A Simulation Study of Post-Adjustment Bootstrapping of Final Weights in the General Social Survey

Michael Wendt[1], David Paton[2]
[1] Social and Aboriginal Statistics Division, Statistics Canada
[2] Canadian Institute for Health Information

Abstract

The General Social Survey is a cross-sectional survey that gathers social information on Canadians, now in its twenty-second annual cycle. As part of processing in recent cycles, bootstrap weights have been provided to researchers for variance estimation. Creation of final survey weights and these bootstrap weights depends upon detailed design information. In early cycles, final weights have been provided but no bootstrap weights are available. Additionally, some design information has been lost over the years. Nevertheless, enough information is available to attempt to "recreate" bootstrap weights by working backwards from final weights extant to a point in the processing steps where bootstrap samples are selected and then working forward. In this document, we provide details of a simulation study on two newer cycles comparing estimates based on bootstrap weights already created with those under this new proposed method. The simulation study provides evidence that the proposed method would be appropriate for earlier cycles.

Key Words: survey weights, bootstrap weights, estimation in complex surveys

1. Introduction

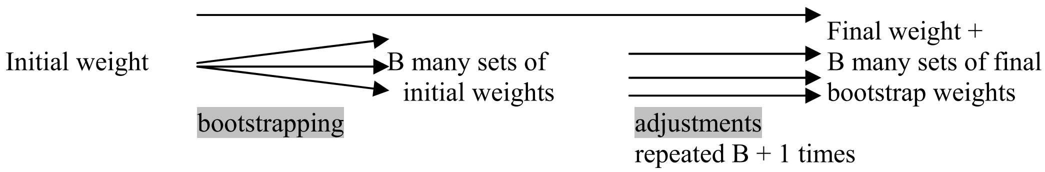
The General Social Survey (GSS) currently provides bootstrap weights to users for estimation of variance. The respondent-level Public Use Micro-data Files contain from 200 to 500 such weights depending on the annual cycle. Construction of these bootstrap weights depends on detailed design information. In essence, the process is to first select bootstrap samples from the original sample using the original sampling design[i], derive bootstrap sampling weights, based on the initial sampling weights (which represent the probability of selection of each telephone number[ii]). After that, we perform various weight adjustments such as turning the weights into person-level weights, adjustments for non-response and calibration to certain known totals on the bootstrap weights.

In the GSS, bootstrap weights are available for GSS-8 and GSS-10 to GSS-20. Researchers find these straightforward to use and much software exists that can incorporate bootstrap weights for variance estimates of many types of estimates. The question arises: would it be possible to construct bootstrap weights for GSS-1 to GSS-7 and GSS-9? After extensive research, we have determined that it would not be feasible to construct such weights for those other cycles. Detailed data for the design information has been lost over the years[iii]. However, final weights for each cycle do exist and they encapsulate this design information. In addition, some extra information about stratification and post-stratification is available for all cycles. It turns out that enough information exists in all cycles that we can modify the question above and ask two questions instead:

- Can we use the information that exists and somehow "un-adjust" the final weights to recreate a version of the original sample and initial weights, bootstrap the result and perform the adjustments over these newly bootstrapped weights?

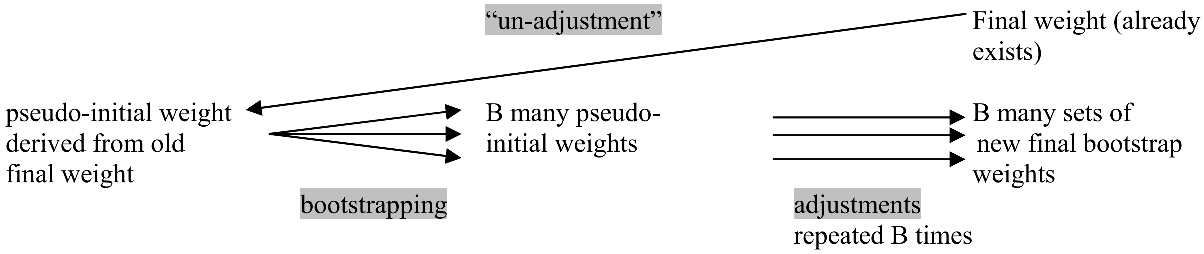
- Would such weights give us a useful variance estimator?

