			Selected SCID cases				Completed SCID cases							
	2008 12-month NSDUH cases[2]		Sample A (WHODAS)		Sample B (SDS)		Sample A (WHODAS)		Sample B (SDS)		Total			
Variable	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	χ^2 (df)	*P*-value
Gender														
Male	89296	45.7	47024	45.7	46574	46.0	58249	51.4	50305	45.1	108553	48.3	0.52 (1)	0.47
Female	106266	54.3	55839	54.3	54643	54.0	55182	48.6	61188	54.9	116370	51.7		
Race/ethnicity														
White, NH	131964	67.5	69942	68.0	73555	72.7	73253	64.6	81481	73.1	154734	68.8		
Black, NH	21712	11.1	11211	10.9	8056	8.0	14146	12.5	11219	10.1	25365	11.3	0.93 (3)	0.43
Other	12518	6.4	9225	9.0	9495	9.4	11196	9.9	3292	3.0	14488	6.4		
Hispanic	29369	15.0	12485	12.1	10111	10.0	14836	13.1	15500	13.9	30336	13.5		
Age														
18–25	28151	14.4	14559	14.2	14255	14.1	16533	14.6	16405	14.7	32938	14.6	0.26 (2)	0.77
26–49	88054	45.0	44956	43.7	46146	45.6	46760	41.2	53072	47.6	99833	44.4		
50+	79358	40.6	43348	42.1	40817	40.3	50136	44.2	42016	37.7	92152	41.0		
Education														
<High School	31766	16.2	11450	11.1	8585	8.5	9569	8.4	10827	9.7	20396	9.1		
High School graduate	61539	31.5	32673	31.8	38635	38.2	34800	30.7	39744	35.6	74545	33.1	0.24 (3)	0.87
Some college	49864	25.5	30655	29.8	22879	22.6	34509	30.4	28534	25.6	63043	28.0		
College graduate	52393	26.8	28085	27.3	31119	30.7	34552	30.5	32387	29.0	66939	29.8		

df = degrees of freedom, *N* = frequency, NH = non-Hispanic, SCID = Structural Clinical Interview for DSM-IV, SDS = Sheehan Disability Scale, WHODAS = eight-item World Health Organization Disability Scale.

[1] The overall NSDUH analysis weight was used for the NSDUH cases. The overall NSDUH analysis weight multiplied by the inverse of the SCID selection probability was used for the selected SCID cases. The MHSS combined unadjusted sample weight was used for completed SCID cases. The MHSS combined unadjusted sample weight included the following weights: overall NSDUH analysis weight; inverse of the SCID selection probability; non-response adjustment for clinical interview; and post-stratification adjustments by gender, race/ethnicity, and age.

[2] This includes all cases for persons aged 18 or older.

Discussion

As part of an ongoing initiative to monitor the prevalence of SMI among adults, the goals of the Mental Health Surveillance Study (MHSS) conducted by SAMHSA were to incorporate candidate measures of SMI in the 2008 NSDUH survey, to identify and select respondents from the NSDUH to receive a structured clinical psychiatric interview within 4 weeks, and ultimately to calibrate the data collected in the NSDUH to the clinical interview data, resulting in estimates of the prevalence of SMI based on the full NSDUH sample. In 2008, a total of 46 180 NSDUH interviews and 1502 clinical interviews were completed with adults aged 18 years and older.

Clinical interviews were conducted by master's and doctoral level mental health professionals who had been carefully and extensively trained to administer the semi-structured clinical interview over the telephone. The study protocol included comprehensive instructions for identifying and managing distressed respondents as well as ongoing supervision and inter-rater training exercises for the clinical interviewers.

