Design and field procedures in the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A)

RONALD C. KESSLER,[1] SHELLI AVENEVOLI,[2] E. JANE COSTELLO,[3] JENNIFER GREIF GREEN,[1] MICHAEL J. GRUBER,[1] STEVEN HEERINGA,[4] KATHLEEN R. MERIKANGAS,[5] BETH-ELLEN PENNELL,[4] NANCY A. SAMPSON[1] & ALAN M. ZASLAVSKY[1]

1 Department of Health Care Policy, Harvard Medical School, Boston, MA, USA
2 Division of Developmental Translational Research, National Institute of Mental Health
3 Center for Developmental Epidemiology, Department of Psychiatry and Behavioral Sciences, Duke University Medical School
4 Survey Research Center, Institute for Social Research, University of Michigan
5 Section on Developmental Genetic Epidemiology, Intramural Research Branch, National Institute of Mental Health

Abstract

An overview is presented of the design and field procedures of the US National Comorbidity Survey Replication Adolescent Supplement (NCS-A), a US face-to-face household survey of the prevalence and correlates of DSM-IV mental disorders. The survey was based on a dual-frame design that included 904 adolescent residents of the households that participated in the US National Comorbidity Survey Replication (85.9% response rate) and 9244 adolescent students selected from a nationally representative sample of 320 schools (74.7% response rate). After expositing the logic of dual-frame designs, comparisons are presented of sample and population distributions on Census socio-demographic variables and, in the school sample, school characteristics. These document only minor differences between the samples and the population. The results of statistical analysis of the bias-efficiency trade-off in weight trimming are then presented. These show that modest trimming meaningfully reduces mean squared error. Analysis of comparative sample efficiency shows that the household sample is more efficient than the school sample, leading to the household sample getting a higher weight relative to its size in the consolidated sample relative to the school sample. Taken together, these results show that the NCS-A is an efficient sample of the target population with good representativeness on a range of socio-demographic and geographic variables. *Copyright © 2009 John Wiley & Sons, Ltd.*

Introduction

This paper presents an overview of the design and field procedures of the National Comorbidity Survey Replication Adolescent Supplement (NCS-A), a national survey of DSM-IV mental disorders among adolescents (ages 13–17) in the US. The survey was fielded between February, 2001 and January 2004 as a late add-on to the National Comorbidity Survey Replication (NCS-R; Kessler and Merikangas, 2004), a national household survey of adults. The NCS-A was carried out at the request of the National Institute of Mental Health (NIMH) to meet a request from Congress to provide national data on the prevalence and correlates of mental disorders among US youth. Based on the limited budget, it was decided that a survey of children, which would require parents and teachers to be the main respondents, was infeasible, but that adolescents could be surveyed with a small amount of supplemental information obtained from self-administered parent questionnaires. This was the study design used in the NCS-A. An overview of the rationale for the study is presented elsewhere (Merikangas *et al.*, 2009).

In order to keep the study within budget, we had to use the same interviewers as the NCS-R. Given the heavy training burden on these interviewers, it was decided to use a modification of the NCS-R interview schedule with adolescents rather than the instrument developed in an earlier program of NIMH-funded methodological research (Lahey *et al.*, 1996). The NCS-A collaborators at Yale and NIMH took the lead in making these instrument modifications. As the NCS-R was carried out entirely in English, the NCS-A, too, was limited to English-speaking adolescents.

The number of adolescents residing in NCS-R households was too small to generate the target sample of 10 000 respondents. The sample was consequently supplemented by adding a school-based sample. This had lower costs than household screening (Johnston *et al.*, 2007). The final sample, then, was based on a dual-frame design (Groves and Lepkowski, 1985; Lepkowski and Groves, 1986) in which one sample was recruited from the NCS-R households and the other from a representative sample of schools in the same communities as the NCS-R households. All schools (public and private, schools for gifted children, therapeutic schools, etc.) were included in their true population proportions. A stratified probability sample of students was selected from each school to participate in the survey.

Survey mode

The NCS-A interview was administered face-to-face to adolescents in their homes using laptop computer-assisted personal interviews (CAPI) by professional survey interviewers from the Survey Research Center (SRC) of the Institute for Social Research at the University of Michigan. The decision to use CAPI rather than paper-and-pencil (PAPI) interviews was based on the fact that the interview schedule had many complex skips that create opportunities for interviewer error. These errors are avoided in CAPI. CAPI is also cost-effective when the sample size is as large as in the NCS-A, as the costs of programming are less than the labor needed to keypunch PAPI responses. Parents were asked to complete paper-and-pencil self-administered questionnaires (PSAQ) while their children were being interviewed. In the school sample, Principals and Mental Health Coordinators were asked to complete a self-administered questionnaire (SAQ) describing the school and its mental health resources.

As the NCS-A asked a number of embarrassing questions, audio computer-assisted self-administered interviewing (A-CASI) might have been used instead of CAPI. A-CASI allows respondents to enter answers 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. Considerable evidence shows that A-CASI can lead to significantly higher reports of some illegal and embarrassing behaviors, although the evidence is more mixed for responses to questions about emotional problems (Tourangeau and Smith, 1998; Turner *et al.*, 1998; Turner *et al.*, 1992). Our decision not to use A-CASI was based on the fact that it was not used in the NCS-R, which would have made it difficult to use it in the NCS-A. The decision not to use A-CASI in the NCS-R, in turn, was based on a concern about non-comparability of responses for purposes of trending with the baseline NCS.

The decision to use SAQ rather than interviewer administered surveys to collect parent data was based largely on financial constraints. As the vast majority of the PSAQ data were collected while interviewers were in the homes of respondents completing the adolescent interviews, the marginal costs of the PSAQ data was quite low. Trade-offs were that the PSAQ response rate was lower than if parents had been interviewed and that the amount and subtlety of data collected from parents were limited by the use of SAQ. But these were consequences that were unavoidable based on financial constraints.

In the case of the SAQ data collected from school Principals and Mental Health Coordinators, the number of respondents was small enough that the increased cost of face-to-face data collection was not an issue, but the SAQ was found to be logistically the most efficient way to

collect these data because of the difficulty finding enough time for the Principals and Mental Health Coordinators to complete interviews. In cases where completed SAQ information could not be obtained, respondents were offered the opportunity to provide the information in a telephone interview or in-person interview.

Fieldwork organization and procedures

As noted earlier, the NCS-A fieldwork was carried out by the same professional SRC national field interview staff that carried out the NCS-R. There were 197 interviewers supervised by a team of 18 experienced regional supervisors. A study manager located at the central SRC facility in Michigan oversaw the work of the supervisors and their staff. After sample selection (see later), each interviewer received a folder for each target household. An advance letter was sent to the household a few days before the initial interviewer contact attempt explaining the study and providing an 800 number for questions prior to the interviewer visiting their household. This mailing also included a brief brochure that posed and answered the questions often asked by survey respondents (e.g. How did you select my child? Will all answers be confidential? What will be done with the answers?).

