Year 15 Fragile Families Survey Weight Adjustment

Lauren A. Kennedy^ & Andrew Gelman* *^School of Social Work, Columbia University, New York.*
**Department of Statistical Sciences and Department of Political Science, Columbia University, New York.*

This memorandum details the weight creation procedure for the Fragile Families and Child Wellbeing Study Wave 6 conducted at Year 15. This memorandum connects closely with previous weighting memorandums (Carlson, 2008, Si and Gelman (2014)). At this wave, we create survey weights for the primary caregivers, children, and subset of children who participated in the home visit for the cities, national, and national excluding city X.

In this memorandum we cover the creation of survey weights for the primary caregivers, children and families who participate in the home visit for the Year 15 follow-up of the Fragile Families and Child Wellbeing Study. In this memorandum we detail the procedure used to create the survey weights for the primary caregivers, children and home visits. Unlike previous waves, no weights were created specifically for the mothers or fathers. This reflected a change in survey design in the Year 15 whereby only a single primary caregiver was asked to complete the survey, rather than both members of the parental unit. The outline of this memo is a follows; first we will detail the overall methodology, taking care to compare and contrast approaches taken at Year 15 to those taken at previous years. We will then focus on the finer details of creating survey weights for each of the three core target groups (primary caregiver, child and home visit), for each of the three target populations (national, national excluding city X and city).

Overall weighting procedure

In this section we discuss the overall weighting strategy for each of the primary caregiver, child and home visit weights. Broadly each set of weights were created by first considering the eligibility of selection, then adjusting for non-response, followed by post-stratification and construction of replicate weights.

Before this, however, we discuss the anchor. As noted by Carlson (2008), the ultimate sampling unit for the study is the birth, of which the baseline mother interview is used a a proxy. Our aim is to adjust the sample at this wave (Year 15) so that it is representative of this original sampling unit, which in turn was adjusted to be representative of the population of interest (national, cities, or national excluding city X). Following this reasoning through, Carlson (2008) use baseline mother characteristics to adjust the baseline weights for non-response and post-stratification at the wave of interest.

However, Si and Gelman (2014), choose a different approach with Year 9 (wave 5). Instead of adjusting the baseline weights for the wave of interest, they instead adjust the weights of the most previously observed wave, which was in turn adjusted to the baseline. The reasoning is that this most recent wave contains the most recent known information, and will likely be better to adjust for non-response. Whilst we acknowledge the elegance of this solution, new challenges created by the Year 15 survey design resulted in our choice to revert to the method used by Carlson (2008). Of particular consideration in this decision was the movement from mother and father interviews to primary caregiver interviews.

The plan is for each population of interest (national, national excluding city X, and cities) to adjust the mother's baseline weight for this population so that the sample at Year 15 is representative of the original population of interest.

Eligibility of selection

In the creation of weights for this Year 15 follow-up we note a slight discrepancy in the criteria for eligibility between baseline to Year 5 and Year 9. In the baseline to Year 5 survey weights, the only individuals deemed to be ineligible were those where the child had deceased or those entries that were duplicates (Carlson, 2008, pg. 15). In the Year 9 survey weights, ineligible was defined as those entries where the child had deceased or some *other ineligible* was recorded. After some discussion, the decision was made that only those entries where the child had been recorded as deceased would be considered ineligible for the Year 15 weights.

Adjusting for non-response

Both Carlson (2008) and Si and Gelman (2014) adjusted for non-response given eligibility using a two stage process. We follow this same approach for the primary caregiver and child weights. For the home visit weights a more complicated procedure is used, which we describe in a latter section. For the present we focus on the two stage process:

1. Adjust for ability to locate. Survey administrators attempted to locate all children believed to be eligible in the wave.
2. Adjust for non-response given location. Among those who were located, only a subsample chose to respond to the survey. We further adjust the weights of those individuals who were located and responded to account for potential bias in response

To make these adjustments, both Carlson (2008) and Si and Gelman (2014) uses a large number of variables that describe baseline mother characteristics and a form of variable selection (Carlson (2008) uses a step-wise regression, Si and Gelman (2014) use LASSO (Friedman, Hastie and Tibshirani, 2010)) to predict location or response in the wave of interest. In the Year 15 weights we follow with this in-principle use of variable selection to reduce over-fitting, but instead use a type of Bayesian regularization using a regularized horseshoe prior (Piironen, Vehtari et al., 2017), implemented through the package rstanarm(Goodrich et al., 2018).

