Assessment of Weighting Methodology for the National Comorbidity Survey

R. J. A. Little, S. Lewitzky, S. Heeringa, J. Lepkowski, and R. C. Kessler

The authors studied weighting adjustments for the National Comorbidity Survey (1990–1992), a large-scale national epidemiologic investigation of the prevalence, risk factors, and consequences of psychiatric morbidity and comorbidity in the United States. Weighting adjustments for differential selection within households, new construction, unit nonresponse, and poststratification were examined separately and in combination. Specific issues addressed included the magnitude of the bias incurred from ignoring the weights, the added variance from weighting and how well this was predicted by simple formulae, and the performance of methods for trimming the weights. The weights after trimming and poststratification appeared to work well. It is suggested that the added variance from weighting be carefully monitored in similar surveys. Alternatives to the use of trimming for controlling variance are worth exploring.

In estimating prevalences from epidemiologic survey data, people with the characteristic of interest may be more or less likely to participate in the survey than people without it. The sample proportion is then a biased estimator of the true proportion. The standard statistical approach to adjustment for this potential bias is to estimate the probability of each respondent's participating in the survey and to assign each observation a weight inversely proportional to that probability (1). This form of weighting is quite different from weighting in inverse proportion to variance in regression analysis, where the objective is variance reduction rather than reduction of bias from differential inclusion probabilities.

Although the use of survey weights for analytic inferences such as regression is still debated (1–6), most statisticians advocate the use of weights for estimates of prevalence and other descriptive means and totals. A common disadvantage of weighting is an increase in the variance of the estimator. Under simplified independent sampling and constant variance assumptions, the added variance can be expressed as

\[ V_w = V_u (1 + s_w^2) \]

(see Kish (1) and Potter (7), for example), where \( V_u \) is the variance of the unweighted prevalence estimate, \( V_w \) is the variance of the weighted prevalence, and \( s_w^2 \) is the sample variance of the weights, assumed to be scaled to average unity. This formula does not apply exactly to stratified multistage surveys involving poststratification but has proven to be a useful rule of thumb in some surveys (for example, see Verma et al. (8)). Since weighting is questionable when added variance dominates bias reduction (1), modifications of the weights to reduce variance, such as trimming (7), collapsing weight strata (9, 10), or shrinking (6, 11), are common in practice.

Does weighting improve inferences in practice? This depends on the nature of the weighting adjustments in particular surveys and how they relate to the survey outcomes. Korn and Graubard (12, 13) studied the effects of weighting in large health surveys. Here, we study weighting adjustments for the National Comorbidity Survey (NCS), a major epidemiologic investigation of the prevalence, risk factors, and consequences of psychiatric morbidity and comorbidity in the United States (14). Specific questions addressed include: How much bias is incurred from ignoring the weights? How much variance is added from weighting, and is it justified by reduced bias? How well is the added variance predicted using equation 1? How suc-
cessful are weight trimming methods in reducing the increase in variance while at the same time adjusting for bias? Weight components are considered individually as well as in combination in addressing these questions. Implications of our answers for other mental health surveys are discussed.

The scope of this paper is limited to estimators of the prevalences of psychiatric disorders, nationally and in one population subgroup. We describe four "participation factors" that affect the probability that a given subject is sampled and participates in the NCS. We describe how these probabilities and the corresponding weights are estimated, poststratified, and trimmed. We then examine the effect of the weights, individually and in combination, on the bias, variance, and mean squared error of our estimates.

MATERIALS AND METHODS

The National Comorbidity Survey

The NCS was a general population survey carried out in 1990–1992 to examine the prevalences, risk factors, and consequences associated with various psychiatric disorders (14). It was based on a stratified, multistage area probability sample of persons aged 15–54 years in the noninstitutionalized civilian population of the 48 coterminous US states. A total of 8,098 respondents participated in the survey, resulting in a response rate of 82.4 percent. Our analysis focused on estimates of prevalence of the 14 main psychiatric disorders diagnosed for the full sample: alcohol abuse or dependency, depression, social phobia, conduct disorder, drug abuse, simple phobia, agoraphobia, dysthymia, antisocial behavior, anxiety disorder, panic disorder, antisocial personality, mania, and hypomania. Each disorder was treated as a dichotomous variable. Diagnoses were based on the diagnostic system of the Diagnostic and Statistical Manual of Mental Disorders, Third Edition, Revised (DSM-III-R) (15), generated from a modified version of the Composite International Diagnostic Interview (16), a structured interview designed to be used by trained interviewers who are not clinicians.