Upon making in-person contact, the interviewer answered questions before obtaining written informed consent from the parent and written informed assent from the adolescent. In the household sample, one random adolescent was selected when more than one resided in the household using a computer-based method in which the names of all resident adolescents in the household were entered into the computer and a routine programmed into the computer selected the random respondent. In the school sample, the adolescent was identified by the school roster. If more than one adolescent in a household was selected in the school sample, which occasionally happened by chance, both were invited to participate. Only after the parent provided signed informed consent was any contact made with the adolescent. Interviews were never conducted with a non-emancipated adolescent unless at least one parent or guardian was present in the home during the interview. However, no parent consent or parent questionnaire was requested in the small number of cases where an emancipated minor was interviewed. Adolescents were given $50 as a token of appreciation for participating in the survey interview, while parents were given $50 for completing the SAQ. School Principals and Mental Health Coordinators were also given $50 each to complete the SAQ describing the school and its mental health resources.

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

Interviewer training and field quality control

Each professional SRC interviewer is required to complete a two-day General Interviewer Training (GIT) course before working on any SRC survey. In addition, experienced interviewers have to complete GIT refresher course at the beginning of every new survey in which they work. Each NCS-A interviewer additionally received a five-day training specific to the NCS-A. Several steps were taken to ensure quality of fieldwork. Sample households were selected centrally to avoid interviewers recruiting respondents from preferred neighborhoods. The computerized Composite International Diagnostic Inverview (CIDI) had a built-in clock to record speed of data entry, making it difficult for interviewers to shorten interviews by skipping sections or filling in sections quickly. Supervisors reviewed each interview within 24 hours of completion to check for a wide range of errors. Supervisors contacted a random 10% of interviewed households to confirm address, enumeration, random selection procedures, interview length, and a random sample of question responses. Completed CAPI interviews were sent electronically to supervisors every night for this purpose. In cases where problems were detected, interviewers were instructed to re-contact the respondent to obtain the missing data.

The sample design

Household sample selection procedures

As noted earlier, the NCS-A household survey was conducted as a supplement to the NCS-R. The NCS-R households that included adolescents were included in the NCS-A. The school sample was recruited from the same sample of counties as the NCS-R. A comprehensive government list of schools was used for selection. The household sample also included adolescents who were not currently enrolled in school. Selection of NCS-R households is described in detail elsewhere (Kessler *et al.*, 2004) and will not be repeated here other than to note that the households were based on a three-stage clustered area probability sampling design that was representative of households in the continental US.

School sample selection procedures

The school sample was selected using the same methods as other SRC school-based surveys (Johnston *et al.*, 2007).

Although school-based samples miss adolescents who have dropped out of school, approximately 96.6% of US adolescents in the age range 13–17 are students (see http://www.census.org), which means that the under-coverage involves only 3.4% of the population in the target age range. In addition, the NCS-R household sample included non-students, which provided some information about how they differ from students. However, the number of non-students was so small in the household sample ($n = 25$) that no precise inferences could be made about this segment of the population. The analysis consequently focused on students both in the household sample and in the school sample. This exclusion is important to keep in mind when considering relatively uncommon disorders that might be highly concentrated among non-students, such as bipolar disorder, where even the exclusion of a mere 3.4% of the population might lead to meaningful under-estimation of prevalence.

In the school sample, a representative sample of middle schools, junior high schools, and high schools was selected with probabilities proportional to the size of the student body in the classes relevant to the target sample (i.e., ages 13–17), in each of the counties or county clusters that made up the primary sampling units (PSUs) of the nationally representative NCS-R sample. The schools were selected from a master file of all licensed schools in each PSU. All accredited schools were eligible, including private and residential schools. In some cases where there were several small schools in a geographic area, those schools were combined to form a cluster that was treated as a single school for purposes of sampling.

Recruitment began by contacting school districts with letters that described the purpose of the study. With the district's approval, individual school Principals were contacted and asked to provide rosters from which to contact student families for study participation. Schools were provided $200 as a token of appreciation for this cooperation. Within each school, a random sample of 40–50 eligible students was selected for sampling. This was done using a systematic selection procedure implemented by the survey firm staff member who obtained access to the school roster. This procedure began with a random start and a systematic selection of every nth student in the roster beginning at the random start, where both the random start and the number n are controlled by a computer program and is used by the survey firm staff member to build the sample. Toward the end of the recruitment period when more schools were needed to complete the study, school payment was increased to $300.

A total of 320 schools participated in the survey. Sample selection began with a target sample of 289 schools initially contacted for participation, of which only 81 agreed. The primary reason given for refusal was reluctance to release student information for research studies. Some schools even had policies against giving out student information. Districts that required formal research proposals usually granted our request eventually, but sometimes with the stipulation that they would only release student information if they first had parental written consent. Schools of the latter typed were generally rejected based on the fact that active initial consent has been shown in previous research to result in a very low response rate (Johnston *et al.*, 2007). In cases where there were no replacement schools readily available, though, this requirement was accepted because there was no choice. This occurred in roughly 15% of schools. As shown later, the response rate was dramatically lower in this sub-sample, which are referred to later as blinded schools because the survey team was blinded to the identities of the sample students until after signed consent was obtained by the school Principals.

Based on the low initial school-level response rate and often protracted time frame of recruitment, multiple replacement schools were recruited for some refusal schools. Replacement schools were selected using standard procedures to match the initial refusal schools in terms of school size, geographic area, and demographic characteristics (Kish, 1987). The fact that the sample ended up with 320 schools rather than the original 289 reflects this expansion of recruitment. In cases where multiple replacement schools were included in the sample for one original school, the total number of interviews targeted in the replacement schools added up to the number targeted for the original school.

A question can be raised whether the high level of replacement of schools led to bias in estimates. As the household sample included respondents who were students in schools that refused to participate in the survey, this question can be investigated empirically. As reported elsewhere, this analysis shows that the use of replacement schools did not introduce bias into estimates of either disorder prevalence or treatment, the two classes of outcomes included in the comparative analysis of students from refusal schools and replacement schools (Kessler *et al.*, 2009a).

