Mental Health Estimates Computed Directly from the Clinical Sample of the Mental Health Surveillance Study and Measures of their Standard Errors

Phillip S. Kott[1], Jonaki Bose[2], Sarra Hedden[2], Dan Liao[1], Art Hughes[2]

Abstract

A clinical follow-up study to the National Survey on Drug Use and Health (NSDUH) collected information on specific mental disorders among adults that could be used to provide national and state estimates of serious mental illness. Specifically, a nationally representative subsample of adult respondents to the NSDUH was interviewed by trained clinicians over the telephone using a psychiatric diagnostic interview between 2008 and 2012. In order to estimate the prevalence of mental health disorders among adults in the U.S., weights were created for the clinical subsample. The weighting procedures included a nearly pseudo-optimal "poststratification" to non-mutually exclusive control totals from the NSDUH interview. This use of data from the entire NSDUH sample in weight creation resulted in estimates with increased accuracy. Both the nearly pseudo-optimal poststratification and improved standard error measures for the resulting estimates were completed using the WTADJX procedure in SUDAAN 11.

Keywords: Nearly pseudo-optimal, standard error, bounded weights, WTADJX, mental health

1. Introduction

The National Survey on Drug Use and Health (NSDUH), conducted by the Substance Abuse and Mental Health Services Administration (SAMHSA), is one of the primary sources of data for population-based prevalence estimates of substance use and mental health indicators in the United States. The NSDUH interview includes several self-administered indicators of mental health, such as assessments of lifetime and past year major depressive episode (MDE), past month and past year general psychological distress and associated functional impairment, as well as past year suicidality. Additionally, from 2008 to 2012, SAMHSA added a clinical component to the Mental Health Surveillance Study (MHSS). Clinicians administered semi-structured diagnostic interviews to a subsample of NSDUH adult respondents to assess the presence of selected mental disorders.

The purpose of the MHSS clinical component was to develop a statistical model to apply to the full NSDUH sample that would generate serious mental illness (SMI) prevalence estimates among adults (aged 18 years or older) at national and State levels and to monitor the prevalence of SMI over time. The most recent prevalence estimates of SMI among adults that have been generated using NSDUH data are available in the 2012 NSDUH mental health findings report (CBHSQ, 2013).

In addition to producing a model for the NSDUH to yield model-based estimates of SMI among adults (CBHSQ, in press), the 2008-2012 MHSS clinical data can be used to generate prevalence estimates of past year mental disorders among the adult civilian, non-institutionalized population across a wide spectrum of diagnostic categories, including mood disorders (major depressive disorder [MDD], bipolar I disorder, and/or dysthymic disorder), anxiety disorders (posttraumatic stress disorder [PTSD], panic disorder with and without agoraphobia, agoraphobia without history of panic disorder, social phobia, specific phobia, obsessive compulsive disorder [OCD], and/or generalized anxiety disorder [GAD]), eating disorders (anorexia nervosa and/or bulimia nervosa), substance use disorders (alcohol abuse, alcohol dependence, illicit drug abuse, and/or illicit drug dependence), intermittent explosive

1. RTI International, 6110 Executive Blvd., Suite 902, Rockville, MD 20852, U.S.A
2. Substance Abuse and Mental Health Administration, 1 Choke Cherry Rd, Rockville, MD, 20857, U.S.A.

disorder, adjustment disorder, as well as psychotic symptoms (delusions and/or hallucinations). Karg *et al.* (in press) presents the past 12-month prevalence estimates of specific mental disorders using the MHSS clinical data.

This paper focuses on how the prevalence estimates and their standard errors were derived from the 2008-2012 MHSS clinical sample. In particular, it describes how the prevalence estimates in Table 1 (featured at the end of this Introduction) covering the 2008-2012 time period were computed using sampling weights that had undergone a number of calibration adjustments with an emphasis on the last one – the annual calibration of the clinical sample to the NSDUH control totals (i.e. the "poststratification" adjustment). It then discusses several alternative methods for measuring the standard errors of those estimates, which are also presented in the table. All these methods use linearization variance estimators. The first ("not corrected") ignores the impact of the calibration weighting entirely, which is what has been done historically with NSDUH main-sample estimates. The second method ("fully corrected-internal") is how SAMHSA computed standard errors for the 2008-2012 MHSS clinical sample, such as Karg *et al.* (in press), that accounted for the annual poststratification adjustment in weighting.

The third method ("fully corrected-external") is how many users may choose to compute standard errors with the 2008-2012 NSDUH adult clinical interview data file that is available to the public. It is easier to implement than the second and returns very similar results. The fourth and fifth methods were also investigated, but for reasons discussed later in this paper, found less suitable for use.

The concentration in this paper will be on the statistical rather than measurement issues. That is to say, the statistical analyses to be discussed in this paper accept the diagnostics made by the mental-health professional during the clinical interview using the Structured Clinical Interview for the DSM-IV-TR Axis I Disorders, Research Version, Non-patient Edition (SCID-I/NP) (First, Spitzer, Gibbon, & Williams, 2002) as accurate.

In addition, these analyses treat the yearly clinical samples as pure random samples – of both adults and time periods – with probabilities of selection accurately captured by the sample weights before the final poststratification). These selection probabilities incorporate the self-selection of unit response (rather than nonresponse) and the impact of the deliberate exclusion from the clinical sample of adults who responded to the NSDUH main interview in Spanish. SAMHSA compensated for the latter by assigning clinically-interviewed Hispanics a probability of having responded to the NSDUH main interview in English based on their characteristics. Older, less educated Hispanics with fewer years in the U.S. tended to respond to the NSDUH main interview in Spanish. When such Hispanics participated in the clinical interview, they were assigned relatively small selection probabilities and consequently relatively large sampling weights so that they effectively represented those Hispanics who responded to the NSDUH main interview in Spanish.

