

**National Congregations Study
Weights for Combined Wave I and Wave II Datasets**

Author: Stephanie Eckman, National Opinion Research Center
 Edited by Mark Chaves, Duke University (last edited on 7/17/08)

There are eight weight variables, each of which allows for a different kind of analysis. The weight variables are:

- W1: Weight for all cases, including the panel cases, ignoring duplicate nominations. This weight allows users to analyze the data at the congregation level. This weight is analogous to WEIGHT2 on the 1998 data file.
- W2: Weight for all cases, including the panel cases, taking account of duplicate nominations. This weight allows users to analyze the data at the congregation level.
- W3: Weight for all cases, including the panel cases, which allows users to analyze the data at the attendee level. This weight is analogous to WEIGHT1 on the 1998 data file.
- W4: Weight for all cases nominated in the 2006 GSS, ignoring duplicate nominations, which allows users to analyze the data at the congregation level.
- W5: Weight for all cases nominated in the 2006 GSS, taking account of duplicate nominations, which allows users to analyze the data at the congregation level.
- W6: Weight for all cases nominated in the 2006 GSS, which allows users to analyze the data at the attendee level.
- W7: Weight for panel cases, ignoring duplicate nominations, which allows users to analyze the data at the congregation level.
- W8: Weight for panel cases, which allows users to analyze the data at the attendee level.

SHORTCUT: USE W2 TO DESCRIBE THE AVERAGE CONGREGATION. USE W3 TO DESCRIBE THE CONGREGATION OF THE AVERAGE ATTENDER.

There are seven sets of completed cases:

Set	Description	Number of Cases
A	1998 data for 1998-nominated congregations that were not selected for the panel.	909
B	1998 data for 1998-nominated congregations that were randomly selected for the panel.	325
C	2006 data for the 1998-nominated congregations that were randomly selected for the panel and not re-nominated in 2006	252
D	2006 data for the 2006-nominated congregations established before or during 1998.	1,194
E	2006 data for the 2006-nominated congregations established after 1998.	50

F	2006 data for the 2006-nominated congregations that were also in the 1998 sample and were randomly selected for the panel.	4
G	2006 data for the 2006-nominated congregations that were also in the 1998 sample but were not randomly selected for the panel	6

For each case in these sets, there are eleven variables relevant to weighting:

- a. YEAR: variable indicating year of data collection (1998 or 2006)
- b. PANEL: dummy variable indicating whether the case was part of the panel survey or not
 - i. If SET='A', PANEL=0
 - ii. If SET='B', 'C', or 'F', PANEL=1
 - iii. Otherwise missing
- c. SET: character variable taking values A – G, as given above
- d. The weight variables, W1 – W8

For each weight variable, we outline the calculation steps, using the following notation:

$$W_{i,step}^{set}$$

where: *i* refers to the weight variable being calculated (1-8 as given above)

set refers to the set of cases (A-G as given above)

step refers to the step in the calculation (0,1,..., final)

W1: Weight for all cases, including the panel cases, ignoring duplicate nominations. This weight allows users to analyze the data at the congregation level.

This weight variable ignores the duplicate nominations in both 1998 and 2006. It is designed to allow researchers to replicate their results from 1998 when WEIGHT2 from the 1998 data file, which ignored duplicate nominations, is applied. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

1. Calculate baseweights for all congregations nominated by GSS 1998 respondents.

The probability of selection of a congregation is proportional to the number of members it has: a large congregation has a higher probability to be nominated, because there is a higher probability that one or more of its members will be in the GSS sample.

$$\Pi_c \propto S_c$$

where the subscript c indexes the congregations. S_c is the size of the congregation as reported by the congregation itself: variable NUMADLTS in the 1998 data file.¹

The baseweight of each congregation is then the inverse of this probability of selection.

$$W1_0^{A,B,C} = \frac{1}{\Pi_c} = \frac{1}{S_c}$$

This baseweight ignores any duplicate nominations. It is identical to an unscaled version of WEIGHT2 in the 1998 data file.