Sample disposition

The NCS-A sample disposition is reported in Table 1. The overall adolescent response rate was 75.6%, for a total of

Table 1 NCS-A sample disposition

	Household		Unblinded school		Blinded school		Total	
	Percentage	*n*	Percentage	*n*	Percentage	*n*	Percentage	*n*
I. Adolescents								
Interview	85.9	904[a]	81.8	8912	22.3	332	75.6	10 148
Refusal	11.3	119	14.7	1604	76.4	1137	21.3	2860
Circumstantial	2.4	25	1.9	211	2.9	13	1.9	249
No contact	0.4	4	1.5	165	0.4	6	1.3	175
II. Parents								
Full questionnaire	52.4	551[b]	52.4	5703	15.9	237	48.3	6491
Short-form questionnaire	18.5	195[b]	16.0	1744	3.7	55	14.8	1994
Either	70.9	746	68.4	7447	19.6	292	63.0	8485
Total		1052		10892		1488		13432

[a] Twenty-five of the household survey respondents were not students. The remaining 879 are students.
[b] Fifteen of the parents who completed a questionnaire eight full questionnaire, seven short-form questionnaire were the parents of adolescents who were not students.

10 148 completed interviews. This is made up of response rates of 85.9% (*n* = 904) in the household sample, 81.8% (8912) in the unblinded school sample, and 22.3% (*n* = 332) in the blinded school sample. Non-response was largely due to refusal (21.3%), which in the household and unblinded school samples came largely from parents rather than adolescents (72.3% and 81.0%, respectively). The refusals in the blinded school sample, in comparison, came almost entirely (98.1%) from parents failing to return the signed consent postcard.

The much higher refusal rate in the blinded school sample than the other samples was due to the fact that in blinded schools active written parental consent, in the form of a signed return postcard in response to a letter mailed by the school Principal, was required before the school would release the names and addresses of sample adolescents to the research team. Some 74.9% of parents in blinded schools failed to return these postcards, while another 1.5% of cases were omitted because of refusal on the part of either the parent (0.9%) or the adolescent (0.6%) to participate after a parent had signed the informed consent postcard. As in the blinded school sample, the majority of refusals in both the household sample (72.3%) and the unblended school sample (81.0%) came from parents rather than adolescents.

Consistent with parents being less cooperative than adolescents, the response rate to the parent SAQ was considerably lower than in the adolescent survey: 63.0% com-pared to 75.6%. The parent SAQ response rate could not be higher than the adolescent response rate by design, as parent SAQs were collected only for adolescents who completed interviews. The conditional parent response rate given adolescent response did not differ substantially between the household sample (82.5%; 70.9/85.9), the unblinded school sample (83.6%; 68.4/81.8), and the blinded school sample (87.9%; 19.6/22.3).

Weighting

As noted earlier in the paper, the most recent Census data show that 96.6% of US adolescents in the age range 13–17 are students. It would consequently have been expected that about 31 non-student respondents would be in the household sample (i.e. 3.4% of 904). The actual number was 25. This is too few to support extrapolation to the population of the roughly half million non-student adolescents in the US. The non-student respondents were consequently excluded from the bulk of the analyses, which concentrated on the 10 123 respondents who were students. Weighting focused on the student population. As the sample design involved a dual-frame approach, a distinct weighting scheme was used to make each sample representative of adolescents in the US household population on the cross-classification of a wide range of socio-demographic and geographic variables. The two weighted samples were then merged for purposes of analysis.

The household sample

The household sample weighting was the simpler of the two in that weights had already been developed for the NCS-R household sample. The NCS-R weights are described elsewhere (Kessler *et al.*, 2004) and will not be discussed here. The first step was to add these weights to the adolescent data and adjust them for differential probability of selection of adolescents as a function of number of other adolescents in the household. These doubly-weighted data were then compared with nationally representative Census data on basic socio-demographic characteristics for purposes of post-stratification. Two data files were used for this purpose. The first was the 2000 Census Public Use Microdata Sample (PUMS; http://www.census.gov/support/pumsdata.html) of a 5% sample of the entire US population. Data were extracted from the PUMS for adolescents who were students at the time of the Census. The second was a small area geo-code data file prepared by a commercial firm that aggregated 2000 Census data to the level of the Block Group (BG) for each of the 208 790 BGs (http://www.geolytics.com/resources/us-census-2000.html). These BG-level data were linked to the data record of each NCS-A respondent, while the national distributions for the population on these same BG-level variables were generated by weighting the BG-level data by the population of eligible adolescents in each BG.

A wide range of variables available in the NCS-A as well as in the PUMS or the BG-level data file was selected to post-stratify the NCS-A data. (Details available on request.) In addition, some information was available about variables not in the Census files available for the NCS-A household sample, as the NCS-R was completed in the households of all NCS-A respondents and non-respondents. In particular, comparisons and weighting were made for discrepancies between the DSM-IV/CIDI disorders reported by the adult NCS-R respondents in the households of NCS-A respondents and non-respondents.

The post-stratification weight was created by using an exponential weighting function to make the distributions of post-stratification variables in the adjusted weighted sample agree with the distributions in the external datasets. Specifically, the weight for case k was of the form

$$W_k^\star = W_k \exp(\beta' \mathbf{x}_k), \qquad (1)$$

where $W_k^\star$ is the adjusted weight, W_k is the weight before adjustment, $\mathbf{x}_k$ is the vector of characteristics associated with case k (derived either from the survey data or from

the BGD) including a one for the intercept, and β is a vector of coefficients calculated to satisfy the condition

$$\sum W_k^\star \mathbf{x}_k = \mathbf{X}, \qquad (2)$$

where $\mathbf{X}$ is the vector of population distributions of the post-stratification variables selected from the PUMS and BPS datasets. This procedure is a version of raking calibration, commonly used to adjust surveys to match census data (Deville *et al.*, 1993), but generalized in this case to allow for adjustment using continuous as well as categorical variables. A program written in the R programming language was used to estimate β and to create these calibrated weights. The weights resulted in the distributions of the post-stratification variables in the weighted sample being identical to those in the population datasets, while maintaining the associations among these variables found in the sample.

Some sense of the extent to which post-stratification affected variable distributions can be seen by comparing the distributions of selected post-stratification variables in the sample before versus after weighting. (Table 2) For the most part, the ratios of proportions based on final (F) weights, which equal the actual population proportions found in the databases used for post-stratification, to the corresponding proportions without post-stratification weighting (U) were in the range 0.8–1.2. This means that proportions typically changed by less than 20% of their base. There were some exceptions, though, as illustrated by the fact that the proportion of the population who defined themselves as neither being Non-Hispanic White, Non-Hispanic Black, or Hispanic is only 61% as high in the population (5.0%) as in the unweighted sample before post-stratification (8.2%).