Regardless of the method of variable selection/regularization, the overall procedure remains the same. Before creating this model, we impute the variables that we will include in the propensity model. We use the statistical package Amelia (Honaker, King and Blackwell, 2011), available in R[1].

Using a logistic regression with either location or response as the outcome variable, we use this set of imputed variables as covariates. Variables that were included in this model are listed in Appendix A. To reduce the computational demand introduced by using a sampled Bayesian method, we first standardize these variables. To achieve regularization and prevent overfitting, we use a regularized horseshoe prior on the β parameters in this model. This intuitively allows us to encode information about the model (namely that some covariates will predict the outcome variable whilst others will not), but does rely on some user input (namely a guess as to the number of covariates expected to be related to the outcome variable).

[1] A small number of variables (m1a4,m1a13,m1b2) were not imputed with Amelia due to collinearity with other variables. Instead these variables were imputed in a naive manner using frequency tables.

Following this regression the propensity for each individual to respond is predicted. Following the rationale outlined by Si and Gelman (2014), we group these propensities into deciles, and adjust based on these cells. As noted by Si and Gelman (2014), this technique reduces the impact of highly variable propensity scores.

Poststratification

Following adjustment for non-response, we then adjust for four key demographic variables: marital status, education, ethnicity and age. For the city weights these variables were adjusted within city. Due to the smaller sample size at this wave, we found that there were some levels of these variables that were not filled at the city level. To adjust this we collapsed some levels for the city post-stratification adjustment. See the tables included in the detailed weight sections for a summary of the number of levels for the poststratfication adjustment at the national, national excluding city X and city weights.

Following the same technique used by Si and Gelman (2014), we too use the survey package (Lumley, 2004). We specify the survey design in the same way, namely one stage cluster sampling and nested stratified sampling. Instead of using the weighted estimates of the previous wave weights to obtain proxy marginal tables for raking, we instead use the baseline weights.

Trimming

Following these adjustments, the survey weights tended to have extreme weights. Using the method suggested by Si and Gelman (2014), we noted that when the weights were trimmed overall estimates seemed biased[2]. We propose a slight adjustment that we feel reduces bias whilst reducing the extreme weights. To our knowledge this method has not been previously proposed as a method for trimming weights[3] so we describe it here for completeness.

The method used in the *survey* package in R uses a method for trimming weights that windsorizes at a particular criterion. Si and Gelman (2014) used a different criteria depending on marital status. We use a single criteria of the 97.5 quantile of the weight distribution. As the package in R currently implements this trimming, the weights are Windsorized at the 97.5th quantile with the excess weight redistributed evenly across all of the weights. We feel the certain heterogeneity within the weights create more bias than desired.

Instead of using this method, we use a different approach to redistribute the excess trimmed weight. After trimming we renormalize the weights so that the sum weights remain the same before and after trimming. We found that this technique reduced bias. The reason for this is that very small weights were increased proportional to the relative size of the weight. The alternative method increases the very small weights by the same amount as the very large weights.

Replicate weights

Replicate weights were created to allow users to estimate standard error when using the survey weights that reflect the complex survey design. We note that in the Year 9 survey weights were created using a Jackknife approach based upon the *natpsu* and *citypsu* representing the primary

[2]No more information is given because it would provide information about variables that have been removed from the public dataset to protect anonymity. Further information if needed is given in a file detailing code requirements and technical specifications, but designed to be distributed only for those with access to the full dataset.