Descriptive analysis of the demographic characteristics of the clinical interview sample indicated that the sample was balanced and consistent with the overall NSDUH sample. A 76% unweighted completion rate among those that agreed to the clinical interview was achieved by the study, a commendable rate given the shortness of the 4-week data collection period and the lack of in-person follow-up

Given the success in the execution of the 2008 study, SAMHSA will continue to include the K6 and WHODAS scales in the main NSDUH interview and to collect clinical interview data from a sub-set of NSDUH respondents to monitor the prevalence of SMI among adults in the USA. This will allow additional analysis of SMI at the state-level, as well as investigations into the prevalence and impact of milder forms of mental illness (e.g. with mild to moderate functional impairment). These continuing calibration activities and the ongoing nature of the study are significant contributions to mental health surveillance in the USA.

Acknowledgments

This work was prepared by the Division of Population Surveys, Office of Applied Studies, Substance Abuse and Mental Health Services Administration, US Department of Health and Human Services, and by RTI International (a trade name of Research Triangle Institute). Work by RTI was performed under Contract No. 283-2004-00022. The authors thank Mary Ellen Marsden for her helpful review.

Declaration of interest

The authors have no competing interests.

References

Alcohol Drug Abuse and Mental Health Administration. *Alcohol, Drug Abuse, and Mental Health Administration (ADAMHA) Reorganization Act, P.L. 102-321, S. 1306.*, 1992.

Aldworth J., Colpe L.J., Gfroerer J.C., Novak S.P., Chromy J.R., Barker P.G., Barnett-Walker K., Karg R.S., Morton K.B., Spagnola K. (2010). National Survey on Drug Use and Health's Mental Health Surveillance Study: calibration analysis. *Int J Methods Psychiatr Res*, **19**(Suppl. 1), 61–87.

American Psychiatric Association. (1994). *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV)*, American Psychiatric Association.

Cohen J. (1960). A coefficient of agreement for nominal scales. *Educ Psychol Measurem*, **20**, 37–46, DOI: 10.1177/001316446002000104

Colpe L.J., Epstein J.F., Barker P.R., Gfroerer J.C. (2009). Screening for serious mental illness in the National Survey on Drug Use and Health (NSDUH). *Ann Epidemiol*, **19**, 210–211, DOI: 10.1016/j.annepidem.2008.09.005

Fennig S., Craig T., Lavelle J., Kovasznay B., Bromet E.J. (1994). Best-estimate versus structured interview-based diagnosis in first-admission psychosis. *Comprehensive Psychiatry*, **35**, 341–348, DOI: 10.1016/0010-440X(94)90273-9

First M.B., Spitzer R.L., Gibbon M., Williams J.B.W. (2002). *Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition*. (SCID-I/NP), New York State Psychiatric Institute, Biometrics Research Department.

Furukawa T.A., Kessler R.C., Slade T., Andrews G. (2003). The performance of the K6 and K10 screening scales for psychological distress in the Australian National Survey of Mental Health and Well-Being. *Psychol Med*, **33**, 357–362, DOI: 10.1017/S0033291702006700

Jackson J.S., Neighbors H.W., Nesse R.M., Trierweiler S.J., Torres M. (2004). Methodological innovations in the National Survey of American Life. *Int J Methods Psychiatr Res*, **13**, 289–298, DOI: 10.1002/mpr.182

Jordan B.K., Karg R.S., Batts K.R., Epstein J.F., Wiesen C. (2008). A clinical validation of the National Survey on Drug Use and Health assessment of substance use disorders. *Addictive Behav*, **33**, 782–798, DOI: 10.1016/j.addbeh.2007.12.007

Kessler R.C., Abelson J., Demler O., Escobar J.I., Gibbon M., Guyer M.E., Howes M.J., Jin R., Vega W.A., Walters E.E., Wang P., Zaslavsky A., Zheng H. (2004). Clinical calibration of DSM-IV diagnoses in the World Mental Health (WMH) version of the World Health Organization

(WHO) Composite International Diagnostic Interview (WMHCIDI). *Int J Methods Psychiatr Res*, **13**, 122–139, DOI: 10.1002/mpr.169

Kessler R.C., Andrews G., Mroczek D., Üstün T.B., Wittchen H.-U. (1998). The World Health Organization Composite International Diagnostic Interview Short Form (CIDI-SF). *Int J Methods Psychiatr Res*, **7**, 171–185, DOI: 10.1002/mpr.47

Kessler R.C., Barker P.R., Colpe L.J., Epstein J.F., Gfroerer J.C., Hiripi E., Howes M.J., Normand S.L., Manderscheid R.W., Walters E.E., Zaslavsky A.M. (2003). Screening for serious mental illness in the general population. *Arch Gen Psychiatry*, **60**, 184–189.