The school sample

Weighting for the school sample was based on weights that controlled for three sets of variables. The first set was extracted from the Quality Education Data (QED) database, a commercially-produced database of the characteristics of all primary and secondary schools in the US (http://www.qeddata.com), controlling to population totals of these variables (weighted by school enrollment) adjusted for discrepancies between the schools included in the sample and the population of all schools in the US. A wide range of school characteristics were examined that included such variables as size, grades covered, type of school (e.g. public versus private, special needs school, K-8 school, junior high school, high school), average size of classroom, average student/teacher ratio, and presence

Table 2 Unweighted and weighted distributions of selected NCS-A post-stratification variables among adolescent student respondents in the NCS-A household sample *n* = 879

	Unweighted (*U*)		Weighted with final post-stratification weights (*F*)		Coefficients[a]	
	Percentage	Standard error	Percentage	Standard error	β	*F/U*[b]
Sex						
Male	48.6	1.4	51.2	1.6	0.11	1.05
Female	51.4	1.4	48.8	1.6	0.00	0.95
Age						
13	22.4	1.5	20.7	1.5	0.00	0.92
14	20.3	1.1	20.9	1.4	0.09	1.03
15	17.6	1.5	20.2	1.8	0.15	1.15
16	21.3	1.5	19.5	1.7	−0.03	0.92
17	18.4	1.3	18.8	1.6	0.00	1.02
Race/ethnicity						
Non-Hispanic White	57.2	2.7	65.6	2.6	0.00	1.15
Non-Hispanic Black	17.1	1.4	15.1	1.4	0.01	0.88
Hispanic	17.5	1.4	14.4	1.7	−0.54	0.82
Non-Hispanic other race	8.2	1.8	5.0	1.0	−0.77	0.61
Family income BG						
Low	24.9	2.2	22.5	2.2	−0.38	0.90
Low-average	25.4	2.6	23.6	3.3	−0.43	0.93
Hi-average	24.6	2.4	25.5	2.7	−0.14	1.04
High	25.1	2.6	28.3	3.5	0.00	1.13
Urbanicity						
Large-mid metro	44.5	4.0	44.5	4.5	0.00	1.00
Urban fringes	29.1	4.0	24.4	3.9	−0.01	0.84
Large-small town-rural	26.4	3.3	31.1	4.8	0.05	1.18
12-Month DSM-IV/CIDI Diagnosis of NCS-R Participant						
Any mood disorder[c]	10.6	1.1	9.4	1.3	0.05	0.89
Any anxiety disorder[d]	17.9	1.2	18.8	1.5	−0.01	1.05
Any impulse-control disorder[e]	5.7	0.8	4.7	0.9	−0.11	0.82

[a] Coefficients are from the exponential log-linear raking model.
[b] The ratio of the unweighted value to the value in the final weight.
[c] Any mood disorder consists of DSM-IV/CIDI diagnoses of bipolar disorder, major depressive disorder, or dysthymic disorder.
[d] Any anxiety disorder consists of DSM-IV/CIDI diagnoses of generalized anxiety disorder, panic disorder, social phobia, or specific phobia.
[e] Any impulse-control disorder consists of a DSM-IV/CIDI diagnosis of IED only Part I diagnoses from the NCS-R could be assessed here.

versus absence of various school programs. The other two sets of variables were the same PUMS and BG-level datasets used in the household sample. The same statistical approach to weighting was used as in the household sample. The within-household probability of selection weights used in the household sample, though, were not needed in the school sample, as schools and students within schools were selected with probabilities proportional to the size of the eligible student body.

As with the household sample, post-stratification did not have dramatic effects on distributions of the post-stratification variables in the school sample. (Detailed

results available on request.) For the most part, relative proportions based on final (*F*) weights compared to unweighted (*U*) data were in the range 0.75–1.25. This means that proportions typically changed by less than 25% of their base. For example, the proportion of adolescents who are non-Hispanic White was estimated to be 55.5% before post-stratification compared to the actual population distribution of 65.6%, a relative increase of 18% (i.e. 65.6/55.5) on this proportion after post-stratification. This general pattern of relatively modest adjustments in proportions held for the vast majority of the post-stratification variables included in the analysis.

Weight trimming

When weights vary greatly relative to the mean, estimates tend to have large standard errors. This, in turn, leads to inefficiency in estimation. It is possible to deal with this problem by trimming extreme weights. There is a trade-off in doing this, though, as weight trimming can lead to bias in 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 (i.e. the increase in bias is greater than the decrease in imprecision).

It is possible to study this trade-off between bias and efficiency empirically in order to select an optimal weight trimming scheme by calculating the mean squared error (MSE) of estimates of substantive importance. This was done by evaluating the effects of weight trimming on 10 prevalence estimates: lifetime and 12-month prevalence estimates of any DSM-IV/CIDI mood, anxiety, externalizing, substance use, and any disorder. As described in detail elsewhere (Kessler *et al.*, 2009b), the DSM-IV diagnoses generated in the NCS-A combine parent and adolescent reports and have good concordance with independent diagnoses based on semi-structured research diagnostic interviews with parents and adolescents by blinded clinical interviewers in an NCA-S clinical reappraisal study. In order to evaluate the effects of weight trimming on prevalence estimates based on the CIDI interviews, MSE for variable *Y* at trimming point *p* was defined as

$$\mathrm{MSE}_{Y_p} = B_{Y_p}^2 + \mathrm{Var}(Y_p), \qquad (3)$$

where B_{Yp} is the bias of the prevalence estimate at that trimming point and $\mathrm{Var}(Y_p)$ is the variance of *Y* at trimming point *p*. An unbiased estimator of $B_{Y_p}^2$ is

$$\hat{B}_{Y_p}^2 = (\hat{B}_{Y_p})^2 - \mathrm{V\hat{a}r}(\hat{B}_{Y_p}), \qquad (4)$$

where $\hat{B}_{Yp}$ is an unbiased estimator of bias and $\mathrm{V\hat{a}r}(\hat{B}_{Yp})$ is the estimated variance of $\hat{B}_{Yp}$. This means that and unbiased estimator for Equation 3 can be rewritten as

$$\mathrm{M\hat{S}E}_{Y_p} = (\hat{B}_{Y_p})^2 - \mathrm{V\hat{a}r}(\hat{B}_{Y_p}) + \mathrm{V\hat{a}r}(\hat{Y}_P). \qquad (5)$$