In this document, we present the results of a simulation study on a proposed method of calculating bootstrap weights for the GSS past cycles, when only the final weights exist and some other, limited information is available. The following diagram exhibits the current process that creates a final set of survey weights as well as the sets of final bootstrap weights (B represents the number of sets of bootstrap weights and actions are in gray):



The adjustments for the bootstrap weights reproduce the adjustment from the (one) initial weight to this final weight and do this based on each bootstrap replicate sample.

We propose to do the following:



Note that the adjustments would only be repeated B many times (to the initial bootstrap weights) as the final weight already exists and does not need to be recreated.

2. Proposed Method

To better describe our proposed method, a brief overview of the GSS weighting process is useful. GSS is a Random Digit Dialling Survey. Some telephone numbers of those originally attempted reach households. When a household is reached, one eligible person is chosen[iv] and that person may or may not respond. The weighting process begins with an initial weight, which represents the probability of selection of the household via that telephone. This weight is adjusted for non-response (generally, by a simple ratio of counts) and then turned into a person-level weight by multiplying by the number of eligible people in the household and dividing by the number of (personal use) telephones. These raw person-level weights are calibrated, via the raking ratio algorithm, to known population totals at the province-sex-age group and regional office-province-month of data collection levels[v]. Bootstrapping is applied to the full sample of households before the non-response adjustment. Variations in the bootstrap samples make the non-response adjustment different for different sets of bootstrap weights.

As noted, for GSS-1 to GSS-7 and GSS-9, bootstrap weights have not been constructed, though we do have one set of final weights and some, but not enough, of the information needed to fully reconstruct the weighting process can be found. Based on the information available in these earlier cycles, we created a method and simulated it using GSS-18 and GSS-20 data to observe the potential of the proposed method. This means that we compared the old (= actual) bootstrap weights with simulated new bootstrap weights. Both resulting weights and estimates were compared. The estimates used in estimating typical design effects were used for the comparisons.

The proposed method to get new bootstrap weights from old final weights is as follows:

1. start with the old final weights

2. turn them back into “telephone weights”; we cannot undo the raking but we can divide by the number of eligible people and multiply by the number of telephones

3. apply (provincial-level) non-response at the stratum-level to create a “households” file with respondents and pseudo-non-respondents; respondents are the records we have; pseudo-non-respondents are records with a stratum identifier only; for each stratum, enough of these are created based on the known provincial-level of non-response; this gives us a pseudo-initial sample

4. create pseudo-initial weights by distributing the stratum sums of the telephone weights equally among all stratum cases (both responding and non-responding cases)

5. bootstrap the pseudo-initial sample and pseudo-initial weights

6. adjust each set of bootstrap weights for non-response; the bootstrap sub-samples will contain varying numbers of respondents and pseudo-non-respondents so the adjustment for non-response would vary across bootstrap samples

7. turn these weights back to person-level by multiplying by the number of eligible people in the household and dividing by the number of telephones

8. rake these weights by province-sex-age group and regional office-stratum [vi]; these levels have been suitably collapsed so that there are at least 15 records in each control grouping

9. these final raked, non-response adjusted, person-level sets of bootstrap weights will be considered our new bootstrap weights.

Steps 1 to 4 represent turning the final weights into something as close as possible to initial weights. Steps 5 to 8 mimic the current weighting process.

3. Simulation Study, National Estimates

In this simulation study, we used GSS-18 and GSS-20 data. GSS-18 has 200 sets of final bootstrap weights and GSS-20 has 500. Throughout, we shall use the term "old" to refer to these weights or estimates using these weights. For GSS-18, we looked at 305 estimates and for GSS-20, we considered 514 estimates. These estimates were all of characteristic variables representing the categorical variables used in the respective design effect calculations. For example, in GSS-18, the variable ACMYR is "main activity of the respondent in the last 12 months" has values like 1 = "working at a paid job or business," 2 = "looking for a job," etc. For each of these, we constructed a characteristic variable, ACMYR_1 (ACMYR = 1, yes or no), ACMYR_2, ACMYR = 2, yes or no), etc. For each of the characteristic variables, we estimated the proportion of yes values in the population and used the bootstrap weights to estimate[vii] the variance of this proportion. In fact, two sets of estimates and weights were created and compared: the old weights and estimates, using the bootstrap weights extant, and the "new" weights and estimates, using the method described in the previous section.[viii]

The idea of our simulation study is that if variances produced by either path, old or new, are similar enough for important variables such as those used in design effect calculations in both GSS-18 and GSS-20, then, inasmuch as procedures for earlier cycles were similar to those currently in use, we could safely apply the proposed method to GSS-1 to GSS-7 and GSS-9 to provide pseudo-bootstrap weights for these cycles that would be useful for analysts. GSS-18 and GSS-20 are sufficiently different in terms of response rates, topic, and other factors to represent a broad enough spectrum of cycles and so we would feel comfortable extended our results for these two to earlier cycles.