More details on the probability sampling and weighting process can be found in Liao *et al.* (in press). Briefly, annual clinical sample weights were the product of six factors: the respondent's NSDUH main-interview analysis weight, a coverage adjustment to compensate for NSDUH main-survey respondents who completed that survey in Spanish, the inverse of the probability the respondent was also selected for the clinical sample (the selection probability into the clinical sample was an independent function of an adult's NSDUH main-interview responses, which varied across the years), a refusal adjustment to compensate for NSDUH respondents selected for clinical evaluation who did not wish to be recontacted, a second nonresponse adjustment to compensate mostly for those who agreed to be recontacted but were unavailable for the clinical evaluation (this included a few who agreed to be recontacted for the evaluations, but refused to respond when recontacted), and a poststratification adjustment to increase the efficiency of direct estimates from the clinical sample. Strictly speaking this adjustment is a calibration to total computed from the NSDUH main-interview respondents, but the procedure in SUDAAN 11 (RTI International 2012) we use calls it a "poststratification" (even though the totals computed from the NSDUH main-interview responses were not for mutually exclusive groups).

This document focuses on that last annual weighting adjustment. Before proceeding, other features of the MHSS require some discussion. The first is that the adult NSDUH main sample in 2008

was randomly divided into two halves. One half sample, denoted the 2008A sample, was assigned functional impairment questions based on an abbreviated version of the World Health Organization Disability Assessment Schedule (WHODAS; Rehm et al., 1999). The other half sample, the 2008B sample, was assigned questions based on the Sheehan Disability Scale (SDS; Leon, Olfson, Portera, Farber, & Sheehan, 1997). Both halves received psychological distress questions based on the Kessler 6 scale (K6; Kessler et al., 2003). From 2009 onward, only the WHODAS and K6 questions were used on the NSDUH main survey.

Weights were constructed separately each year, treating the 2008A and 2008B clinical samples as if they represented distinct years. Although ideally it would be preferable to construct direct estimates for the clinical sample every year, the small respondent sample sizes: 759 in 2008A, 741 in 2008B, 520 respondents from 2009, 516 in 2010, 1,495 in 2011, and 1,622 in 2012, produced yearly estimates with standard errors deemed to be unacceptably large. The clinical sample began in 2008 and was discontinued after 2012; so no additional years of data are available.

Consequently, the clinical samples were combined across the years to generate prevalence estimates of mental disorders. Because the sample size, sampling allocation, and weight adjustments for the clinical sample differed from year to year, gains in statistical efficiency could be realized by scaling the weights.

The scaling factors were determined by focusing on the standard errors of prevalence estimates for SMI, any mental illness, and the occurrence of major depressive episode in the previous year. They were 0.06 for 2008A and 2008B, 0.04 for 2009, 0.14 for 2010, 0.35 for 2011, and 0.35 for 2012. A discussion of the assumptions underlying the use of these factors and their implications on the estimation of prevalences for specific mental disorders over the 2008 to 2012 time period is contained in Section 2.

Section 3 describes how the clinical sample was calibrated to the NSDUH main sample each year in a nearly pseudo-optimal fashion (Kott 2011). Section 4 shows how WTADJX routine in SUDAAN 11 (RTI 2012) was used to estimate yearly standard errors for prevalences. As noted earlier, this section treats the weights before the final calibration as pure probability-sampling weights. Kott and Day (2013) argue that this treatment will, if anything, tend to overestimate standard errors.

Section 5 describes how the standard-error measures for prevalence estimates in Table 1 were calculated and discusses the table's implications. Like when calculating the prevalence estimates themselves, the weights for each year (2008A, 2008B, 2009, 2010, 2011, and 2012) were scaled in the computation of the standard-error measures.

Clinical MHSS data sets being made available to qualified researchers will not contain identifiers for the year of the interview. This is one reason why the final two standard-error measures in Table 1 were calculated. Both pretend that there was a single calibration across all years with the NSDUH main-sample calibration targets either calculated within the WTADJX procedure ("internal") or not ("external"). The latter simplifies standard-error computation since it does not require that data from the full adult NSDUH main respondent sample be included in the calculations.

Standard errors for prevalence estimates computed from the NSDUH main sample were produced using 900 variance strata, each containing two variance primary sampling units (PSUs; these are sometimes called "replicates"). For producing standard errors for estimates from the clinical sample, the NSDUH variance strata were collapsed into 100 MHSS variance strata so that there would be at least one respondent within each variance PSU. For analyzing the clinical sample a year at a time, a further random collapsing into 50 variance strata was carried out. This further collapsing was used in developing the clinical weights and the model the serious mental illness (see Liao *et al.*, in press). Section 6 compares the standard errors computed using this set of variance strata with some of those in Table 1.

Section 7 provides further discussion of the statistical results in this document. Another issue requiring brief mention is that by using WTADJX to measure standard errors, one loses the ability to conduct a Wald/F test when comparing prevalences across three or more groups as WTADJX does not have that capability. That is why Bonferroni-adjusted *t* tests were used when comparing prevalences across age groups in Karg et al. (in press).

2. The Scaling Factors

For the scaling of the weights across years to be most relevant for prevalence estimates, one needs to either (2) assume the underlying mental-health prevalence being estimated is constant across the years from 2008 to 2012 or (2) treat the target of estimation as the weighted mean of the annual prevalences, where the weight applied to each year is its scaling factor times its relative population size.

Mathematically, the true average prevalence from 2008 to 2012 can be expressed as

$$\bar{Y} = \frac{N_{2008}\bar{Y}_{2008} + N_{2009}\bar{Y}_{2009} + N_{2010}\bar{Y}_{2010} + N_{2011}\bar{Y}_{2011} + N_{2012}\bar{Y}_{2012}}{N_{2008} + N_{2009} + N_{2010} + N_{2011} + N_{2012}},$$

where N_t and $\bar{Y}_t$ are, respectively, the adult population size and the prevalence in year *t*. The assumption-free target of the scaled estimates is instead:

$$\bar{Y}_{scaled} = \frac{(.12)N_{2008}\bar{Y}_{2008} + (.04)N_{2009}\bar{Y}_{2009} + (.14)N_{2010}\bar{Y}_{2010} + (.35)N_{2011}\bar{Y}_{2011} + (.35)N_{2012}\bar{Y}_{2012}}{(.12)N_{2008} + (.04)N_{2009} + (.14)N_{2010} + (.35)N_{2011} + (.35)N_{2012}}. \quad (1)$$

We investigated the reasonableness of the former assumption that the 43 underlying mentalhealth prevalence estimated were constant from 2008 to 2012 by computing the 5 yearly estimates for each variable (combining the 2008A and 2008B samples) and then the standard errors of the 10 paired comparisons (e.g., the 2008 estimate for past-year explosive disorder minus the 2010 estimate) using the fully-corrected internal version of the standard-error measure.