2. Calculate baseweights for all congregations nominated by GSS 2006 respondents.

The 2006 GSS design included subsampling of households, so not all households have the same weight. To approximate what was done above, we used the minimum of the weights of the nominating households in the numerator. For example, if a congregation is nominated by both a subsampled household and a non-sampled household, this weight counts the congregation as nominated by the non-sampled household.

$$W1_0^{D,E} = \frac{\min_{i \in c} (W_i^{2006})}{S_c}$$

This weight adjusts for GSS subsampling of households, but not for the within-household respondent selection probabilities, or for multiple nominations of a single congregation. (The GSS weight variable referred to here is W1 in the GSS weighting documentation.) S_c here is the variable NUMADLTS in the 2006 data file.

3. Identify congregations in the 2006 sample that were established after 1998 nominations and set aside. These congregations had no chance of selection in the previous round and must stand in for all new congregations.

Let α be the sample estimate of the percent of congregations that were established after 1998, calculated after weighting sets D and E by $W1_0$. We used this proportion as a quality check in a later step.

$$\alpha = \frac{\sum W1_0^E}{\sum W1_0^D + \sum W1_0^E}$$

¹ See the appendix for details on how this variable was imputed when it was missing.

The final weight of the cases in set E is equal to the baseweight.²

$$W1_1^E = W1_0^E$$

4. The 1998 panel cases and the 2006 cases that were established before 1998 are each a national probability sample of older congregations. To combine these two samples, we developed a trade-off parameter:

$$W1_1^D = \lambda * W1_0^D$$

$$W1_1^C = (1 - \lambda) * W1_0^C$$

We calculated the optimal lambda which equalizes the contributions to effective sample size from each sample. See the appendix for details on the derivation of the optimal lambda.

5. We next reduced the weights of set C and D cases so that their sum is equal to the weighted number of older cases recruited from the 2006 respondents.

$$W1_2^C = \gamma * W1_1^C$$

$$W1_2^D = \gamma * W1_1^D$$

$$\text{where } \gamma = \frac{\sum W1_0^D}{\sum W1_1^D + \sum W1_1^C}$$

This adjustment is necessary so that when data from the older congregations in this round are combined with data from the newer congregations (set E from step 3), the weighted percent of new congregations in the combined sample equals the weighted estimate of the percent of new congregations in the population, α . We checked that $\alpha^* = \alpha$.

$$\alpha^* = \frac{\sum W1_2^E}{\sum W1_2^C + \sum W1_2^D + \sum W1_2^E}$$

6. Rescale sets C, D and E

² The same could be done with congregations nominated by Spanish-only respondents since the 1998 round of the GSS did not interview in Spanish and the 2006 round did. However, not all members of these congregations would be non-English speakers, and thus these congregations probably had a chance of selection in the previous round. We decided against any adjustment of congregations nominated by Spanish-speaking respondents.

Many data analysis programs assume that the sum of the weights is equal to the sample size. Thus it is good practice to rescale the weights to the total number of cases, to ensure correct calculation of standard errors and confidence intervals. Without changing the relative weights between the cases, we rescaled the weights for sets C, D and E so that the sum of the weights is equal to the number of cases.

$$\beta = \frac{|C| + |D| + |E|}{\sum W1_2^C + \sum W1_2^D + \sum W1_2^E}$$

$$W1_{final}^C = \beta * W1_2^C$$

$$W1_{final}^D = \beta * W1_2^D$$

$$W1_{final}^E = \beta * W1_2^E$$

7. Rescale sets A and B

For the same reasons, we rescaled the weights for sets A and B so that the sum of the weights is equal to the number of cases.

$$\varphi = \frac{|A| + |B|}{\sum W1_2^A + \sum W1_2^B}$$

$$W1_{final}^A = \varphi * W1_0^A$$

$$W1_{final}^B = \varphi * W1_0^B$$

W1 is analogous to WEIGHT2 in the 1998 data file, and it is identical to it for the original 1,236 cases from 1998. If users select YEAR = 1998 and weight by W1, results should be identical to those obtained when weighting by WEIGHT2 in the 1998 data file.