For each of GSS-18 and GSS-20, 20 simulations were performed. That is, for each cycle, 20 different complete collections of new bootstrap weights and new estimates were compared to old weights and old estimates. Over the 20 simulations, results were quite similar.

Table 1 presents results for one arbitrarily chosen simulation in each cycle. Comments are provided after the table. In the top part, the number of old and new publishable estimates are given (i.e., CV < 16.5%). The number of estimates that would be "lost" is given next. This means a CV < 16.5% becomes a CV of ≥ 16.5%. The bottom part provides a numerical comparison of old and new variances. For each new publishable estimated proportion, we computed the old bootstrap standard error and the new bootstrap standard error and computed the ratio of new to old:

R = new standard error of estimate / old standard error of estimate.

The idea is that if this ratio is close to 1, then variances (for the chosen estimates, at least) are pretty much equivalent, leading us to believe that the methods would produce similar results in significance tests, for example. The mean and standard deviation of R and a five number summary for each of the simulations are on the bottom.

Table 1: Comparison of old and new estimates for one simulation in each of GSS-18 and GSS-20

Item	GSS-18	GSS-20
Total number of point estimates	305	514
Number of old publishable estimates	262	412
Number of new publishable estimates	263	411
Would lose from old to new	0	2
Would lose from new to old	1	1
Average(R) among new publishable	1.004	0.987
Std(R) among new publishable	0.073	0.0453
Min(R) among new publishable	0.844	0.875
Q1(R) among new publishable	0.950	0.958
Median(R) among new publishable	0.996	0.986
Q3(R) among new publishable	1.052	1.018
Max(R) among new publishable	1.207	1.150
Percentage of R < 1 among new publishable	52.3	60.6

Table 1 may be summarized simply as "both types of variance estimates are similar." Indeed, the number of publishable point estimates (those with a CV < 16.5%) doesn't vary much in either method. "Switching" from old to new, or from new to old, would have little impact on the *set* of publishable point estimates. In fact, looking at the actual CVs one sees that even the "lost" estimates go from a CV near 16.5 to a CV only slightly over 16.5. In short, if only the new type of weights were available and not the old, the loss, *in terms of counts among these estimates*, would not be that dramatic.

We also looked at the potential magnitude of change with new type replacing old type. The idea is that we could try to determine "if we only had the new, what could we potentially lose in accuracy?" We cannot answer the question completely but Table 1 provides some detail. For example, among the 262 new publishable GSS-18 estimates, the ratio of new to old bootstrap standard error is very close to one. Specifically, the R's range from 0.844 to 1.207 with an average of 1.004 and slightly more of them (52.3%) were less than 1 than 1 or more. Among, the 411 new publishable GSS-20 estimates, a higher percentage of the ratios of new to old standard error were less than 1[ix]. Nevertheless, the numbers are mostly close to 1. The following diagrams exhibit scatter-plots of standard errors (among new publishable estimates) for one of the twenty simulations. One sees that most of the R's are very near 1.