A difference (e.g., between the 2008 and 2010 estimates of a prevalence) was decided as being statistically significant if the smallest of the 10 *p*-values per variable was less than .01. There was a less than 10% chance of this happening under the null hypothesis of an unchanging prevalence across the five years. Note that .01 is a Bonferroni adjustment applied to .1 (i.e., .01 = .1/10, with 10 being the number of paired comparisons per variable).

Three of the lowest differences were statistically significant, which is about what should be expected with 43 variables (i.e., less than 4.3). There were 430 (43 x 10) paired comparisons in all. Had we alternatively Bonferroni-adjusted the lowest *p*-value of the 430 (.00030), it would not have been significant at the .1 level.

This means the clinical data was consistent with the null hypothesis of each prevalence staying constant from 2008 to 2012. Note, however, that yearly sample sizes were small so our failure to reject the null hypothesis may have more to do with lack of power than the underlying truth of the null hypothesis.

3. Nearly Pseudo-Optimal Calibration

The MHSS clinical samples were calibrated separately in each year before the decision was made to restrict analyses to the joint 2008-2012 data set. In this section, attention is focused on the adult NSDUH sample in a single year (with the 2008A and 2008B samples treated as if they were sampled from different years).

Let *S* denote the NSDUH main adult respondent sample, w_k the weights attached to main-survey respondent *k*, and q_k the respondent's clinical-sample weight after all adjustments for coverage and nonresponse but before the final calibration to the NSDUH main sample. By convention, $q_k = 0$ when adult *k* is a respondent to the NSDUH main interview but is either not sampled for the clinical interview or did not respond if sampled for some reason.

Let $a_k = q_k/w_k$. Given a vector of calibration variables $\mathbf{z}_k$ to be defined shortly and a scalar $D = .04(\sum_S w_k)$, the final adjustment factor for clinical-interview respondent had this form:

$$f_k = \frac{\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}\right)}{1 + \left[\exp\left(\frac{U_k}{U_k - 1} a_k \mathbf{z}_k{}^T \mathbf{g}\right) - 1\right] \Big/ U_k}, \tag{2}$$

where **g** was chosen by successive linearizations (Newton's method) to satisfy the calibration equation:

$$\sum_S w_k \mathbf{z}_k = \sum_S q_k f_k \mathbf{z}_k, \tag{3}$$

and $U_k = D/q_k$ assures that no f_k is greater than U_k, which means that no final weight $\omega_k = q_k f_k$ exceeds D (*i.e.*, 4% of the total of the weights). In fact, we first trimmed a few q_k to D before applying f_k: one in 2008A, one in 2008B, one in 2010, and three in 2009. The explanation for this and other choices inherent in equation (2) are contained in the following paragraphs.

The w_k in the NSDUH main respondent sample have been calibrated so that their sum equals the adult population size. By first trimming (an asymptotically ignorable number of weights) and then restricting the final clinical weights to be no greater than T, we are assuring that no single observation dominates a prevalence estimate, which is an implicit assumption of the asymptotics underlying probability-sampling theory. It turned out that in order for a **g** to be found satisfying equations (2) and (3), U_k in (2) needed to be replaced by $1.25(T/q_k)$ for the 2008A clinical sample.

The vector $\mathbf{z}_k$ consisted of the following components, chosen to reduce the standard errors of the prevalence estimator for serious and any mental illness:

- indicators for six categories of gender (male and female) by age (18 to 25, 26 to 34, 35 or older) categories,
- indicators for four race/ethnicity categories (Hispanic, non-Hispanic white, non-Hispanic black, other),
- an indicator for past-year suicidal thoughts,
- indicators from the NSDUH main interview for a past year and lifetime major depressive episode,
- interaction terms between an alternative K6 score and the three age categories, and
- interaction terms between an alternative WHODAS score (or an alternative SDS score for the 2008B sample) and the three age categories.

See Liao *et al.* (in press, Chapter 2) for details on the alternative K6, WHODAS, and SDS scores.

The a_k in equation (2) renders the adjustment factors nearly pseudo-optimal (Kott 2011). If each $1/a_k$ were equal to the Poisson (i.e., independent across elements) probability that adult k is a respondent in the clinical sample given s/he is a respondent to the NSDUH main interview, then asymptotically optimal adjustment factors satisfying the calibration equation (3) would have the form: $f_k^{PO} = 1 + (a_k - 1)\mathbf{z}_k{}^T \mathbf{g}$. These factors can be negative and are unbounded.

A set of bounded, nonnegative adjustment factors asymptotically identical to the f_k^{PO} are

$$f_k^{NPO} = \frac{\exp\left(\frac{U_k}{U_k - 1}[a_k - 1]\mathbf{z}_k{}^T \mathbf{g}\right)}{1 + \left\{\exp\left(\frac{U_k}{U_k - 1}[a_k - 1]\mathbf{z}_k{}^T \mathbf{g}\right) - 1\right\} \Big/ U_k},$$

since $\mathbf{g} = \mathbf{O}_P(1/\sqrt{n})$ under mild conditions we assume to hold. Since all $a_k >> 1$, $f_k^{PO} \approx f_k^{NPO} \approx f_k$.

(Replacing f_k by f_k^{NPO} would reduce standard-error estimates in Table 1 by an average of less than 0.003% (log[fully-corrected standard-error measure/improved standard-error measure] × 100%) with a maximum decrease of roughly 0.01%; this measure in discussed further in Section 5.)

The adjustment factors produced by equation (2) can never be negative. As it happens, no final weight was less than 1. If it were necessary, we could have assured that all $\omega_k \geq 1$, by replacing equation (2) with

$$f_k^* = L_k + \frac{\exp\left(\frac{U_k - L_k}{(1-L_k)(U_k-1)} a_k \mathbf{z}_k{}^T \mathbf{g}\right)}{1/(1-L_k) + \left[\exp\left(\frac{U_k - L_k}{(1-L_k)(U_k-1)} a_k \mathbf{z}_k{}^T \mathbf{g}\right) - 1\right] / (U_k - L_k)},$$

where $L_k = 1/q_k$. Because, q_k is never less than 203 in the MHSS clinical sample, this is not necessary.