W2: Weight for all cases, including the panel cases, taking account of duplicate nominations. This weight allows users to analyze the data at the congregation level.

W2 is very similar in derivation from W1. The only difference comes in the calculation of the baseweights in steps 1 and 2. This weight variable does not ignore the duplicate nominations in both 1998 and 2006. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

1. Calculate baseweights for all congregations nominated by GSS 1998 respondents

The probability of selection of a congregation is proportional to the number of members it has: a large congregation has a higher probability to be nominated,

because there is a higher probability that one or more of its members will be in the GSS sample.

$$\Pi_c \propto \frac{S_c}{\sum_{i \in c} W_i^{1998}}$$

where the subscript c indexes the congregations. The numerator is the size of the congregation as reported by the congregation itself: variable NUMADLTS in the 1998 dataset. The denominator is the sum of the weights of all GSS 1998 respondents that nominated this congregation. Because the 1998 GSS used an equal probability sample of households, and because we are ignoring the selection of respondents within households, this term is equal to the number of nominations a congregation received.

The baseweight of each congregation is then the inverse of this probability of selection.

$$W2_0^{A,B,C} = \frac{1}{\Pi_c} = \frac{\sum_{i \in c} W_i^{1998}}{S_c}$$

2. Calculate baseweights for all congregations nominated by GSS 2006 respondents

The 2006 GSS design included subsampling of households, so the sum of the weights of the nominating respondents cannot be ignored.

$$W2_0^{D,E} = \frac{\sum_{i \in c} W_i^{2006}}{S_c}$$

The numerator sums the weights of all GSS respondents who nominated a congregation. (Again, it is the weight variable W1 in the GSS weighting documentation that we refer to here.)

Steps 3 – 7 are unchanged from W1. All parameters ($\alpha, \alpha^*, \beta, \gamma, \phi, \lambda$) were recalculated for W2.

W3: Weight for all cases, including the panel cases, which allows users to analyze the data at the attendee-level.

This weight will be used by analysts who wish to make statements about the attendees of the selected congregations rather than about the congregations themselves. It is designed to allow researchers to replicate their results from 1998 when WEIGHT1 from the 1998

data file is applied. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

W3 is analogous to WEIGHT1 in the 1998 data file, and it is identical to it for the original 1,236 cases from 1998. If users select YEAR = 1998 and weight by W3, results should be identical to those obtained when weighting by WEIGHT1 in the 1998 data file.

1. Calculate baseweights for all congregations nominated by GSS 1998 respondents

The baseweight of each congregation is the sum of the number of nominations each congregation received.

$$W3_0^{A,B,C} = \sum_{i \in c} W_i^{1998}$$

where c indexes the congregations and i indexes the GSS respondents.

Because the 1998 GSS sample was equal probability sample of households (and we are ignoring within-household respondent selection), all respondent weights are equal to one. Thus the baseweight for each congregation is simply the number of nominations it received.

2. Calculate baseweights for all congregations nominated by GSS 2006 respondents

The 2006 GSS design included subsampling of households, so the sum of the weights of the nominating respondents cannot be ignored.

$$W3_0^{D,E} = \sum_{i \in c} W_i^{2006}$$

This formula adjusts for GSS subsampling of households, but not for the within-household respondent selection probabilities.

Steps 3 – 7 are unchanged from W2. All parameters ($\alpha, \alpha^*, \beta, \gamma, \phi, \lambda$) were recalculated for W3. The interpretation of some of these parameters is different as well. For example, α is here the percent of all congregation attendees that attend congregations established after 1998.

W4: Weight for all cases nominated in the 2006 GSS, which allows users to analyze the data at the congregation level.