[3]We are currently conducting research and simulation studies to investigate the strengths and weaknesses of this method, but for reasons discussed in the non-public document, we believe this method to be appropriate for this scenario. We will update this document in the future when this research is published

sampling unit (PSU) and *natstratum* and *citystratum* representing the strata structure. Previous waves had chosen not to use this information but instead use a method of random groups (Wolter, 1985) whereby random groups are created to be used in lieu of the actual stratum. Both methods were considered for the Year 15 replicate weights, but eventually the random groups method was decided upon. The deciding factor was that although there was little risk to identifying the strata from the replicate weights, it would be preferable to obscure this information entirely, which the method of random groups allows.

Following the example of Carlson (2008), we use 33 random groups for the national replicate weights and 10 random groups for the city replicate weights. While this method worked well for the primary caregiver weights and child weights, the home visit weights were a sufficiently small sample that the size of each group in the national weights was very small if 33 groups were used. For these replicate weights we used 10 random groups for both the national and city weights

Checks

We conduct a number of checks on the validity of the survey weights, broadly following those proposed by Carlson (2008). The first pass of checks ensures that the sum of the weights match the population total. If this passes, we also check the sum weight in each city[4], and then the same checks for each of the poststratification variables. As these weights are trimmed following the raking step, we do not necessarily expect to see the same numbers for each check. If the weighted counts are dramatically different to those observed in the population, then we would investigate further.

We also run a number of common sense checks to check whether the groups and/or individuals who we would expect to have weights of zero do. In particular, we check that individuals in the four cities not included in the national weights do not have national weights, the individuals in city X and these four cities do not have national-X weights, and the city weights all have weights provided the case is eligible, located and complete.

Primary caregiver weights

Weights for the cases where the primary caregiver completed the primary caregiver survey were named with the following convention. Weights are available for all cases where a) a baseline mother weight was available and b) the primary caregiver was eligible, located and completed the survey.

Table 1: Naming convention for the primary caregiver Wave 6/Year 15 weights

Population	Base.Weight	Replicate.Weight
National	p6natwt	p6natwt_rep1 - p6natwt_rep33
National excl. X	p6natwt	p6natwtx_rep1 - p6natwtx_rep33
City	p6citywt	p6citywt_rep1 - p6citywt_rep10

[4]The national weights do not poststratify on city, so we would not expect the counts to be exact. However, it would also be undesirable to dramatically overweight or underweight a city relative to the original weights.

National weights

There were 3404 cases who both had a baseline national weight and were eligible for the Year 15 wave. Of these cases, 3052 cases were located. Of those who were located, 2642 cases completed the primary caregiver interview. Weights were only calculated for cases where the primary caregiver interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Previous waves (waves 1-4) used a variable selection technique for this stage, and so report the significant covariates. We, like Si and Gelman (2014), use a regularization technique, which doesn't easily produce the "significant" covariates but does protect against overfitting when predicting the propensity to respond. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. We then group cases by the quantiles of the propensity, and then adjust. These propensity scores are multiplied by the original wave 1 weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the population marginal distributions for each of these variables. We estimate the population marginal distribution using the weighted count (wave 1 mother national weights) of these variables at baseline.

Table 2: Baseline national population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	642
Unmarried	450591	2038

Table 3: Baseline national population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	106
Some HS	211988	730
HS or equiv	338497	835
Some College	214319	691
College +	253467	318

Table 4: Baseline national population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	706
Black, Non-hispanic	430161	679
Hispanic	254738	1193

Ethnicity	Population.Count	Wave6.Count
Other	93211	101

Table 5: Baseline national population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	85
18-19	89690	412
20-24	283787	983
25-29	294845	583
30-34	252185	367
35-40	128291	191
40+	29058	59

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 6: Summary of untrimmed survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.568	27.62	101.499	422.13	359.936	18572.24

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming (rather than the Windsorizing technique employed by Si and Gelman (2014)). We are currently researching the benefits of this approach, but we note that it seems particularly beneficial in reducing bias due to specific features of this data. We trim at the 97.5% quantile, and summarize below:

Table 7: Summary of trimmed national primary caregiver survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.708	30.049	110.111	422.13	388.851	7845.822

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

National weights (excluding city X)

There was 3083 cases with both a baseline national (excluding city x) weight and were eligible for the Year 15 wave. Of these cases, 2776 cases were located. Of those who were located, 2415 cases completed the primary caregiver interview. Weights were only calculated for cases where the primary caregiver interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 natx weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the population marginal distributions for each of these variables, where the population is the set of all large cities excluding city X. We estimate the population marginal distribution using the weighted count (wave 1 mother natx weights) of these variables at baseline.