Another way to look at the weight-adjustment function in equation (2) is to draw a distinction between the vector of *calibration variables*, $\mathbf{z}_k$, in equation (3) and the vector of *model variables* in equation (2), $\mathbf{x}_k = a_k \mathbf{z}_k$. In this formulation f_k is a function of $\mathbf{x}_k{}^T \mathbf{g}$. The *f* needs to be subscripted by *k* so that the U_k can vary.

4. Standard Error Estimation with WTADJX

We can express a calibration-weighted total $t = \sum_S \omega_k y_k$, where ω_k is the calibration weight for adult *k*, as $t = \sum_S w_k \mathbf{z}_k{}^T \mathbf{b} + \sum_S \omega_k e_k$, where, for technical reasons explained in Kott and Liao (2012), the quasi-randomization regression coefficient is

$$\mathbf{b} = (\sum_S q_k f_k [(U_k - f_k)/(U_k - 1)] a_k \mathbf{z}_k \mathbf{z}_k{}^T)^{-1} \sum_S q_k f_k [(U_k - f_k)/(U_k - 1)] a_k \mathbf{z}_k y_k, \qquad (4)$$

while $e_k = y_k - \mathbf{z}_k{}^T \mathbf{b}$. This decomposition is effectively what WTADJX does. Each $\mathbf{x}_k = a_k \mathbf{z}_k$ in **b** can be viewed as a vector of model variables, while $\mathbf{z}_k{}^T$ in both **b** and e_k can be viewed as a (transposed) vector of calibration variables.

Since f_k is close to 1 (because **g** converges to 0 as the sample grows large), nothing would be lost asymptotically by replacing $\partial f_k / \partial(\mathbf{x}_k{}^T \mathbf{g}) = f_k [(U_k - f_k)/(U_k - 1)]$ in **b** with 1 or f_k. The interjection of $\partial f_k / \partial(\mathbf{x}_k{}^T \mathbf{g})$ into **b** is only needed when the WTADJX calibration-weighting program is used to adjust for nonresponse or coverage errors. In either of those circumstances, f_k is *not* asymptotically identical to 1 (in fact, $1/f_k$ estimates either the probability that *k* responds or the expected number of times *k* appears in the sampling frame).

For analytical purposes, the NSDUH main adult sample has a stratified multistage design with ignorably small first-stage selection probabilities and the clinical sample is Poisson. As a result, the standard error of *t* can be estimated using the "with-replacement" linearization variance estimator by noting $t = \sum_S w_k h_k$, where $h_k = \mathbf{z}_k{}^T \mathbf{b} + (\omega_k / w_k) e_k$. The standard error of an estimated mean can be computed in an analogous manner since $\sum_S w_k = \sum_S \omega_k$ by our calibration equations (the sex/age categories exhaust the population).

Getting WTADJX with ADJUST = POST to compute these standard errors takes some innovation. First, let $S^{(1)}$ denote that sample of *S* for which $q_k > 0$. Then, create the data set $S^{(2)} = S + S^{(1)}$. This new data set contains two versions of the adults originally in $S^{(1)}$, which are treated as distinct elements of $S^{(2)}$ from the same variance PSU.

For the weight variable in WTADJX (WEIGHT), use q_k for elements originally from $S^{(1)}$ and w_k for elements from *S*. WTADJX allows different model and calibration variables. For the calibration variables (CALVARS), use the components of $\mathbf{z}_k$ from the last section, multiplying each by 1 for elements from $S^{(1)}$ and by -1 for elements from *S*. As a result, only the weights for elements from in $S^{(1)}$ are adjusted while the final sum of weights can be specified as 0 for all calibration variables (POSTWGT). For the model variables (in the MODEL statement: MODEL _ONE_ = [model variables]) multiply the components of $\mathbf{z}_k$ by 0 for elements from *S* and by a_k for elements from $S^{(1)}$, so that only the elements from $S^{(1)}$ are used in computing **b**.

The y_k (which appeared in the VAR statement) for elements from *S* are set to missing. As a result, these elements are treated as if they were outside the domain of interest but still have an impact on variance estimation. In particular, in the computation of the variance of $t = \sum_S \omega_k y_k$, an element *k* from $S^{(1)}$ contributes $y_k - \mathbf{z}_k{}^T \mathbf{b}$, while an element from *S* contributes $0 - (-\mathbf{z}_k{}^T)\mathbf{b} = \mathbf{z}_k{}^T \mathbf{b}$.

Finally, the WTMAX statement is used to truncate the q_k to *T*, while the UPPERBD statement interjects the U_k into the weight adjustments.

5. The Standard-Error Measures in Table 1

As noted in the introduction, when the clinical samples were combined across years, weights were scaled using the factors: 0.06 for 2008A and 2008B, 0.04 for 2009, 0.14 for 2010, 0.35 for 2011, and 0.35 for 2012. In computing the standard errors for estimates computed with these scaled weights using WTADJX, the weights before the poststratification step (i.e., before the trimming) that appeared in the WEIGHT statement had to be scaled by the same factors.

The variance strata and variance PSUs used in WTADJX remained the same: MHVESTR and MHVEREP, which were designed for analysis of clinical data combined across years. All the variables in the MODEL and CALVARS step were crossed by a categorical indicator for year (CALV1, which ranges from 1 to 6) because the clinical samples were calibrated yearly.

The above explains how the "fully corrected-internal" column of standard-error measures in Table 1 was computed. The "not corrected" column was computed using the DESCRIPT procedure in SUDAAN 11 with the scaled final clinical sample weights in the WEIGHT statement. There is no MODEL or CALVARS statement in DESCRIPT, and only data from the clinical sample were needed for the calculations.