W4 is most similar to W1. W4 will be used by researchers who wish to analyze the newly nominated congregations only. Note that only cases in sets D and E have non-missing values of W4. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

1. Calculate baseweights for all congregations nominated by GSS 2006 respondents.

The 2006 GSS design included subsampling of households, so not all households have the same weight. To approximate what was done in the last round, we use the minimum of the weights of the nominating households in the numerator. For example if a congregation is nominated by both a subsampled household and a non-subsampled household, this weight counts the congregation as nominated by the non-subsampled household.

$$W4_{final}^{D,E} = \frac{\min_{i \in c} (W_i^{2006})}{S_c}$$

This weight adjusts for GSS subsampling of households, but not for the within-household respondent selection probabilities or for multiple nominations of a single congregation.

2. Rescale weights.

As in steps 6 and 7 above, we scaled the weights so that the sum of the weights is equal to the number of cases.

W5: Weight for all cases nominated in the 2006 GSS, which allows users to analyze the data at the congregation level.

W5 differs from W4 in the same way that W2 differs from W1: it brings in an accounting for multiple nominations of the same congregation. Note that only cases in sets D and E have non-missing values of W5. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

1. Calculate baseweights for all congregations nominated by GSS 2006 respondents.

The 2006 GSS design included subsampling of households, so the sum of the weights of the nominating respondents cannot be ignored.

$$W2_{final}^{D,E} = \frac{\sum W_i^{2006}}{S_c}$$

The numerator sums the weights of all GSS respondents who nominated a congregation.

2. Rescale weights

As in steps 6 and 7 above, we scaled the weights here so that the sum of the weights is equal to the number of cases.

W6: Weight for all cases nominated in the 2006 GSS, which allows users to analyze the data at the attendee-level.

W6 is most similar to W3. W6 will be used by researchers who wish to analyze the newly nominated congregations only. This weight will allow repetition of analyses run on the 1998 dataset. Note that only cases in sets D and E have non-missing values of W6. For the purposes of this weight, cases in Set F and Set G are considered to be in Set D.

1. Calculate baseweights for all congregations nominated by GSS 2006 respondents.

The 2006 GSS design included subsampling of households, so the sum of the weights of the nominating respondents cannot be ignored.

$$W2_0^{D,E} = \sum_{i \in c} W_i^{2006}$$

The numerator sums the weights of all GSS respondents who nominated a congregation.

2. Rescale weights.

As in steps 6 and 7 above, we scaled the weights here so that the sum of the weights is equal to the number of cases.

W7: Weight for panel cases, ignoring duplicate nominations, which allows users to analyze the data at the congregation-level.

For the purposes of this weight, cases in Set F and G are considered to be in Set B. Even though Set G cases were not randomly selected for the panel, we included them in the panel since there are so few of them. It would be prudent for users of the panel data to check that their results are not affected by the inclusion of the six cases in Set G.

Only cases in Sets B, C, F, and G have non-missing values of W7. In addition, only congregations that completed the survey in both the 1998 and the 2006 rounds will have a non-missing value on W7.

1. Calculate baseweights for all congregations nominated by GSS 1998 respondents.

The probability of selection of a congregation is proportional to the number of members it has: a large congregation has a higher probability to be nominated,

because there is a higher probability that one or more of its members will be in the GSS sample.

$$\Pi_c \propto S_c$$

where the subscript c indexes the congregations. S_c is the size of the congregation as reported by the congregation itself: variable NUMADLTS in the 1998 data file.³

The baseweight of each congregation is then the inverse of this probability of selection.

$$W7_0^{B,C} = \frac{1}{\Pi_c} = \frac{1}{S_c}$$

This baseweight ignores any duplicate nominations.

2. Rescale weights.

As in steps 6 and 7 above, we scaled the weights here so that the sum of the weights is equal to the number of cases.

W8: Weight for panel cases, which allows users to analyze the data at the attendee-level.

For the purposes of this weight, cases in Set F and G are considered to be in Set B. Again, it would be prudent for users of the panel data to check that their results are not affected by the inclusion of the six cases in Set G.

Only cases in Sets B, C, F, and G have non-missing values of W8. In addition, only congregations that completed the survey in both the 1998 and the 2006 rounds have non-missing values on W8.