Kranzler H.R., Kadden R.M., Babor T.F., Tennen H., Rounsaville B.J. (1996). Validity of the SCID in substance abuse patients. *Addiction*, **91**, 859–868, DOI: 10.1080/09652149640068

Kranzler H.R., Kadden R.M., Burleson J.A., Babor T.F., Apter A., Rounsaville B.J. (1995). Validity of psychiatric diagnoses in patients with substance use disorders: is the interview more important than the interviewer? *Comprehensive Psychiatry*, **36**, 278–288, DOI: 10.1016/S0010-440X(95)90073-X

Landis J.R., Koch G.G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, **33**, 159–174.

Leon A.C., Olfson M., Portera L., Farber L., Sheehan D.V. (1997). Assessing psychiatric impairment in primary care with the Sheehan Disability Scale. *Int J Psychiatry Med*, **27**, 93–105, DOI: 10.2190/T8EM-C8YH-373N-1UWD

Lohr S.L. (1999). *Sampling: Design and Analysis*, Duxbury Press.

Novak S.P., Colpe L.J., Barker P.G., Gfroerer J.C. (2010). Development of a brief mental health impairment scale using a nationally representative sample in the United States. *Int J Methods Psychiatr Res*, **19**(Suppl. 1), 49–60.

Ramirez Basco M., Bostic J.Q., Davies D., Rush A.J., Witte B., Hendrickse W., Barnett V. (2000). Methods to improve diagnostic accuracy in a community mental health setting. *Am J Psychiatry*, **157**, 1599–1605.

Rehm J., Ustun T.B., Saxena S., Nelson C.B., Chatterji S., Ivis F., Adlaf E. (1999). On the development and psychometric testing of the WHO screening instrument to assess disablement in the general population. *Int J Methods Psychiatr Res*, **8**, 110–123, DOI: 10.1002/mpr.61

Segal D.L., Kabacoff R.I., Hersen M., Van Hasselt V.B., Ryan C.F. (1995). Update on the reliability of diagnosis in older psychiatric outpatients using the Structured Clinical Interview for DSM-III-R. *J Clin Geropsychol*, **1**, 313–321.

Shear M.K., Greeno C., Kang J., Ludewig D., Frank E., Swartz H.A., Hanekamp M. (2000). Diagnosis of nonpsychotic patients in community clinics. *Am J Psychiatry*, **157**, 581–587.

Steiner J.L., Tebes J.K., Sledge W.H., Walker M.L. (1995). A comparison of the structured clinical interview for DSM-III-R and clinical diagnoses. *J Nerv Mental Dis*, **183**, 365–369.

Substance Abuse and Mental Health Services Administration. (2005). *Results from the 2004 National Survey on Drug Use and Health: National Findings* (Office of Applied Studies, NSDUH Series H-28, DHHS Publication No. SMA 05-4062), Rockville, MD.

Zanarini M.C., Frankenburg F.R. (2001). Attainment and maintenance of reliability of axis I and II disorders over the course of a longitudinal study. *Comprehensive Psychiatry*, **42**, 369–374, DOI: 10.1053/comp.2001.24556

Zanarini M.C., Skodol A.E., Bender D., Dolan R., Sanislow C., Schaefer E., Morey L.C., Grilo C.M., Shea M.T., McGlashan T.H., Gunderson J.G. (2000). The Collaborative Longitudinal Personality Disorders Study: reliability of axis I and II diagnoses. *J Personality Disord*, **14**, 291–299.