Weighting methodology for the NCS

The NCS was designed to yield an equal probability sample of households. However, the probability of participation varied because of four main factors: a within-household factor, a hold-back factor, a new construction factor, and a unit nonresponse factor. For each respondent, weights were estimated for each of these factors and were multiplied together to obtain the overall weight, which was inversely proportional to the estimated probability of participation.

As is common in surveys of this kind, the weights in this process were modified in an attempt to improve their statistical properties. The highest values of individual and final weights were trimmed back to reduce variance. The trimmed final weight was poststratified to adjust for differences between our sample and national data available from a larger survey. Finally, the resulting poststratified weights were again trimmed. Our goal was to examine the impact of this complex weighting process on prevalence estimates.

Within-household weight. Interviewers were instructed to select at random one person aged 15–54 years from each selected household for interview, and only these target respondents were interviewed. Thus, individuals in large households had a smaller chance of selection. Specifically, if a target respondent was successfully interviewed in a household containing k people between the ages of 15 and 54 years, the respondent had a 1/k probability of being selected as the target from that household and therefore received a within-household weight proportional to k. Two adjustments were made to this basic procedure. First, people aged 15–24 years tend to live in large households; to increase the sample size for this group, each household member aged 15–24 was double-listed prior to selection for a random one third of all targeted households. The within-household weight for these respondents was adjusted to account for this double listing. Second, less than 0.3 percent of the 8,098 respondents had within-household weights greater than 6. These respondents had their weights trimmed to a value of 6. The final distribution of the within-household weights is shown in figure 1. Scaled to average unity, weights ranged from 0.52 to 5.87, with a standard deviation of 0.48.

Hold-back weight. The survey was conducted in six phases, or rotations, each in itself a nationally representative random sample. For completion of as much of the final rotation as possible, near the close of fieldwork targeted individuals in the final rotation who had not yet been successfully interviewed were randomly divided into two halves. One half was "held back," and interviewers focused their efforts on the other half. Each interviewed subject in the half not held back was then assigned a hold-back weight proportional to 2, whereas all other respondents received a hold-back weight proportional to 1. Only 3.3 percent of respondents received the larger of the two hold-back weights, so the impact of the hold-back weight on variance was expected to be minor. Scaled to average unity, the weights ranged from 0.97 to 1.94, with a standard deviation of 0.17.
Weighting for the National Comorbidity Survey

New construction weight. As in many multistage household surveys, the primary sampling units and secondary sampling units were selected with probability proportional to 1980 US Census household totals. The number of sample households from a secondary sampling unit thus varied according to the ratio of the current number of households to the 1980 Census measure of size. Excessive interviewer workloads resulted in 34 secondary sampling units, mainly because of large-scale housing construction undertaken since the 1980 Census. To reduce interviewer burden in these secondary sampling units, households were randomly subsampled with rates from one half to one fifth before the sample was released to the field. Respondents then received new construction weights ranging initially from 2 to 5, compared with 1 for other respondents. Less than 1 percent of the 8,098 observations with new construction weights greater than 3 were later trimmed to a value of 3. Since less than 4 percent of the sample had new construction weights more than 1, the impact of this weighting adjustment was expected to be minor. Scaled to average unity, weights ranged from 0.63 to 2.85, with a standard deviation of 0.27.

Nonresponse weight. Previous research has shown that nonrespondents in mental health surveys have higher rates of psychiatric disorders than respondents (17, 18). To assess this issue in the NCS, a nonrespondent survey was carried out, consisting of a brief (20-minute) screening interview administered to a probability sample of nonrespondents. The interview included screening questions selected from the larger interview for the main DSM-III-R disorders, as well as questions on basic demographic data. Of the 831 nonrespondents selected, 353 responded, 128 could not be contacted (mainly because they had moved and could not be traced), and 350 refused to participate. The relatively high interview rate (50 percent) among the contacted participants was achieved by the low burden coupled with a substantial financial incentive ($50).