Each of the three elements in Equation 5 can be estimated empirically for any value of *p* in comparison to an untrimmed estimate (which is assumed to be unbiased), making it possible to calculate MSE across a range of trimming points to determine the trimming point that minimizes MSE for any given variable *Y*. The first term, $(\hat{B}_{Yp})^2$, can be estimated directly as $(Y_p - Y_0)^2$, where Y_0 represents the weighted prevalence estimate of *Y* based on the untrimmed data and Y_p is the weighted prevalence estimate based on data trimmed at trimming point *p*. The other two elements in Equation 5 can be estimated using pseudo-replication (Zaslavsky *et al.*, 2001). In the present case, this was done by generating 84 separate estimates for Y_p at each value of *p* for each of the two samples. The number 84 is based on the fact that the NCS-R sample design has 42 geographic strata (made up of PSUs or, in the case of non-self-representing PSUs, pairs of PSUs) each with two sampling-error calculation units (SECUs; constituting sub-samples within self-representing PSUs and individual PSUs within strata that are made up of multiple non-self-representing PSUs), 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 Y_p is defined as the weighted estimate of *Y* at trimming point *p* in the total sample and $Y_{p(sn)}$ is defined 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 $\mathrm{Var}(Y_p)$ can be estimated as

$$\mathrm{V\hat{a}r}(Y_p) = \mathrm{SUM}_s[(Y_{p(s1)} - Y_p)^2 + (Y_{p(s2)} - Y_p)^2]/2 \qquad (6)$$

$\mathrm{Var}(B_{Yp})$ was estimated in the same fashion by replacing $Y_{p(sn)}$ in Equation 4 with

$$\hat{B}_{Yp(sn)} = Y_{p(sn)} - Y_{0(sn)} \text{ and replacing } Y_p \text{ with } \hat{B}_{Yp(sn)} = Y_p - Y_0$$

This method was used to evaluate the effects of trimming between 1% and 10% of respondents at each tail of

the weight distribution in each of the two samples. Trimming consisted of assigning the weight at the trimming point to all cases with more extreme weights on that tail of the weight distribution. The weighting analysis described in Equation 1 and 2 was replicated anew for each combination of trimming points on the two tails so as to obtain an accurate post-stratification of the weighted sample to the population. Prevalence estimates and their design-based standard errors, which were estimated using the Taylor series method (Wolter, 1985), were then calculated for each of the 10 variables used in the analysis of bias-efficiency trade-off. Inspection of empirical variation in MSE with changes in trimming rules was used to select final trimming rules that were used to generate the results in Table 2.

In both samples, MSE was not strongly affected by trimming. Final trimming rules were consequently chosen that trimmed the minimum proportion of cases while approximating the minimum average MSE across all possibilities considered. In the household sample, no weight trimming was performed for low weights but the highest 2.5% of weights were trimmed. This reduced the coefficient of variation of weights (the ratio of the standard deviation of weights to the mean weight) by about 8%. This was achieved with a roughly 2% increase in MSE due to bias, for a total reduction in MSE of approximately 6%. In the school sample, the bottom 2.9% and upper 0.1% of weights were trimmed, reducing the coefficient of variation of weights by about 9%. This was achieved with a nearly 4% increase in MSE due to bias, for a total reduction in MSE of approximately 5%.

Weighting the parent sample

The weights described so far were developed for the full samples. Weights were similarly calculated for the sub-samples of cases with parent data to make possible analyses requiring these responses. To make these samples nationally representative with respect to the weighting variables, the weighting analyses described earlier was replicated by treating the total sample as the 'population' and the sub-sample of cases with parent SAQ data as the 'sample.' The post-stratification control variables included all those used in the full-sample analyses in addition to the lifetime and 12-month prevalence estimates in the total sample of DSM-IV/CIDI mood, anxiety, impulse-control, and substance disorders. By controlling for the presence of diagnoses adjustments were made for possible tendencies of parents to be either more or less likely to respond to the SAQ when their children had certain types of diagnoses. At the same time the national representa-

tiveness of the full sample with respect to demographic and school characteristics was retained. This re-weighting was carried separately in the household and school samples and, within each of these samples, in the sub-samples with full SAQ data and either full or partial SAQ data. The final trimmed weights from the total sample were included as base weights in these analyses and no further trimming was done when the post-stratification weights were applied to the data.

Combining the weighted household and school samples

The research team plans to carry out substantive analyses of the NCS-A data largely in a consolidated sample that combines the household and school samples. Some decision about relative weighting is needed to do this combining. The obvious approach is to transform the weights in each sample to sum to the number of respondents in the sample and then combine these two weighted data files into a single file. However, this approach implicitly assumes that the two samples have the same efficiency. This assumption turns out to be incorrect, as shown by the fact that the H/S ratio of design-based variance estimates of various descriptive measures in the household sample (H) relative to the school sample (S) is generally lower than the roughly 10.5:1 ratio of the two sample sizes (9244 versus 879) which means that the NCS-A household sample is more efficient for this set of estimates than the NCS-A school sample (Table 3). The reason for this is that the NCS-A household sample has less clustering than the school sample because the number of adolescent student respondents in the household sample ($n = 879$) is smaller than the number of area segments ($n = 1001$). In the case of the school sample, in comparison, the number of adolescent respondents ($n = 9244$) is nearly 30 times larger than the number of schools ($n = 320$), which means that there is considerable clustering at the segment level.

Based on these results, the approach taken to combine the household and school samples into a single larger consolidation sample gave higher weight to the household sample in recognition of the greater efficiency of the household sample than the school component. This approach is based on the goal of combining the two samples into a consolidated dual-frame sample that minimizes the overall MSE of estimates, which is achieved when the two samples are weighted inversely proportional to their MSEs (Lepkowski and Groves, 1986). Based on the results reported in Table 3, this was done by assuming that the variance of estimates average six times higher in

Table 3 Ratios of design-based variance estimates of selected descriptive statistics in the household sample *H* relative to the school sample *S*

	Variance ratios		
	Mean	Trimmed mean[a]	Median
I. DSM-IV/CIDI disorder prevalence estimates[b]			
Prevalence	6.9	6.2	7.5
II. Predictive effects of socio-demographic variables on disorder prevalence[c]			
Age	7.6	6.6	7.1
Sex	5.9	6.0	5.1
Race-ethnicity	4.1	3.2	2.6
Parent education	5.1	5.0	4.8

[a] Mean among observations in the 25th–75th percentile range on the distribution of variance ratios.
[b] Lifetime, 12-month, and 30-day prevalence estimates of any DSM-IV/CIDI mood disorder, anxiety disorder, impulse-control disorder, substance disorder, and any disorder.
[c] Based on multivariate logistic regression equations to predict each of the 15 outcomes in Part I of the table.

the household sample than the school sample, which means that we constructed the consolidated sample so that the sum of weights in the school sample was six times that of the sum in the household sample. Combined samples were created using this same weighting approach for the PSAQ student sample and the short-form PSAQ sample.

Analysis with combined and separate samples

Although the bulk of NCS-A analyses will be carried out with the consolidated dataset, we also plan to carry out sensitivity analyses of critical results in the separate household and school sub-samples because a criticism could be raised that the school sample does not represent the population as well as the household sample based on the fact that the majority of the schools originally selected to participate in the NCS-A school sample did not participate and had to be replaced. The household sample, in comparison, had a high adolescent response rate (85.9%). It would be comforting to find that substantive results found in the combined sample could be replicated in the household sample as well as in the school sample.