It is clear from the table that the fully-corrected-internal standard-error measures tend to be smaller than the not-corrected measures. To summarize the differences between the fully-corrected-internal standard-error measure and an alternative measure, the following statistic is calculated for each prevalence estimate:

$$D = \log(\text{alternative } se \text{ measure} / \text{fully-corrected-internal } se \text{ measure}) \times 100\%, \qquad (5)$$

Observe that D is very close to the percentage difference between the two *se* measures when that difference is within 10%. Unlike a standard percentage difference, however, D treats the numerator and denominator values in its internal ratio: *alternative se measure / corrected-internal se measure*

symmetrically. For example, when numerator is twice the denominator, D is roughly 69%, while when the numerator is half the denominator, D is roughly - 69%. Therefore, a positive D value indicates the alternative *se* measure tends to overestimate the *se*; while a negative value indicates the alternative *se* measure tends to underestimate the *se*.

The average D value across the 43 estimates in the table across using the not-corrected *se* measure is 13.5%. The median D *value* is 8.9%, with half the values ranging between 3.1% to 22.9%. Not all the D values were positive, however.

An operational problem with the fully-corrected-internal method is that it requires access to data from all adult respondents to the entire NSDUH main-interview. It also requires categorical year indicators, the clinical weights *before* poststratification, and the T and U_k values in equation (2).

The hope was that the standard-error measure in the last column of Table 1 could be used instead of the fully-corrected-internal one, despite the following changes in how the measure was determined:

1) The final clinical weight appeared in the WEIGHT statement.
2) Neither the model nor the calibration variables were cross classified with the categorical year indicator, for example, a single Hispanic indicator variable served as a calibration variable in place of separate yearly Hispanic indicators.
3) Only clinical sample data were used in the program.

The last meant that calibration targets were supplied by an external source. Scaled versions of the NSDUH main-survey weights were used in computing the calibration-variable targets in the POSTWGT statement.

Observe that since the computation of these alternative standard-error measures started WTADJX with the final clinical weights, the program did not change the weights at all. In addition, model variables (for the MODEL statement) were created by multiplying the calibration variables not by a_k as in Section 3 but by $a_k q_k/\omega_k$, where q_k is the clinical weight of adult k before poststratification an before trimming

values greater than T to T. The result was to produce a new quasi-randomization regression coefficient in equation (4) more similar to the fully-corrected one.

The standard-error measure described above acknowledges some of the calibration but not the separate yearly targets: It is "corrected, but not by year." Furthermore, it uses a source "external" to WTADJX for the calibration-variable targets. An alternative "internal" measure was computed for comparison purposes. It too started with the final clinical weights and used the same model variables, but the target calibration totals were computed within the WTADJX program as they were with the fully-corrected standard-error measure. This captured any additional error caused by the calibration targets themselves being estimated from a sample.

Finally, a fully-corrected-external standard error measure was computed like the corrected-but-not-by-year-external measure but with different calibration and model variables cross-classified with the categorical year indicator. In particular, it was computed using only the MHSS clinical sample, began with the final clinical-sample weight, and did not require knowledge of the T and U_k values.

Table 2 summarizes the biases of the alternative standard-error measures by measuring the differences between each and the fully-corrected-internal standard-error measure using the D statistic in equation (5). On average, correcting, but not by year, removed less than half of the bias in the not-corrected standard-error measure relative to the fully-corrected version (13.5% was reduced to 8.1% or 7.9). Using the external versions of the standard-error measures tended to be slightly higher than their internal analogues. We will explore a possible reason for this surprising result in Section 7.

Tables 1 and 2 show that the external and internal versions of a standard-error measure are usually close. The biggest absolute difference in the two fully-corrected measures was .02 (for alcohol abuse), whose D statistic was within 4.2%. The average percentage difference computed using the D statistic was 0.4%. SAMHSA only publishes estimated mental-health prevalences and their standard errors to one decimal place.

6. Using Different Stratum Identifiers

As noted in the introduction, the standard-error measures in Table 1 were computed using 100 variance strata. These are identified on MHSS data sets by MHVESTR, with the two variance PSUs within each identified by MHVEREP. An alternative set of 50 variance strata, identified by MHVSTR09, were used for the methodological work described in Liao *et al.* (in press). Table 3 displays the impact of using the alternative stratum identifiers on the not-corrected standard-error measures and the fully-corrected-internal standard-error measures.

Using MHVESTR09 in place of MHVESTR resulted in an average decrease in the not-corrected standard-error measure of 0.5% computed using the D statistic and in the fully-corrected-internal measure of 1.5%, both small amounts. Asymptotic theory suggests that using this stratum identifier might have the *opposite* effect because some of the variance-decreasing impact of stratification could be lost due to the additional collapsing of strata. We explore this anomalous result in the following section.

7. Discussion

Since the standard-error measures labeled "internal" were designed to capture the added variability of the calibration targets based on estimates from the NSDUH main survey, it is a bit disconcerting that the results in Table 2 suggests (as was the case) that they were on average smaller than otherwise analogous measures that treated those targets as fixed (the "external" measures). To explore what may have happened, observe that

$$p = \left(\frac{\sum_S \omega_k y_k}{\sum_S \omega_k} - \frac{\sum_S w_k \mathbf{z}_k{}^T \mathbf{b}^*}{\sum_S w_k} \right) + \frac{\sum_S w_k \mathbf{z}_k{}^T \mathbf{b}^*}{\sum_S w_k}$$

implies $Var(p) = Var(p - p_1) + Var(p_1) + 2Cov(p - p_1, p_1)$, where $p_1 = \sum_S w_k \mathbf{z}_k{}^T \mathbf{b}^* / \sum_S w_k$,

and **b*** is the probability limit of **b** as the clinical sample size grows arbitrarily large, which we assumes exists (**b***, unlike **b**, is not a random variable). Because the NSDUH main sample is itself calibrated, $\sum_S \omega_k = \sum_S w_k$.

In the above formulation, $p = \sum_S \omega_k y_k / \sum_S \omega_k$ is a prevalence estimate based on the clinical sample, and *Var*(*p*) its variance. The expression *Var*(*p*) is the full variance estimator used in the fully-corrected internal standard-error measure, $Var(p - p_1)$ is asymptotically the variance of *p* treating the calibration targets as fixed, which is used in the fully-corrected external standard-error measure, and $Var(p_1)$ is the direct contribution to the variance of *p* from the calibration targets themselves being based on a random sample. What has happened is that $2Cov(p - p_1, p_1)$ tended to be negative and dominate $Var(p_1)$.