Participants in the nonrespondent study did, in fact, report higher prevalences of most DSM-III-R disorders than respondents (19). The bias was adjusted by means of a two-step weighting process. First, the 353 participants in the nonrespondent survey were weighted using the method of propensity score adjustment (20) to match the distribution of all NCS nonrespondents on 1) household characteristics recorded at first attempted contact (for example, whether there were “No Trespassing” signs posted) and 2) block-level characteristics from the 1990 Census (for example, percentage of block residents living in poverty and percentage living alone). Second, the respondents were weighted using a second propensity-score adjustment to match the combined distribution of respondents and nonrespondents on variables appearing in both the full and nonrespondent interviews. The final distribution of nonresponse weights is shown in figure 2. Scaled to average unity, they ranged from 0.81 to 82.27, with a standard deviation of 1.04.

Poststratification and posttrimming. The weighted distribution of NCS participants across demographic groups was compared with the distribution estimated

---

**FIGURE 1.** Frequency distribution of within-household weights in the National Comorbidity Survey, 1990-1992. Numbers over bars are the actual histogram frequencies.

**FIGURE 2.** Frequency distribution of nonresponse weights in the National Comorbidity Survey, 1990-1992. Numbers over bars are the actual histogram frequencies.
from the 1989 National Health Interview Survey (NHIS) (21). The NHIS measured this distribution more precisely, since it was based on a much larger sample with a very high response rate. Poststratification was applied to adjust the NCS distribution across demographic groups to match the NHIS distribution. To the extent that the prevalence of a given psychiatric disorder is correlated with these demographic variables, poststratification improves the NCS estimate of the national prevalence of this disorder (10, 22). A demographic variable was included in the poststratification if 1) reliable NHIS data were available, 2) its data distribution in the NHIS differed from the distribution in the NCS, and 3) the variable was significantly related to the DSM-III-R disorders. The number of poststrata was limited to eight to avoid excess variance that arose when the NCS respondents were too sparsely distributed across the categories. The final poststrata were:

1. High school education or less, lives with others.
2. High school education or less, lives alone.
3. More than high school education, lives alone, lives in a metropolitan area with >250,000 people.
4. More than high school education, lives with others, lives in a metropolitan area with >250,000 people, not in Midwest, not Hispanic.
5. More than high school education, lives with others, lives in a metropolitan area with >250,000 people, not in Midwest, Hispanic.
6. More than high school education, lives with others, lives in an urban Metropolitan Statistical Area with >250,000 people, in Midwest.
7. More than high school education, lives with others, lives in an urban non-Metropolitan Statistical Area with >250,000 people, in Midwest.
8. More than high school education, lives in a metropolitan area with <250,000 people.

For any given demographic category, the poststratification weight is proportional to the ratio of the proportion of people in that category estimated by the NHIS to the proportion in that category estimated by the NCS, where the latter includes the four weighting adjustments discussed above. In order to minimize seasonality effects, poststratification was applied separately to each of the 12 monthly samples in a year. The resulting poststratification weights were then multiplied by the product of the four weights described above to yield poststratified weights. A final trimming reduced all weights falling into the highest 1.5th percentile to the 98.5th percentile. The distribution of the final weights after poststratification and trimming is shown in figure 3.

They ranged from 0.10 to 5.67, with a standard deviation of 0.97.

**Variance estimation**

The variance of estimates was estimated using the jackknife repeated replications method, which accounts for the complex sample design (2, 23). The replications were $2H = 84$ sampling error computation units, consisting of 46 primary sampling units sampled with a probability less than 1 and 38 random groupings of secondary sampling units within primary sampling units selected with a probability of 1. These 84 sampling error computation units were classified into $H = 42$ relatively homogeneous randomly ordered pairs according to geography and urbanicity. For example, one replicate contained two urban northeastern primary sampling units, another contained two farm belt rural primary sampling units, and so on.

Let $\hat{Q}_w$ be the estimate of a prevalence $\Theta$ for some weighting scheme $w$. For $j = 1, \ldots, H$, let $\hat{Q}_v(j)$ be the estimate computed like $\hat{Q}_w$ (including poststratification if $\hat{Q}_w$ is poststratified), but with the weight for the first of the two randomly ordered primary sampling units in the $j$th replicate doubled and the weight of the second primary sampling unit in the $j$th replicate set to zero. The estimated variance of $\hat{Q}_w$ is then

$$\hat{\sigma}_w^2 = \sum_{j=1}^{H} (\hat{Q}_w(j) - \hat{Q}_w)^2.$$
Methodology for assessment of weighting adjustments