1. Calculate baseweights for all congregations nominated by GSS 1998 respondents.

The baseweight of each congregation is the sum of the number of nominations each congregation received.

$$W8_0^{B,C} = \sum_{i \in c} W_i^{1998}$$

where c indexes the congregations and i indexes the GSS respondents.

³ See the appendix for details on how this variable was imputed when it was missing.

Because the 1998 GSS sample was equal probability sample of households (and we are ignoring within-household respondent selection), all respondent weights are equal to one. Thus the baseweight for each congregation is simply the number of nominations it received.

2. Rescale weights.

As in steps 6 and 7 above, we scaled the weights here so that the sum of the weights is equal to the number of cases.

Appendix: Imputation of NUMADLTS

The number of regularly attending adults, NUMADLTS, is an integral variable in the calculation of the NCS weights. These data are collected from congregations during the NCS interview. When this variable is missing we must impute it from available data.

The method of imputation depends on the data that are available for a given congregation. We used the first of the following methods that we could:

1. If both NUMTOTAL (the total number of congregation members) and NUMREGLR (the number of regularly attending members) are non-missing, we used regression imputation to estimate NUMADLTS. That is, for the cases where all three variables are non-missing, we estimated a regression equation that predicts the log of NUMADLTS from the logs of NUMTOTAL and NUMREGLR. Then, for cases where NUMADLTS is missing, we estimated it using the coefficients from the regression equation.
2. If only NUMTOTAL is non-missing, we used regression imputation with the log of this variable only.
3. If only NUMREGLR is non-missing, we used regression imputation with the log of this variable only.
4. If both NUMTOTAL and NUMREGLR are missing, we used the variable collected in the nominating round of the GSS. This variable, CONGNUM, is the GSS respondent's estimate of the number of regularly participating adults at his/her congregation. We know from the 1998 NCS that there is some bias in this estimate and that the error is larger for larger congregations, so we again took a log when fitting the regression model and deriving the imputation parameter.
5. If none of the above methods were available, we used mean imputation to fill in the missing values of NUMADLTS.

The above procedure was followed for imputing NUMADLTS in the 2006 data. A somewhat less sophisticated approach was used for imputing NUMADLTS in the 1998 data.

The variable IMPSIZE flags the cases with imputed values on NUMADLTS.

Appendix: Derivation of Optimal Lambda Parameter

From: O'Muircheartaigh, Colm, and Steven Pedlow. 2002. "Combining Samples vs. Cumulating Cases: A Comparison of Two Weighting Strategies in NLSY97." 2002 Proceedings of the American Statistical Association, Survey Research Methods Section [CD-ROM], Alexandria, VA: American Statistical Association, pp. 2557-2562.

To maintain the characteristic that the weights from both samples together sum to the population size (rather than each sample independently), the CX weights were multiplied by λ ($0 < \lambda < 1$), and the SU weights were multiplied by $1 - \lambda$ in producing estimators based on both samples together:

$$\hat{\mu} = \lambda \hat{\mu}_c + (1 - \lambda) \hat{\mu}_s$$

in which $\hat{\mu}_c$ represents a statistic derived from the CX sample and $\hat{\mu}_s$ represents the corresponding statistic from the SU sample. Because the two samples are independent, the optimum λ for a weight of this form is proportional to the relative effective sample size in the CX sample:

$$\lambda = \frac{n_c / d_c}{n_c / d_c + n_s / d_s}$$

$$1 - \lambda = \frac{n_s / d_s}{n_c / d_c + n_s / d_s}$$

in which n_c and n_s are the nominal sample sizes for the CX and SU samples and d_c and d_s represent the design effects for the estimators from each sample. It is inconvenient to use the design effects themselves, since they will vary from one variable to the next. Instead, a general factor was used (one plus the squared coefficient of variation of the weights within each sample), as was done for NLS79; this factor captures the impact of unequal weighting on the sample efficiency:

$$\hat{d}_c = 1 + CV(W_i \in CX)^2$$

$$\hat{d}_s = 1 + CV(W_i \in SU)^2$$