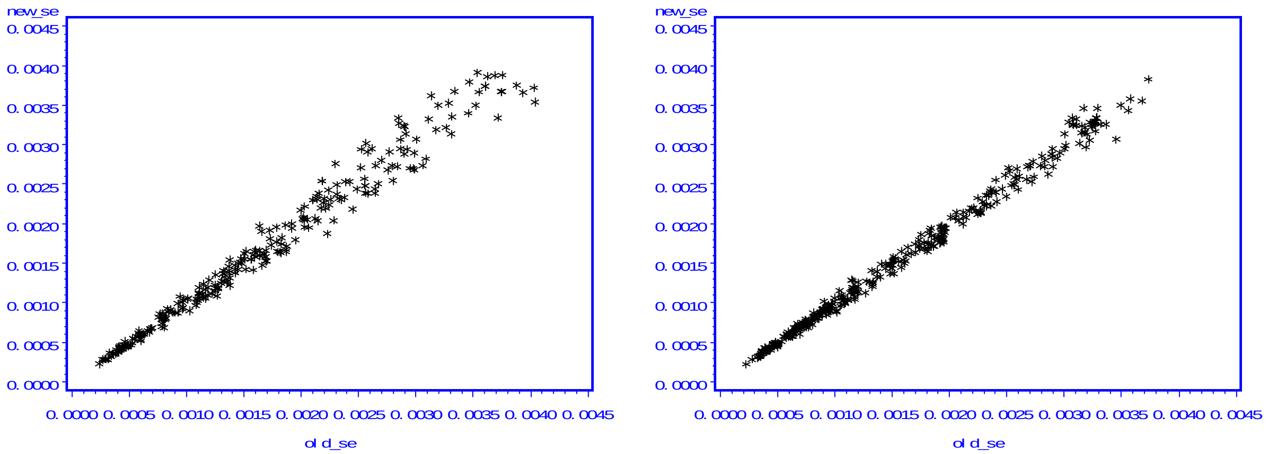


Figure 1: Scatter-plots for "old" vs. "new" standard errors for GSS-18 (left) and GSS-20 (right)

It seems, for these two simulations at least, we could suggest that if new bootstrap weights were constructed for the early cycles, researchers could be reasonably confident in the computed variances.

Additionally, we performed a detailed comparison of one set of old weights with the corresponding set of new weights. Certainly, estimated population totals among certain groups would be the same, since the weights were calibrated. But, more importantly, the old and new weights were very similar (in terms of a five number summary, for example).

Recall, we actually performed 20 simulations for each cycle. Table 1 above gives the results for only one simulation each. A complete listing of results is presented in the Appendix. Some variation was observed across the 20 simulations. For example, one simulation in GSS-20 had 18 old publishable estimates that would not be new publishable. Further investigation revealed these to be borderline (near CV = 16.5%) cases. Generally, though, the number of new publishable and old publishable estimates was nearly the same. The ratios of old standard error to new standard error were generally near 1. The widest range in GSS-18 was R from 0.807 to 1.331 and, for GSS-20, the widest range was R from 0.818 to 1.118. In both cycles, in many simulations, more of these ratios were less than 1 than 1 or more[x]. This does not necessarily mean that the new standard error of a given estimate (not on our list) is probably lower than the old standard. Again, we would simply suggest to users that they be cautious about making rejection or non-rejection decisions when p-values of tests are near the chosen rejection boundary.

4. Simulation Study Provincial Estimates

The above results for National-level estimates seem promising and point to the usefulness of the method. Many researchers require a finer level of detail. We performed another simulation study based on GSS-18 data using the 305 estimates above crossed with the ten provinces. That is, we did a simulation study and compared old and new estimated variances for 305 x 10 = 3050 point estimates. Results similar to those of Table 1 above were found:

Table 2: Comparison of old and new estimates for one simulation in GSS-18, provincial-level estimates

Item	GSS-18
Total number of point estimates	3050
Number of old publishable estimates	1869
Number of new publishable estimates	1872
Would lose from old to new	22
Would lose from new to old	25

Item	GSS-18
Average(R) among new publishable	1.001
Std(R) among new publishable	0.070
Min(R) among new publishable	0.777
Q1(R) among new publishable	0.951
Median(R) among new publishable	0.998

Item	GSS-18
Q3(R) among new publishable	1.046
Max(R) among new publishable	1.286
Percentage of R < 1 among new publishable	49.75

From Table 2, one can see slightly more "volatility" between the old and new estimated variances. Nevertheless, it seems that the old and new methods would produce very close results.

5. Recommendations and Conclusions

The recommendation is to perform this new bootstrapping method for the GSS cycles for which bootstrap weights do not currently exist. Before going into details of this recommendation, we describe the current situation[xi] by cycle. It is summarized in Table 3:

Table 3: Current Status of GSS Variance Estimation Ability

Cycle	Status
1, 2, 6, 9	No bootstrap weights exist; direct[xii] variance estimation can be performed; can do limited kinds of variance estimates (variance for total, mean, ratio of two variables)
3, 4, 5	No bootstrap weights exist; no variance programs exist; not enough cluster information exists to produce "good" direct variance estimation programs; can only compute estimates but no "good" variance estimators
7	No bootstrap weights exist; direct variance estimation program can be found; could do limited kinds of variance estimates (variance for total, mean, ratio of two variables)
8, 10 - 18	200 bootstrap weights exist; can compute many different kinds of estimates and their variances using BOOTVAR, SUDAAN, STATA, etc.
19, 20	500 bootstrap weights exist; can compute many different kinds of estimates and their variances using BOOTVAR, SUDAAN, STATA, etc.