Design effects

Although the effects of weighting and 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, 1965) for a number of variables of interest. The DE is the square of the ratio of the design-based standard error of a descriptive statistic divided by the simple random sample standard error. The design-based standard error can be calculated using a number of methods (Wolter, 1985), each of which takes into consideration information about the clustering and weighting of the data. 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. DEs due to clustering are usually a good deal larger in estimating prevalence and other first-order statistics than more complex statistics, as 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. DEs due to weighting are also usually somewhat smaller for multivariate than bivariate descriptive statistics because DEs are due not only to the variance of the weights but also to the strength of the association between the weights and the substantive variables under consideration.

Because means typically have higher DEs than other statistics, evaluations of DEs typically focus on the estimation of means. However, we also examined associations of three socio-demographic variables (age, sex, and a dichotomy for non-Hispanic White race-ethnicity versus all others) with the disorder clusters. The latter included 30-day, 12-month, and lifetime prevalence of any DSM-IV/CIDI anxiety disorder, mood disorder, impulse-control disorder, substances disorder, and any disorder (five classes of disorder in each of three time frames). The DEs for prevalence are in the range 1.4–1.6 in the household sample, 3.1–4.6 in the school sample, and 3.3–4.5 in the combined sample (Table 4). The DEs for the associations of socio-demographic variables with the disorders in the household sample are similar to those for the prevalence estimates (1.4–1.7), while those in the school sample are lower than for the prevalence estimates (2.9–3.5). The same is true for the DEs for the associations in the combined sample (2.4–2.9). The DEs are consistently lower for estimates involving 30-day disorders than 12-month or lifetime disorders because less common outcomes generally have lower DEs because multiple cases of

Table 4 Design effects[a] for prevalence estimates of DSM-IV/CIDI disorder clusters[b] and for associations[c] between socio-demographic variables and these clusters in the NCS-A household sample, school sample, and combined sample

	Household	School	Combined
Prevalence estimates of DSM-IV/CIDI disorders[b]			
30-day	1.5	3.1	3.3
12-month	1.6	4.6	4.1
Lifetime	1.4	4.3	4.5
Socio-demographic associations[c]			
30-day	1.7	2.9	2.4
12-month	1.4	3.5	2.9
Lifetime	1.4	3.5	2.9
Total	879	9244	10 123

[a] Design effects are the squares of the ratios of the standard errors of design-based estimates and estimates based on the assumption of simple random sampling. See the text for a more detailed discussion of the substantive interpretation of design effects.
[b] The five DSM-IV/CIDI disorder clusters considered in each of three time frames 30-day prevalence, 12-month prevalence, and lifetime prevalence are any anxiety disorder, any mood disorder, any impulse-control disorder, any substance disorder, and any disorder.
[c] Associations were estimated in logistic regression equations that used information about respondent age, sex, and race-ethnicity non-Hispanic White versus all others to predict each of the five outcomes in each of the three time frames.

these outcomes seldom occur in a single SECU, leading to low clustering. Because of this fact, we can expect the DEs associated with the prevalence and correlates of individual disorders to be lower than those reported here for disorders clusters.

It is important to recognize that the above calculations did not take into consideration the fact that post-stratification weighting improves the extent to which the sample is representative of the population with respect to post-stratification variables compared to a simple random sample. As a result, design effects are over-estimated to an unknown degree in the results reported in Table 4. This bias could be corrected by using a pseudo-replication simulation approach to estimate DE and building in the post-stratification to each replicate. When we use pseudo-replication to estimate design effects, as we do for highly non-linear statistics where the linearization

assumption of the Taylor series method might be violated, we use the jackknife repeated replications (JRR) method of pseudo-replication (Kish and Frankel, 1974). As described in more detail elsewhere (Kessler *et al.*, 2004), we work with 76 JRR pseudo-samples in the NCS-R and NCS-A. This means that we estimate coefficients of interest 76 separate times, once in each pseudo-sample, and then use information about the distribution of the coefficient across the pseudo-samples to estimate design effects. The positive effects of post-stratification could be built into this procedure by developing post-stratification weights for each pseudo-sample, which would decrease variation across the pseudo-samples to some degree and reduce the empirical estimates of design effects appropriately. We did not do this, though, based on the fact that it would be labor-intensive to develop 76 separate post-stratification weighting schemes and our past experience has been that this exercise only has modest effects in decreasing estimates of design effects.

Model-based versus design-based estimation

The weights described earlier were developed in order to support a program of substantive data analysis based on 'design-based' estimation of descriptive and inferential statistics; that is, estimation that attempts to make the sample representative of the population with respect to weighting variables and to make standard errors of survey estimates accurate by using information about the sample design (i.e. clustering and weighting) to adjust for discrepancies between the sample and the population in estimating descriptive statistics and to adjust for discrepancies between the sample design and a simple random sample in estimating inferential statistics (Wolter, 1985). The alternative to design-based estimation is 'model-based' estimation, in which inferences are made by building a statistical model that attempts to include all variables needed to adjust for discrepancies between the sample and the population, including controls for weights and sample clusters (DuMouchel and Duncan, 1983).

If clusters or weights are judged based on appropriate analyses not to contribute meaningfully to the prediction of substantive outcomes and not to have meaningful correlations with substantive predictors in model-based analyses, these design variables can be deleted as controls in the prediction equations, leading to an increase in the precision of estimates and to substantially better precision than in design-based analyses (Gelman, 2007). Meaningful interactions between substantive predictors and variables that define either clusters or weights can be included in prediction equations. However, the

interpretation of these coefficients becomes very complex when many such interactions exist, in which case design-based estimation becomes more attractive.

It has been our experience in the past that many complex interactions exist between substantive predictors of mental disorders and design variables (i.e. clusters and weights) in community epidemiological surveys. As a result, we have based the bulk of our substantive analyses of design-based rather than model-based methods. However, these experiences have largely been based on surveys of adults. To investigate this possibility in the NCS-A, we carried out preliminary analyses of associations of NCS-A clusters (strata and SECUs) and weights as predictors of the same 10 DSM-IV/CIDI classes of mental disorders as considered in the analysis of weight trimming. We also examined these design variables as modifiers of the predictive effects of several basic socio-demographic predictors of these disorders, including age, sex, race-ethnicity, and parental education (Table 5). The χ^2 values of main predictive effects show that the clusters (83 dummy predictor variables for strata-SECUs) are significant predictors of all 10 outcomes, while weights are not. Weights are involved, though, as significant modifiers of the associations between socio-demographics and the outcomes in 17.5% of the cases examined (i.e. 7/40 of the associations of the four socio-demographic variables with the 10 outcomes). This is much more than the 5% we would expect on the basis of chance alone. Significant interactions of socio-demographics with clusters are even more common, occurring in 60% of the cases examined (24/40). These results strongly suggest that substantial complexities would arise in attempting to use model-based methods to estimate substantive associations with the NCS-A data.