Variance comparisons. For a prevalence \( \Theta \) of the population, let \( \hat{\Theta}_u \) be an unweighted estimate from the sample and let \( \hat{\Theta}_w \) be an estimate for some weighting scheme. Let \( V_u \) and \( V_w \) denote the variances of \( \hat{\Theta}_u \) and \( \hat{\Theta}_w \), and the bias may depend on the choice of estimator \( \hat{\Theta}_w \). Results with regard to bias and mean squared error seem justified, although it destroys exact unbiasedness. Results with regard to bias and mean squared error, it is necessary to estimate bias and mean squared error, it is necessary to estimate MSE from the sample and let \( \hat{\Theta}_u \) be an estimate for some weighting scheme. Let \( \hat{\Theta}_u \) be an estimate for some weighting scheme—the unweighted estimator, for example—with unknown bias \( B_1 \). Let \( V_0 \) and \( V_1 \) denote the variances of \( \hat{\Theta}_0 \) and \( \hat{\Theta}_1 \), respectively, and let MSE_0 and MSE_1 denote the mean squared errors; thus, by definition, MSE_1 = \( B_1^2 + V_1 \), and by assumption MSE_0 = \( V_0 \). In addition, \( E(\hat{\Theta}_1 - \hat{\Theta}_0) = B_1 \) and \( E((\hat{\Theta}_1 - \hat{\Theta}_0)^2) = B_1^2 + V_{01} \), where \( V_{01} \) is the variance of \( \hat{\Theta}_1 - \hat{\Theta}_0 \). Let \( V_0 \), \( \hat{V}_1 \), and \( \hat{V}_{01} \) be jackknife repeated replications estimates of \( V_0 \), \( V_1 \), and \( V_{01} \). Then, \( \hat{B}_1 = \hat{\Theta}_1 - \hat{\Theta}_0 \) is a consistent estimator of the bias of \( \hat{\Theta}_1 \), MSE_0 = \( \hat{V}_0 \), and

\[
\text{MSE}_1 = \hat{V}_1 + \max\{\hat{B}_1^2 - \hat{V}_{01}, 0\} \tag{2}
\]

is a consistent estimate of the mean squared error of \( \hat{\Theta}_1 \). Equation 2 is a refinement of the mean squared error estimates presented by Kish (1) in that the subtraction of \( \hat{V}_{01} \) in the last term corrects for upward bias of \( \hat{B}_1^2 \) as an estimate of the squared bias. The last term in equation 2 is constrained to be nonnegative, since it is estimating a nonnegative quantity. This adjustment seems justified, although it destroys exact unbiasedness. Results with regard to bias and mean squared error should be interpreted cautiously, since they depend on the choice of estimator \( \hat{\Theta}_0 \), and the bias may not be very precisely determined.

RESULTS

Combined weighting adjustments

Table 1 summarizes estimated overall prevalences of the 14 disorders for four weighting options, namely 1) unweighted, 2) weighted without poststratification or trimming, 3) poststratified, weighted with no trim-
ming, and 4) the final weight, with poststratification and trimming. The disorders are listed in the order of decreasing prevalence, as estimated by option 4 above. Estimated standard errors are shown, together with predicted standard errors for the weighted methods based on equation 1. Root mean squared errors are also tabulated according to the methods presented above, with method 3 assumed to be unbiased. If this assumption is incorrect, root mean squared error results are biased in favor of method 3.

Other things being equal, variance increases as a fraction of mean squared error as the sample size is reduced. Since weighting tends to reduce bias and increase variance, one would expect it to be more useful for estimates based on the entire sample than for estimates based on a subdomain. The impact of weighting on subdomain estimates was examined by repeating the analysis of table 1 for subjects aged 15–24 years, a restriction that reduced the sample size from 8,098 to 1,769. Results are shown in table 2. Below, we discuss the results presented in tables 1 and 2 with reference to our original research questions.

How much bias is incurred from ignoring the weights? Weighting has a modest effect on the estimated prevalences, from a substantive viewpoint. Unweighted estimates are generally within 10 percent of the final weighted estimates, and differences are minor compared with other potential effects relating to how a disorder is measured. For example, estimates of depression can vary greatly depending on how the condition is operationalized (24). Nevertheless, weighting adjustments for some disorders—for example, depression and dysthymia—are quite large when compared with the standard errors of the estimates, so weighting may still be useful for reducing bias and mean squared error. The differences in weighted and unweighted estimates are significantly different from zero for depression and dysthymia ($P = 0.005$ and $P = 0.018$, respectively).