The negativity of $Cov(p - p_1, p_1)$ was largely a happy byproduct of inserting the a_k in equation (2), which was done to increase statistical efficiency. Recall that a_k is a product of all the weighting factors applied to the NSDUH main sample analysis weight w_k before poststratification . Removing the a_k before computing the fully-corrected internal version of the standard-error measures caused, on average, a 2.7% increase in the estimated coefficient of variation (CV). The estimated means also changed slightly, hence the use of estimated CV here (and the *D* statistic). As can be seen in Table 4, not all CV measures decreased from inserting the a_k. In fact, over a quarter decreased, and the median decrease was only 0.5%, which is still positive

Removing the a_k from equation (2) but otherwise mimicking the production of the fully-corrected-internal version of the standard-error measures tends to make $Cov(p - p_1, p_1)$ disappear. When external targets replaced internally-computed ones with the a_k in equation (2) removed from both, the standard-error measure decreased 0.4% on average. The median decrease was of roughly the same size.

It is important to appreciate that most of the variance of *p* comes from $Var(p - p_1)$ whatever the sign of $Cov(p - p_1, p_1)$. One should also keep in mind that the square of the fully-corrected internal version of the standard-error measure is only an *asymptotically* unbiased estimator for *Var*(*p*) under a host of assumptions. In addition, this variance estimator itself has a variance.

It is surprising that the standard-error measures computed with MHVSTR09 were slightly larger, on average, that the measures computed with MHVESTR09 when asymptotic theory suggests they should have been, if anything, slightly smaller. One needs to remember that the clinical samples were finite while asymptotic theory treats them as arbitrarily large. Moreover, there is a lot of noise in the variances estimators computed either way.

One of the assumptions implicitly made by the fully-corrected internal version of the standard-error measure was that the calibration targets themselves were pure probability estimates based on the NSDUH main sample. Some of these targets, like the populations of the six age/gender categories, were in fact, provided by the Census Bureau, while the others benefited from the calibration weighting in the NSDUH main sample. With that in mind, it is comforting to note that whether or not the calibration targets were treated as fixed had only a very modest impact on the standard-error measures.

Another assumption was that the weighting adjustment for the clinical sample before poststratification not only removed selection biases due to undercoverage and nonresponse, but actually estimated the probabilities of a Hispanic responding to the main NSDUH survey in English and an adult participating to the clinical interview exactly. In fact, the presence of these additional calibration-weighting adjustments would tend to bias the fully-corrected -internal standard-error measure *upward* (i.e., make them overestimate the true standard error). A detailed argument for this can be found in Kott and Day (in press). Briefly, if the residuals $e_k = y_k - \mathbf{z}_k{}^T\mathbf{b}$ were correlated with the covariates used in the coverage adjustment or one of the two nonresponse adjustments, then using calibration-weighting techniques in those adjustments would incorporate more information about the under-covered and/or nonrespondents in the prevalence estimation – and thus decrease the variance of the resulting prevalence estimator – than would be reflected in the standard-error measure.

Returning to the results displayed Tables 1 and 2 and keeping in mind that standard-error measures are themselves estimates subject to both bias and variance, there appears little argument against using the external version of the fully-corrected standard-error measure. The corrected-but-not-by-year measures, in contrast, appear not to capture adequately the reduction in standard error due to the poststratification of the clinical sample.

References

Center for Behavioral Health Statistics and Quality (in press). *National Survey on Drug Use and Health: 2008-2012 Adult Clinical Interview Data File and Codebook*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

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, NY: Biometrics Research, New York State Psychiatric Institute.

Karg, R.S., Bose, J., Batts, K.R., Forman-Hoffman, V.L., Liao, D., Hirsch, E., Pemberton, M.R., Colpe, L.J., and Hedden, S.L. (in press). *CBHSQ Data Review: Past Year Mental Disorders among Adults in the United States: Results from the 2008–2012 Mental Health Surveillance Study*. Rockville, MD: Substance Abuse and Mental Health Services Administration.

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, *Archives of General Psychiatry, 60*, 184-189.

Kott, P.S. (2011). A Nearly Pseudo-optimal Method for Keeping Calibration Weights from Falling Below Unity in the Absence of Nonresponse or Frame Errors, *Pakistan Journal of Statistics,* 27, 391-396.

Kott, P.S. and Day, C. (in press). Developing Calibration Weights and Standard-Error Estimates for a Survey of Drug-Related Emergency-Department Visits, Journal of Official Statistics, forthcoming.

Leon, A. C., Olfson, M., Portera, L., Farber, L., & Sheehan, D. V. (1997). Assessing Psychiatric Impairment in Primary care with the Sheehan Disability Scale, *International Journal of Psychiatry in Medicine, 27*(2), 93-105.

Liao, D., Kott, P., Aldworth, J., Yu, F., Morton, K., Chen, P. Shook-Sa, B., Kott, P., & Davis, T. (in press). *2012 Mental Health Surveillance Study: Design and estimation report* (prepared for the Substance Abuse and Mental Health Services Administration, Contract No. HHSS283201000003C, RTI 0212800.001.112.008.006). Research Triangle Park, NC: RTI International.

Rehm, J., Üstün, 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, *International Journal of Methods in Psychiatric Research, 8*, 110-123.

RTI International (2012). *SUDAAN Language Manual, Release 11.0.* Research Triangle Park, NC: RTI International.

Substance Abuse and Mental Health Services Administration (2013). *Results from the 2012 National Survey on Drug Use and Health: Mental Health Findings*. NSDUH Series H-47, HHS Publication No. (SMA) 13-4805. Rockville, MD: Substance Abuse and Mental Health Services Administration.