Optimizing the design for a fixed budget

We have discussed budget constraints several times earlier as providing a rationale for various design decisions. It is worth noting in this regard that survey methodologists have developed formal procedures for optimizing survey designs for a fixed budget (Kish, 1987). These procedures, though, require prior information to be available on the accuracy of data collected from alternative sources (in our case, adolescents, parents, and possibly even teachers), using various procedures (in our case, self-report questionnaires, fully-structured diagnostic interviews, and semi-structured clinical interviews), the associations among these reports, and the relative costs of collecting data of each sort (Groves, 1989). We did not have access to such data in designing the NCS-A. In addition, we had

the constraint that the sample of adolescents had to be 10 000, constraining us in our design options.

It is possible, though, to carry out an analysis of design optimization *post hoc* in an effort to guide future researchers. We did this for the adult NCS-R survey and found that the optimal design to estimate the prevalence of clinical diagnoses (that is, diagnoses based on the SCID clinical reappraisal interviews rather than on the CIDI) would have reduced the sample of CIDI interviews from roughly 9000 to roughly 7000 and increased the number of SCID clinical reappraisal interviews to about 2000 (Kessler *et al.*, 2004). It is noteworthy that the optimal NCS-A was not to eliminate CIDI interviews entirely and to carry out only SCID interviews. This is because the CIDI was found to contain information that predicted SCID diagnoses strongly at a cost considerably less than the cost of administering a SCID interview.

We are constrained in carrying out a similar *post hoc* analysis of design optimization in the NCS-A because we have no information about the implications of the most obvious design change: carrying out either face-to-face or telephone interviews with parents that assessed all the DSM-IV disorders considered in the survey rather than using self-administered parent questionnaires to assess only a subset of these diagnoses. It might be that the optimal fixed-cost design would have been one that reduced the sample size below (perhaps substantially so) the target of 10 000 and included interviews with parents. It is also possible that optimal allocation of resources to minimize mean-squared error of K-SADS diagnoses would have resulted in a decrease in the number of respondents (both parents and youth) administered fully-structured CIDI interviews and increased the number that received semi-structured K-SADS clinical reappraisal interviews. It would be valuable for formal analyses of these alternatives to be undertaken using available data from existing surveys where all these elements are in place. We suspect that this exercise will show that the optimal design for estimating prevalence based on clinical assessments and estimating correlates of clinical diagnoses would be one that included interviews with parents and a somewhat lower ratio of CIDI to K-SADS interviews than in the NCS-A.

Overview

This paper presented an overview of the NCS-A survey design and field procedures. The design allowed us to gather data from a national sample of adolescents and schools weighted to be representative of the population on a wide range of socio-demographic, school, and

Table 5 Chi-square values of the main effects and interactions of design effects and socio-demographic variables in predicting lifetime and 12-month disorder in the NCS-A $n = 10\,123$[a]

	Lifetime disorders					Twelve-month disorders					df
	Mood[b]	Anxiety	Impulse-control[b]	Substance	Any[b]	Mood[b]	Anxiety	Impulse-control[b]	Substance	Any[b]	
I. Main effects of clusters and weights											
Clusters	140.3*	260.7*	157.8*	356.1*	186.9*	126.3*	210.3*	144.5*	262.9*	175.2*	83
Weights	0.1	1.7	0.4	0.1	3.7	0.3	2.9	2.0	0.0	3.4	1
II. Main effects of demographics											
Age	79.8*	10.5*	12.8*	589.7*	79.5*	53.7*	12.4*	1.2	363.0*	54.2*	1
Sex	52.4*	140.8*	68.0*	32.6*	2.0	62.1*	179.6*	32.9*	30.9*	11.6*	1
Race-ethnicity	18.7*	50.5*	18.9*	79.7*	26.9*	24.1*	49.8*	12.4*	62.4*	19.3*	3
Parent education[c]	8.3*	57.9*	76.9*	23.9*	63.6*	9.6*	61.9*	64.4*	13.8*	59.9*	3
III. Interactions involving weights											
Age	0.0	4.3*	0.1	4.9*	0.6	0.4	7.0*	0.5	4.5*	3.1	1
Sex	0.3	0.1	2.3	3.9*	2.1	0.9	0.2	1.1	4.0*	0.3	1
Race-ethnicity	3.0	3.2	1.7	0.5	0.4	1.1	2.0	1.7	1.5	0.8	3
Parent education	5.4	5.9	1.6	1.3	0.5	3.0	10.6*	0.5	1.8	0.2	3
IV. Interactions involving clusters											
Age	198.2*[d]	94.1	114.3*	777.2*[d]	99.0	168.1*[d]	91.1	147.5*[d]	522.4*[d]	86.2	83
Sex	193.7*[d]	99.4	75.3	380.2*[d]	71.3	188.5*[d]	82.7	80.1	287.4*[d]	87.3	83
Race-ethnicity	284.9*[d]	447.3*[d]	307.8*[d]	298.5*[d]	320.5*[d]	285.1[d]	386.7*[d]	274.9[d]	178.2[d]	327.9*[d]	249
Parent education	352.2*[d]	254.9	367.7*[d]	162.2[d]	430.2*[d]	291.9*[d]	479.6*[d]	326.5*[d]	347.5*[d]	419.2*[d]	249

[a] There are $n = 10\,123$ observations with weights and design effects. Each effect is looked at through a bivariate model.
[b] These disorders use information parent SAQs and consequently are limited to the 6483 cases with completed PSAQs
[c] Parent education is defined as the maximum number of years of education of either parent. Categories are: 0–11 years, 12 years, 13–15 years, 16+ years
[d] Due to sparse data, some models were run with a forward regression logistic model that stepped in two-way interactions included at an entry criterion of $p = 0.95$ i.e. even very small coefficients were entered. The number of degrees of freedom in these models is lower than in the saturated model.
* Significant at the 0.05 level.

geographic characteristics. Although less desirable than interviewer-administered survey data, the parent SAQ data were obtained very cost-effectively and provide valuable collateral information about family history, developmental milestones, and externalizing disorders of the adolescent respondents. The SAQ data provided by Principals and Mental Health Coordinators, furthermore, provide information that might prove to be valuable in expanding our understanding of the ways in which school characteristics influence detection and response to adolescent mental disorders. Innovative methods of post-stratification, weight trimming, and combination of the household and school samples improve the representativeness and efficiency of the consolidated sample.