How much added variance is incurred from weighting? Weighting without poststratification or trimming results in marked increases in estimated standard errors for all disorders except drug use, simple phobia, agoraphobia, and dysthymia. The most extreme case is the estimate for hypomania, which has an estimated standard error 9.9 times that of the unweighted estimate. We initially thought that these large increases might be attributable to instability in the jackknife repeated replications method for estimating variance, but standard errors based on jackknifing of individual cases yielded similar results. The high estimated variances are caused by a few cases with these disorders receiving abnormally high weights.
How well is the added variance from weighting predicted using equation 1? For the entire sample, the predicted standard error of the final weighted estimate from equation 1 is close to the observed standard error for four disorders, but it overestimates the observed standard error for the other disorders. This overestimation is approximately halved when post-stratification is not used. Results for the subclass are more variable, but equation 1 still has a tendency to overpredict the standard error. Without trimming, the large estimated standard errors for some disorders (caused by outlying weights) are not reflected in the predicted standard errors from equation 1, resulting in a tendency for the formula to underpredict the standard error.

How successful are methods for trimming the weights in limiting the increase in variance while at the same time adjusting for bias? For the entire sample, poststratification without trimming tends to reduce the variance of the raw weighted estimator. Nevertheless, in terms of estimated root mean squared error, this estimator still appears to be worse than the unweighted estimator for nine disorders; it is about the same for three disorders and appears to be better for only two disorders. In the age 15–24 subgroup, poststratification without trimming reduces estimated root mean squared error for four of the 14 disorders and increases the estimated root mean squared error for 10. Thus, poststratification does not appear to be very useful for these estimates.

The final weighted estimator after poststratification and trimming has smaller estimated standard errors than the other weighted estimators, suggesting that trimming has been effective. For the entire sample, this estimator appears to be best in terms of estimated root mean squared error, with an average estimated root mean squared error over all disorders of 0.44 compared with 0.51 for the unweighted estimator. However, gains over the unweighted estimator are not uniform across disorders.

The data in table 2 support the idea that weighting is less useful for estimates in subdomains; root mean squared errors after poststratification and trimming are similar to those from the unweighted estimates. One possible reaction would be to restrict the role of weighting to estimates made for the whole sample, although this could introduce undesirable inconsistencies between estimates for the whole sample and for subdomains.

Individual weight components

Tables 3-6 focus respectively on the household, hold-back, new construction, and nonresponse weighting adjustments. In each table, prevalences, standard

<table>
<thead>
<tr>
<th>Alcohol abuse/dependency</th>
<th>Depression</th>
<th>Social phobia</th>
<th>Conduct disorder</th>
<th>Drug abuse</th>
<th>Simple phobia</th>
<th>Agoraphobia</th>
<th>Dysthymia</th>
<th>Antisocial behavior</th>
<th>Anxiety disorder</th>
<th>Panic disorder</th>
<th>Antisocial personality</th>
<th>Mania</th>
<th>Hypomania</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>24.61</td>
<td>17.98</td>
<td>13.08</td>
<td>12.40</td>
<td>12.53</td>
<td>10.82</td>
<td>6.01</td>
<td>7.10</td>
<td>5.57</td>
<td>5.12</td>
<td>3.38</td>
<td>3.22</td>
<td>1.56</td>
<td>0.94</td>
</tr>
<tr>
<td>Hold-back weight only</td>
<td>24.38</td>
<td>17.89</td>
<td>12.65</td>
<td>12.25</td>
<td>12.41</td>
<td>10.69</td>
<td>5.92</td>
<td>7.09</td>
<td>5.51</td>
<td>5.08</td>
<td>3.39</td>
<td>3.19</td>
<td>1.55</td>
<td>0.97</td>
</tr>
<tr>
<td>Final weight</td>
<td>23.55</td>
<td>17.67</td>
<td>13.34</td>
<td>12.95</td>
<td>11.96</td>
<td>11.28</td>
<td>6.68</td>
<td>6.40</td>
<td>5.16</td>
<td>5.15</td>
<td>3.51</td>
<td>3.17</td>
<td>1.67</td>
<td>1.02</td>
</tr>
</tbody>
</table>

### TABLE 5. Effect of new construction weights on estimates, standard errors, and root mean squared errors of prevalences in the National Comorbidity Survey, 1990–1992