Table 1. Alternative Standard Error Measures for Mental-Health Prevalence Estimates: NSDUH Adult Clinical Interview Data File, 2008-2012

Variable	Prevalence Estimate	Standard Error Measures				
		Not Corrected	Fully Corrected		Corrected, but not by Year	
			Internal	External	Internal	External
Lifetime MDD Disorder	19.79	0.95	0.82	0.82	0.87	0.88
Lifetime MDE Disorder	20.68	0.97	0.84	0.84	0.87	0.89
Lifetime Manic Disorder	0.71	0.16	0.15	0.15	0.16	0.16
Lifetime Bipolar Disorder	0.69	0.16	0.15	0.15	0.16	0.16
Lifetime MDE or Manic Disorder	20.78	0.97	0.84	0.84	0.88	0.89
Past-Year MDD Disorder	5.99	0.43	0.34	0.34	0.36	0.36
Past-Year MDE Disorder	6.34	0.44	0.36	0.35	0.38	0.37
Past-Year Manic Disorder	0.31	0.10	0.09	0.09	0.09	0.09
Past-Year Dysthymic Disorder	1.70	0.27	0.25	0.24	0.26	0.26
Past-Year Bipolar Disorder	0.39	0.10	0.09	0.09	0.10	0.10
Past-Year Any Mood Disorder	7.40	0.52	0.39	0.39	0.42	0.42
Psych Screen	0.58	0.15	0.14	0.14	0.15	0.15
Posttraumatic Disorder	0.74	0.10	0.10	0.10	0.10	0.09
Panic Disorder without Agoraphobia	0.89	0.14	0.13	0.13	0.13	0.13
Agoraphobia without History of Panic Disorder	0.21	0.08	0.07	0.07	0.08	0.08
Social Phobia	0.96	0.20	0.15	0.16	0.21	0.21
Specific Phobia	1.61	0.81	0.44	0.44	0.71	0.71
Obsessive-Compulsive Disorder	0.29	0.07	0.06	0.06	0.06	0.06
Generalized Anxiety Disorder	1.79	0.22	0.19	0.20	0.21	0.21
Any Anxiety Disorder	5.65	0.89	0.55	0.54	0.80	0.80
Explosive Disorder	0.39	0.09	0.08	0.09	0.09	0.09
Serious Mental Illness*	3.94	0.29	0.25	0.25	0.26	0.26
Any Mental Illness*	17.95	0.97	0.75	0.75	0.87	0.87

Table 1. Alternative Standard Error Measures for Mental-Health Prevalence Estimates: NSDUH Adult Clinical Interview Data File, 2008-2012 (cont.)

Variable	Prevalence Estimate	Standard Error Measures				
		Not Corrected	Fully Corrected		Corrected, but not by Year	
			Internal	External	Internal	External
Alcohol Abuse	3.07	0.45	0.48	0.50	0.44	0.44
Alcohol Dependence	3.28	0.39	0.38	0.38	0.41	0.41
Alcohol Dependence or Abuse	6.36	0.66	0.65	0.65	0.64	0.64
Illicit Drug Abuse	0.92	0.18	0.17	0.17	0.18	0.18
Illicit Drug Dependence	2.06	0.39	0.44	0.45	0.39	0.39
Illicit Drug Dependence or Abuse	2.98	0.43	0.47	0.48	0.44	0.44
Any Substance Use Disorder (SUD)	7.77	0.70	0.71	0.71	0.68	0.68
Adjustment Disorder	6.89	0.50	0.48	0.49	0.50	0.50
Any Disorders (excluding SUD or Adjustment)	11.49	0.95	0.66	0.65	0.84	0.84
1 Disorder	8.01	0.92	0.60	0.60	0.83	0.83
2 Disorders	1.76	0.19	0.17	0.17	0.18	0.18
3+ Disorder	0.87	0.12	0.11	0.11	0.11	0.11
Any Disorders (excluding SUD)	17.11	1.11	0.85	0.86	0.98	0.98
1 Disorder	10.97	0.97	0.68	0.70	0.88	0.89
2 Disorders	3.18	0.42	0.45	0.46	0.42	0.42
3+ Disorders	1.91	0.28	0.25	0.22	0.27	0.27
Any Disorders	22.52	1.16	0.92	0.93	1.04	1.04
1 Disorder	14.90	0.98	0.75	0.75	0.91	0.92
2 Disorders	4.10	0.47	0.48	0.50	0.45	0.45
3+ Disorders	2.19	0.29	0.27	0.26	0.28	0.28

* The prevalence estimates here are not based on a model like those in Substance Abuse and Mental Health Services Administration (2013). Moreover these estimates scale the contributions from the component years (2008 through 2012) for statistical efficiency, which was neither necessary nor appropriate for the model-based estimates.

Any mood disorder is defined as having major depressive disorder, bipolar disorder (type I only), or dysthymic disorder in the past year.

Substance abuse and dependence are mutually exclusive. If a respondent is classified as having substance dependence (alcohol or illicit drugs), then he cannot be classified as abusing that substance regardless of responses to the abuse criteria questions.

Any disorder is defined as having one of the measured mood disorders, anxiety disorders, substance use disorders (included or excluded as specified in the header), eating disorders, adjustment disorder (included or excluded as specified in the header), or intermittent explosive disorder. A respondent can be classified as having any disorder even if the number of disorders is not able to be determined.

Combined variables are set to "Yes" if one or more source variable is "Yes," to "No" if all of the source variables are "No," and "missing" otherwise. Cases with missing values in the MHSS variables are excluded from the analyses.

There is a summary of the standard-error measures on the following page.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, NSDUH main study and clinical sample, 2008-2012.

Table 2. Summarizing Differences Between Alternative Standard-Error Measure and the Fully-Corrected-Internal Method (in Percent, using the *D* Statistic*)

Method	Mean	Median	First Quartile	Last Quartile	Minimum	Maximum
Not Corrected	13.5	8.9	3.1	22.9	-14.0	61.2
Fully Corrected-External	0.4	0.1	-0.5	1.2	-1.8	4.2
Corrected, but not by Year						
External	7.9	5.1	1.4	11.3	-14.0	48.5
Internal	8.0	5.3	1.9	11.3	-14.2	48.5

* D = log(alternative se measure / fully-corrected-internal se measure) ×100%,

See Table 1 for other definitions.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, NSDUH main study and clinical sample, 2008-2012.