Based on this table, our choices seem to be to use bootstrap weights for GSS-8 and GSS-12 to GSS-20, use direct variance estimation for GSS-1, GSS-2, GSS-6, GSS-7, and GSS-9, and have no variance estimation for GSS-3 to GSS-5. This is somewhat unsatisfying.

Fortunately, some additional information is available. For all cycles, we do have a final (person-level) weight. This weight has, in many cases been adjusted and calibrated. We also have the sampling stratum variable available in all cycles. Additionally, we can construct post-strata by province-sex-age group and regional office-stratum for all cycles.

We have variables to convert back and forth from person-level to telephone-level weights. Non-response rates at the provincial-level have been found for almost all cycles. We are currently investigating their existence for all cycles. In short, we can perform the 9 steps described in section 2 and this would be relatively easy to implement.

Users are familiar with bootstrap weights. Software exists for computing variance estimates for totals, ratios, means, but also for regression coefficients, logistic regression coefficients, and many other entities such as estimation of proportional hazards models, tobit regression, etc. On the other hand, programs exist that compute direct variances for totals and ratios for some cycles but it would be a major undertaking to extend these to regression coefficients, survival analysis, etc.

Another advantage of using bootstrap variance estimates instead of the direct variances is that bootstrap variance includes the variability due to non-response and all the weight adjustments, which the direct variance likely does not.
The simulation study has shown, for two reasonably representative cycles, the new method would provide estimates in many cases with standard errors close to those given by the old, current method. Based on the simulation study, the availability of information, the relative ease of implementation, and the utility of bootstrap weights, we recommend the construction of "new type" bootstrap weights for GSS-1 to GSS-7 and GSS-9.

Appendix: Raw SAS Output

The following is complete listing of 20 GSS-18 and 20 GSS-20 simulations sorted by increasing range(R).

GSS-18

Obs	num_old_pub	num_new_pub	n_oyn	n_ony	n_ci_oin	n_ci_nio	mean_r	std_r	min_r	median_r	max_r	range_r	p_r_u1
1	262	264	0	2	305	305	0.98963	0.066877	0.84305	0.98586	1.17970	0.33664	60.2273
2	262	264	2	4	305	305	0.99247	0.066928	0.82325	0.98653	1.16888	0.34563	58.3333
3	262	263	0	1	305	305	0.99390	0.064104	0.85135	0.98814	1.20558	0.35423	60.4563
4	262	263	0	1	305	305	1.00434	0.072698	0.84372	0.99578	1.20656	0.36284	52.8517
5	262	262	1	1	305	305	0.99878	0.065868	0.84481	0.99760	1.20830	0.36349	52.2901
6	262	263	1	2	305	305	0.99747	0.070915	0.84593	0.99461	1.22999	0.38406	54.3726
7	262	262	2	2	305	305	1.00175	0.065689	0.84195	0.99829	1.23235	0.39040	51.9084
8	262	262	2	2	305	305	0.99948	0.067631	0.84609	0.99149	1.23771	0.39162	54.1985
9	262	264	0	2	305	305	1.00633	0.067174	0.81349	1.00148	1.20546	0.39197	49.2424
10	262	262	0	0	305	305	1.00362	0.065033	0.81435	1.00009	1.21073	0.39637	50.0000
11	262	262	1	1	305	305	1.00652	0.071429	0.83035	1.00639	1.22945	0.39910	47.3282
12	262	264	0	2	305	305	0.99385	0.068462	0.81681	0.99356	1.21991	0.40310	54.9242
13	262	264	0	2	305	305	0.99603	0.073426	0.79830	0.99373	1.20774	0.40944	55.3030
14	262	262	1	1	305	305	1.00013	0.069569	0.81441	0.99627	1.24434	0.42993	52.2901
15	262	262	1	1	305	305	0.99468	0.068181	0.76771	0.99222	1.20371	0.43600	55.3435
16	262	263	0	1	305	305	0.99568	0.071379	0.79771	0.99721	1.24123	0.44352	52.4715
17	262	262	1	1	305	305	0.98669	0.081274	0.78948	0.98512	1.23443	0.44495	56.1069
18	262	263	0	1	305	305	0.99992	0.074383	0.82698	0.99330	1.31272	0.48574	56.2738
19	262	261	3	2	305	305	0.99793	0.071482	0.78918	0.99162	1.31124	0.52206	53.2567
20	262	263	2	3	305	305	1.00155	0.070155	0.80696	1.00512	1.33113	0.52417	48.2890