<table>
<thead>
<tr>
<th>Alcohol abuse/dependency</th>
<th>Depression</th>
<th>Social phobia</th>
<th>Conduct disorder</th>
<th>Drug abuse</th>
<th>Simple phobia</th>
<th>Agoraphobia</th>
<th>Dysthymia</th>
<th>Antisocial behavior</th>
<th>Anxiety disorder</th>
<th>Panic disorder</th>
<th>Antisocial personality</th>
<th>Mania</th>
<th>Hypomania</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unweighted</td>
<td>24.61</td>
<td>17.98</td>
<td>13.08</td>
<td>12.40</td>
<td>12.53</td>
<td>10.82</td>
<td>6.01</td>
<td>7.10</td>
<td>5.57</td>
<td>5.12</td>
<td>3.38</td>
<td>3.22</td>
<td>1.56</td>
<td>0.94</td>
</tr>
<tr>
<td>NC* weight only</td>
<td>24.56</td>
<td>18.00</td>
<td>12.77</td>
<td>12.40</td>
<td>12.61</td>
<td>10.96</td>
<td>6.02</td>
<td>7.25</td>
<td>5.48</td>
<td>5.15</td>
<td>3.50</td>
<td>3.21</td>
<td>1.52</td>
<td>0.98</td>
</tr>
<tr>
<td>Final weight</td>
<td>23.55</td>
<td>17.07</td>
<td>13.34</td>
<td>12.95</td>
<td>11.96</td>
<td>11.28</td>
<td>6.68</td>
<td>6.40</td>
<td>5.16</td>
<td>5.15</td>
<td>3.51</td>
<td>3.17</td>
<td>1.67</td>
<td>1.02</td>
</tr>
</tbody>
</table>

### Standard error

| Unweighted               | 0.75       | 0.54          | 0.59            | 0.38        | 0.56        | 0.47        | 0.3        | 0.26                | 0.27           | 0.17          | 0.18                   | 0.17  | 0.18     | 0.34   | 0.51   |
| Hold-back weight only    | 0.76       | 0.53          | 0.57            | 0.57        | 0.55        | 0.48        | 0.3        | 0.25                | 0.28           | 0.18          | 0.17                   | 0.16  | 0.1     | 0.36   | 0.56   |
| All but hold-back weight | 0.84       | 0.64          | 0.75            | 0.65        | 0.65        | 0.52        | 0.36       | 0.29                | 0.24           | 0.25          | 0.24                   | 0.19  | 0.27     | 0.44   | 0.44   |
| Final weight             | 0.79       | 0.58          | 0.7             | 0.53        | 0.55        | 0.38        | 0.29       | 0.25                | 0.24           | 0.19          | 0.27                   | 0.12  | 0.27     | 0.42   | 0.42   |

### Root mean squared error

| Unweighted               | 0.94       | 0.61          | 0.59            | 0.62        | 0.56        | 0.69        | 0.72        | 0.75                | 0.26           | 0.44          | 0.29                   | 0.18  | 0.34     | 0.51   | 0.56   |
| NC weight only           | 0.68       | 0.61          | 0.6             | 0.52        | 0.54        | 0.7         | 0.91        | 0.25                | 0.39           | 0.22          | 0.19                   | 0.24  | 0.29     | 0.5    | 0.36   |
| All but NC weight        | 1.23       | 0.65          | 1.18            | 0.88        | 0.58        | 0.75        | 0.54        | 0.45                | 1.39           | 1.04          | 1.01                   | 1.16  | 1.01     | 1.26   | 0.87   |
| All weights, no PS       | 1.04       | 0.76          | 1.01            | 0.94        | 0.55        | 0.69        | 0.59        | 0.67                | 1.31           | 1.07          | 0.97                   | 1.06  | 0.95     | 0.85   | 0.85   |
| Final weight             | 0.79       | 0.58          | 0.7             | 0.53        | 0.55        | 0.38        | 0.29       | 0.25                | 0.24           | 0.19          | 0.27                   | 0.12  | 0.27     | 0.44   | 0.42   |

* PS, poststratification.
errors, and root mean squared errors for estimates based on the entire sample are shown for five weighting schemes: 1) unweighted, 2) weighting only by the individual component weight, 3) weighting by all but the individual component weight, 4) weighting by all of the weighting components, and 5) the final weight, including poststratification and trimming. Changes from scheme 1 to scheme 2 and from scheme 3 to scheme 4 show the effects of the component weight, without and with the other weights included; scheme 5 allows comparisons with the final estimators.