Table 4. Summarizing the Impact of Removing the a_k from Equation (2) on the Fully-Corrected Estimator (in Percent, Applying the *D* Statistic to CVs*)

Comparisons	Mean	Median	First Quartile	Last Quartile	Minimum	Maximum
Fully Corrected-Internal Without vs. With the a_k	2.7	0.5	-3.1	5.2	-17.6	32.3
Fully Corrected-External vs. Fully Corrected-Internal Both Without the a_k	-0.4	-0.4	-0.8	0.1	-.2.8	0.9

* D = log(CV using alternative se measure / CV using fully-corrected-internal se measure) ×100%.

See Table 3 for other definitions.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, NSDUH main study and clinical sample, 2008-2012.

Summary of Standard-Error Measures

Not corrected treats the final calibrated weights as if there were design weights.

Fully corrected-external treats the annual calibration targets as if they were provided externally.

Fully corrected-internal computes the annual calibration targets internally and treats them as random.

Corrected, but not by year-external recalibrates the final weights as if calibration was combined across years using external targets.

Corrected, but not by year-internal calibrates the final weights as if the calibration was combined across years using internal targets.

Table 3. Standard Error Measures for Mental-Health Prevalence Estimates Using Different Stratum Identifiers

Variable	Prevalence Estimate	Standard Error Measures			
		Not Corrected		Fully Corrected-Internal	
		MHVESTR	MHVSTR09	MHVESTR	MHVSTR09
Lifetime MDD Disorder	19.79	0.95	0.93	0.82	0.71
Lifetime MDE Disorder	20.68	0.97	0.94	0.84	0.71
Lifetime Manic Disorder	0.71	0.16	0.15	0.15	0.15
Lifetime Bipolar Disorder	0.69	0.16	0.16	0.15	0.15
Lifetime MDE or Manic Disorder	20.78	0.97	0.94	0.84	0.72
Past-Year MDD Disorder	5.99	0.43	0.44	0.34	0.35
Past-Year MDE Disorder	6.34	0.44	0.45	0.36	0.36
Past-Year Manic Disorder	0.31	0.10	0.10	0.09	0.10
Past-Year Dysthymic Disorder	1.70	0.27	0.25	0.25	0.24
Past-Year Bipolar Disorder	0.39	0.10	0.10	0.09	0.10
Past-Year Any Mood Disorder	7.40	0.52	0.52	0.39	0.39
Psych Screen	0.58	0.15	0.15	0.14	0.15
Posttraumatic Disorder	0.74	0.10	0.10	0.10	0.10
Panic Disorder without Agoraphobia	0,89	0.15	0.14	0.15	0.13
Agoraphobia without History of Panic Disorder	0.21	0.08	0.07	0.07	0.06
Social Phobia	0.96	0.20	0.17	0.15	0.12
Specific Phobia	1.61	0.81	0.81	0.44	0.43
Obsessive-Compulsive Disorder	0.29	0.07	0.05	0.06	0.05
Generalized Anxiety Disorder	1.79	0.22	0.20	0.19	0.19
Any Anxiety Disorder	5.75	0.89	0.87	0.55	0.56
Explosive Disorder	0.39	0.09	0.08	0.08	0.08
Serious Mental Illness*	3.94	0.29	0.27	0.25	0.24
Any Mental Illness*	17.95	0.97	0.93	0.75	0.66

Table 3. Standard Error Measures for Mental-Health Prevalence Estimates Using Different Stratum Identifiers (cont.)

Variable	Prevalence Estimate	Standard Error Measures			
		Not Corrected		Fully Corrected-Internal	
		MHVESTR	MHVSTR09	MHVESTR	MHVSTR09
Alcohol Abuse	3.07	0.45	0.45	0.48	0.50
Alcohol Dependence	3.28	0.39	0.40	0.38	0.40
Alcohol Dependence or Abuse	6.36	0.66	0.65	0.65	0.66
Illicit Drug Abuse	0.92	0.18	0.21	0.17	0.20
Illicit Drug Dependence	2.06	0.39	0.42	0.44	0.49
Illicit Drug Dependence or Abuse	2.98	0.43	0.47	0.47	0.51
Any Substance Use Disorder (SUD)	7.77	0.70	0.69	0.71	0.71
Adjustment Disorder	6.89	0.50	0.48	0.48	0.46
Any Disorders (excluding SUD or Adjustment)	11.49	0.95	0.95	0.66	0.62
1 Disorder	8.01	0.92	0.96	0.60	0.58
2 Disorders	1.76	0.19	0.18	0.17	0.16
3+ Disorder	0.87	0.12	0.11	0.11	0.11
Any Disorders (excluding SUD)	17.11	1.11	1.22	0.85	0.87
1 Disorder	10.97	0.97	1.11	0.68	0.77
2 Disorders	3.18	0.42	0.38	0.45	0.42
3+ Disorders	1.91	0.28	0.31	0.25	0.28
Any Disorders	22.52	1.16	1.25	0.92	0.91
1 Disorder	14.90	0.98	1.07	0.75	0.79
2 Disorders	4.10	0.47	0.46	0.48	0.48
3+ Disorders	2.19	0.29	0.32	0.27	0.30

* The prevalence estimates here are not based on a model like those in Substance Abuse and Mental Health Services Administration (2013). Moreover these estimates scale the contributions from the component years (2008 through 2012) for statistical efficiency, which was neither necessary nor appropriate for the model-based estimates.

Any mood disorder is defined as having major depressive disorder, bipolar disorder (type I only), or dysthymic disorder in the past year.

Substance abuse and dependence are mutually exclusive. If a respondent is classified as having substance dependence (alcohol or illicit drugs), then he cannot be classified as abusing that substance regardless of responses to the abuse criteria questions.

Any disorder is defined as having one of the measured mood disorders, anxiety disorders, substance use disorders (included or excluded as specified in the header), eating disorders, adjustment disorder (included or excluded as specified in the header), or intermittent explosive disorder. A respondent can be classified as having any disorder even if the number of disorders is not able to be determined.

Combined variables are set to "Yes" if one or more source variable is "Yes," to "No" if all of the source variables are "No," and "missing" otherwise. Cases with missing values in the MHSS variables are excluded from the analyses.

Source: SAMHSA, Center for Behavioral Health Statistics and Quality, NSDUH main study and clinical sample, 2008-2012.