GSS-20

Obs	num_old_pub	num_new_pub	n_oynn	n_onny	n_ci_oin	n_ci_nio	mean_r	std_r	min_r	median_r	max_r	range_r	p_r_u1
1	412	412	1	1	514	514	0.99193	0.044017	0.88234	0.98794	1.11314	0.23081	59.4660
2	412	413	1	2	514	514	0.99224	0.046443	0.88412	0.99138	1.12100	0.23688	54.9637
3	412	412	1	1	514	514	0.98821	0.040259	0.86938	0.98882	1.10847	0.23910	65.5340
4	412	413	1	2	514	514	0.98905	0.044744	0.86713	0.98375	1.10974	0.24261	60.0484
5	412	412	2	2	514	514	0.98798	0.046589	0.87847	0.98657	1.12686	0.24839	60.9223
6	412	412	1	1	514	514	0.98928	0.043060	0.86713	0.99069	1.12302	0.25588	58.2524
7	412	397	16	1	514	514	0.98854	0.048365	0.85728	0.99122	1.11339	0.25611	59.4458
8	412	411	2	1	514	514	0.99147	0.044788	0.84605	0.99708	1.10396	0.25791	55.4745
9	412	412	1	1	514	514	0.99151	0.046783	0.87028	0.99525	1.12977	0.25949	54.3689
10	412	397	17	2	514	514	0.98976	0.038824	0.87045	0.98786	1.13096	0.26051	58.1864
11	412	414	0	2	514	514	0.99996	0.044746	0.87424	1.00341	1.13796	0.26372	47.5845
12	412	411	2	1	514	514	0.98720	0.045283	0.87546	0.98636	1.14998	0.27452	60.5839
13	412	412	2	2	514	514	0.99520	0.052508	0.87622	0.98891	1.15098	0.27476	56.5534
14	412	413	1	2	514	514	0.98959	0.053460	0.85205	0.99087	1.12725	0.27520	56.9007
15	412	414	0	2	514	514	0.99074	0.044973	0.85496	0.98827	1.13245	0.27750	58.4541
16	412	395	18	1	514	514	0.98994	0.042157	0.85913	0.98779	1.14169	0.28255	63.2911
17	412	412	2	2	514	514	0.98637	0.050920	0.82230	0.99136	1.11035	0.28805	59.2233
18	412	411	1	0	514	514	0.98770	0.042616	0.83759	0.98253	1.12571	0.28811	61.3139
19	412	413	1	2	514	514	0.99772	0.046056	0.85185	0.99848	1.14994	0.29808	53.2688
20	412	411	1	0	514	514	0.98761	0.046546	0.81798	0.99109	1.11788	0.29991	57.6642

References

[i] In the GSS, for confidentiality reasons, we use the "mean bootstrap" with R = 25. For more details of this method, the reader is referred to Yung, W., (1997). Variance Estimation for Public Use Files Under Confidentiality Constraints, in *Proceedings of the Survey Research Methods Section*, Washington, D.C.: American Statistical Association, pp. 434-439.

[ii] GSS is a Random Digit Dialing Survey where telephone numbers are selected randomly. If a household is reached, one eligible member is randomly selected for the interview.

[iii] Just what is no longer available depends upon the cycle. Sometimes, non-households records (needed for detailed non-response adjustment) are not available, for example.

[iv] A household roster is formed and a person (typically over 15) is randomly selected).

[v] In some cycles, a third dimension, reference day for time-use diary, is used.

[vi] Month was included after GSS-6.

[vii] We used a modified version of Statistics Canada's Bootvar to estimate variances.

[viii] In fact, we can compare these two sets of estimates with a third type. In some, but not all, older cycles, "direct" variance estimation programs exist. These compute either totals or ratios and estimate their variances using Taylor linearization (in the case of ratios) and assuming stratified random sampling. In an earlier version of this study, we compared, old, new, and direct, and found results quite similar to the study at hand. For this reason, details of the comparison with direct estimation are not presented here.

[ix] GSS-20 had a higher level of non-response than GSS-18, which may account for this.
[x] This slight tendency to underestimate the variance could be caused by steps 3 and 4 of the proposed method.
[xi] This is at the respondent file level. Some cycles have extra information at the time-use level, episode of victimization level, etc.
[xii] As was mentioned in endnote viii, for some older cycles, "direct" variance estimation programs exist. These estimate their variances using Taylor linearization (in the case of ratios) and assume stratified random sampling.