Of the four component weights, the nonresponse weight has the largest impact on estimated prevalences, raising them by nearly 1 percentage point, on average. The impact of weighting for nonresponse is quite significant for some of the rare disorders that might be associated with higher rates of nonresponse—antisocial behavior, anxiety disorder, panic disorder, antisocial personality—and the estimated prevalence of hypomania is doubled from 0.9 percent to 1.9 percent. However, the estimated variances of estimates weighted for nonresponse are much higher than the estimated variances of the unweighted or trimmed estimates. In addition, the effects of the nonresponse weights on final estimated prevalences are reduced considerably by trimming.

The within-household weight places more weight on sampled individuals in large households. The effect of the weight is to reduce the estimated prevalences of alcoholism, depression, and drug use by approximately 1 percent. Effects on other disorders are more minor. There is some evidence that the within-household weight operates in the opposite direction from the nonresponse weight, so the two weighting adjustments tend to cancel each other out. For example, adding the household weight to the unweighted estimates increases the variance of the prevalences by approximately 10 percent, on average. However, adding the within-household weight after all of the other weights are included has the effect of reducing the variances by approximately 40 percent, on average, reflecting a reduction in the weight variability from inclusion of this component.

The other two weighting components, the new construction weight and the hold-back weight, have much more minor effects on the estimates and their standard errors, presumably because these weights apply only to small proportions of the sample.

DISCUSSION

The effects of weighting on the point estimates were modest from a substantive point of view; the unweighted estimates were far from disastrous. However, it would be dangerous to conclude from this
single application that weights can always be ignored in surveys of mental disorders. Bias has a more insidious effect on inferences than does variance, since it is often left unmeasured, and in general we prefer an approach that corrects for bias provided that the impact on variance is not too great. From this perspective, our general conclusion is that the overall weighting strategy, with poststratification and trimming, was quite successful, resulting in reduced root mean squared error for the estimates made for the entire sample and root mean squared errors comparable with those for the unweighted estimates in the age 15–24 subdomain. In the latter case, the superior theoretical properties of the weighted estimates seem to justify weighting, even though useful reductions in root mean squared error were not found empirically.

An important limitation of this study is that it reports results for a single survey, albeit an important one. Other surveys may have different sample sizes, weighting characteristics, and weight trimming strategies, so it is hard to generalize the findings of the present study. However, some tentative comments can be made. The raw weights without adjustment added much too much variance in the effort to limit bias. This finding suggests that the increased variance from weighting should be carefully monitored. This is easily achieved by directly computing the variances with and without the weights, using a method such as jackknife repeated replications that takes into account other features of the sample design, and checking that the increase in variance is not excessive. The alternative approach of using equation 1 to predict the increase in variance does not appear reliable, based on our empirical evidence.

Methods for reducing added variance from weighting were clearly needed, and the trimming adjustments were effective. However, from a theoretical perspective, trimming is ad hoc, and alternative approaches to variance reduction may yield improved estimates. These alternatives include weight smoothing based on random-effects models (6, 19) or estimation based on predictions from a model that includes the weight as a covariate (25, 26). Our future work will focus on these alternative variance reduction methods.

Another limitation of this study is that it examined estimated rates for the entire sample and for one subdomain. Thus, we did not consider measures of association, such as differences in mean values or regression coefficients. Theory suggests that weighting may be less important for bias reduction in such analyses, since biases in comparisons tend to cancel out. Given the modest support for weighting our study provides for estimated rates, we would not expect weighting of analyses of associations to have much payoff in the setting of the NCS, although that is an empirical question worthy of further exploration.

In this study, the most important components of the weighting adjustments were the nonrespondent and within-household weights. The new construction and hold-back weights had a limited impact on root mean squared error, presumably because they affected a relatively small fraction of the sample. Since weights for new construction generally affect only a small fraction of the sample, we expect our findings on that weight to generalize to other studies. The impact of the hold-back weight would be greater if the strategy of subsampling hard-to-reach cases were applied to the entire sample, rather than to a subset, as was the case in the NCS. More widespread subsampling of hard-to-reach persons has the potential of substantially reducing survey costs, and we are currently studying the relative efficiency of variants of such a design in the context of the NCS.

ACKNOWLEDGMENTS

This work was supported by National Institute of Mental Health grants MH46376 and MH49098, with supplemental support from the National Institute of Drug Abuse and the W. T. Grant Foundation.

REFERENCES

13. Korn EL, Graubard BI. Analysis of large health surveys.