An important limitation of the NCS-A is the relatively low response rate of schools in the school sample and of individual respondents in the blinded school sub-sample. The response rate of adolescents in the household sample was considerably higher. Because of this between-sample difference, we will carry out sensitivity analyses separately in the household and school samples. Consistency of results across the two samples will be taken as an indication of robustness of findings.

Despite the limitation imposed by the low response rate of schools, the data on comparisons of sample and population characteristics at the level of the individual with Census socio-demographic characteristics and at the level of the school with administrative databases are very encouraging regarding the representativeness of the sample. The rich substantive information contained in the NCS-A will allow many analyses to be carried out to increase our understanding of the health and well-being of adolescents in the US.

Declaration of interest statement

Portions of the paper and Table 1 appeared previously in Kessler, R.C., Avenevoli, S., Costello, E.J., Green, J.G., Gruber, M.J., Heeringa, S., Merikangas, K.R., Pennell, B-E., Sampson, N.A., Zaslavsky, A.M. (2009) The National Comorbidity Survey Adolescent Supplement (NCS-A): II. Overview and design. *Journal of the American Academy of Child and Adolescent Psychiatry*. Copyright Lippincott Williams & Wilkins. Used with permission.

Competing interests

Dr Kessler has been a consultant for GlaxoSmithKline Inc., Kaiser Permanente, Pfizer Inc., Sanofi-Aventis, Shire Pharmaceuticals, and Wyeth-Ayerst; has served on advisory boards for Eli Lilly & Company and Wyeth-Ayerst; and has had research support for his epidemiological studies from Bristol-Myers Squibb, Eli Lilly & Company, GlaxoSmithKline, Johnson & Johnson Pharmaceuticals, Ortho-McNeil Pharmaceuticals Inc., Pfizer Inc., and Sanofi-Aventis. The remaining authors report no competing interests.

References

Deville J.C., Sarndal C.E., Sautory O. (1993) Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, **88**, 1013–1020.

DuMouchel W.H., Duncan G.J. (1983) Using sample survey weights in multiple regression analyses of stratified samples. *Journal of the American Statistical Association*, **78**, 535–543.

Gelman A. (2007) Struggles with survey weighting and regression modeling. *Statistical Science*, **22**, 153–164, DOI: 10.1214/088342306000000691

Groves R.M. (1989) *Survey Errors and Survey Costs*, John Wiley and Sons.

Groves R.M., Lepkowski J.M. (1985) Dual frame, mixed mode survey designs. *Journal of Official Statistics*, **1**, 263–286.

Johnston L.D., O'Malley P.M., Bachman J.G., Schulenberg J.E. (2007) *Monitoring the Future National Results on Adolescent Drug Use: Overview of Key Findings, 2006*, NIH Publication No. 07-6202, National Institute on Drug Abuse.

Kessler R.C., Merikangas K.R. (2004) The National Comorbidity Survey Replication (NCS-R): background and aims. *International Journal of Methods in Psychiatric Research*, **13**: 60–68.

Kessler R.C., Avenevoli S., Costello E.J., Green J.G., Gruber M.J., Heeringa S., Merikangas K.R., Pennell B., Sampson N.A., Zaslavsky A.M. (2009a) The National Comorbidity Survey Adolescent Supplement (NCS-A): II. Overview and design. *Journal of the American Academy of Child and Adolescent Psychiatry*, **48**(4), 380–385.

Kessler R.C., Avenevoli S., Greif Green J., Gruber M.J., Heeringa S., Guyer M., He Y., Jin R., Kaufman J., Merikangas K.R., Sampson N.A., Zaslavsky A.M. (2009b) The National Comorbidity Survey Adolescent Supplement (NCS-A): III. Concordance of DSM-IV/CIDI diagnoses with clinical reassessments. *Journal of the American Academy of Child and Adolescent Psychiatry*, **48**(4), 386–399.

Kessler R.C., Berglund P., Chiu W.T., Demler O., Heeringa S., Hiripi E., Jin R., Pennell B.E., Walters E.E., Zaslavsky A., Zheng H. (2004) The US National Comorbidity Survey Replication (NCS-R): Design and field procedures. *International Journal of Methods in Psychiatric Research*, **13**, 69–92.

Kish L. (1965) *Survey Sampling*, John Wiley & Sons.

Kish L. (1987) *Statistical Design for Research*, John Wiley and Sons.

Kish L., Frankel M.R. (1974) Inferences from complex samples. *Journal of the Royal Statistical Society*, **36**, 1–37.

Lahey B.B., Flagg E.W., Bird H.R., Schwab-Stone M.E., Canino G., Dulcan M.K., Leaf P.J., Davies M., Brogan D., Bourdon K., Horwitz S.M., Rubio-Stipec M., Freeman D.H., Lichtman J.H., Shaffer D., Goodman S.H., Narrow W.E., Weissman M.M., Kandel D.B., Jensen P.S., Richters J.E., Regier D.A. (1996) The NIMH Methods for the Epidemiology of Child and Adolescent Mental Disorders (MECA) study: Background and methodology. *Journal of the American Academy of Child and Adolescent Psychiatry*, **35**, 855–864.

Lepkowski J.M., Groves R.M. (1986) A mean square error model for dual frame, mixed mode survey design. *Journal of the American Statistical Association*, **81**, 930–937.

Merikangas K.R., Avenevoli S., Costello E.J., Koretz D., Kessler R.C. (2009) The National Comorbidity Survey Adolescent Supplement (NCS-A): I. Background and measures. *Journal of the American Academy of Child and Adolescent Psychiatry*, **48**(4), 367–369.

Tourangeau R., Smith T.W. (1998) Collecting sensitive information with different modes of data collection. In: *Computer Assisted Survey Information Collection* (eds Couper M, Baker R, Bethlehem J, Clark C, Martin J, Nicholls W, O'Reilly J), pp. 431–454, John Wiley & Sons.

Turner C.F., Ku L., Rogers S.M., Lindberg L.D., Pleck J.H., Sonenstein F.L. (1998) Adolescent sexual behavior, drug use, and violence: Increased reporting with computer survey technology. *Science*, **280**, 867–873.

Turner C.F., Lessler J.T., Devore J.W. (1992) Effects of mode of administration and wording on reporting of drug use. In *Survey Measurement of Drug Use: Methodological Studies* (eds Turner CF, Lessler JT, Gfroerer JC), DHHS Publication No. ADM 92-1929, pp. 177–220, National Institute on Drug Abuse.

Wolter K.M. (1985) *Introduction to Variance Estimation*, Springer-Verlag.

Zaslavsky A.M., Schenker N., Belin T.R. (2001) Downweighting influential clusters in surveys: Application to the 1990 Post-Enumeration Survey. *Journal of the American Statistical Association*, **96**, 858–869.