Table 8: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	590
Unmarried	450591	1858

Table 9: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	86
Some HS	211988	671
HS or equiv	338497	776
Some College	214319	631
College +	253467	284

Table 10: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	605
Black, Non-hispanic	430161	620
Hispanic	254738	1129
Other	93211	94

Table 11: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	85
18-19	89690	371
20-24	283787	891
25-29	294845	532
30-34	252185	336
35-40	128291	176
40+	29058	57

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 12: Summary of untrimmed natx weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.684	33.486	120.987	462.136	406.65	18052.13

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We trim at the 97.5% quantile, and summarize below:

Table 13: Summary of trimmed national primary caregiver survey natx weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.831	36.269	131.041	462.136	442.153	7793.659

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

City Weights

There was 4736 cases both had a baseline city weight and were eligible for the Year 15 wave. Of these cases, 4197 cases were located. Of those who were located, 3643 cases completed the primary caregiver interview. Weights were only calculated for cases where the primary caregiver interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 city weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. Unlike the national weights, we rake to the marginal distributions of these four variables within the cities. This meant that we had finer grain cells to adjust to, which (due to the smaller sample size) led to difficulties with convergence. As a compromise, we pooled some levels of age group, ethnicity and education to increase the size of the cells within city.

For data protection reasons, we do not summarize the marginal distributions within city. However in the following tables we compare estimated baseline combined city marginal distributions to the combined city sample distributions in wave 6 for the pooled variables.

Table 14: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	181631	882
Unmarried	165606	2814

Table 15: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade or Some HS	99971.91	1198
HS or equiv	91402.71	976
Some College	81541.68	1000
College +	74321.83	522

Table 16: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	102381.7	805
Black, Non-hispanic	120995.1	1859
Other	123861.3	1032

Table 17: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<20	44953.32	656
20-24	88552.02	1357
25-34	166342.71	1337
34+	47390.08	346

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "citypsu" as the primary sampling unit and "citystratum" as the stratum structure. The distribution of the weights is summarized below:

Table 18: Summary of untrimmed city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.799	15.165	34.026	93.95	69.747	4271.69

We trim at the 97.5% quantile in each city, and summarize below. Unlike the national and national excluding city X, trimming doesn't have large effect on the overall distribution of the weights. This is partly because the weights are trimmed within city. We do not report the distribution of weights for each city due to data security protocol, but note that even looking at the overall weights trimming does reduce some extreme values.

Table 19: Summary of trimmed primary caregiver survey city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.851	15.577	34.57	93.95	70.575	3987.253

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

Child weights

Weights for the cases where the child completed the survey were named with the following convention. Weights are available for all cases where a) a baseline mother weight was available and b) the case was eligible, located and completed the survey.

Table 20: Naming convention for the child wave 6/Year 15 weights

Population	Base.Weight	Replicate.Weight
National	k6natwt	k6natwt_rep1 - k6natwt_rep33
National excl. X	k6natwt	k6natwtx_rep1 - k6natwtx_rep33
City	k6citywt	k6citywt_rep1 - k6citywt_rep10

National weights

There was 3404 cases who both had a baseline national weight and were eligible for the Year 15 wave. Of these cases, 3052 cases were located. Of those who were located, 2494 cases completed the child interview. Weights were only calculated for cases where the child interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the population marginal distributions for each of these variables. We estimate the population marginal distribution using the weighted count (wave 1 mother national weights) of these variables at baseline.

Table 21: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	617
Unmarried	450591	1915

Table 22: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	98
Some HS	211988	680
HS or equiv	338497	787
Some College	214319	662
College +	253467	305

Table 23: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	677
Black, Non-hispanic	430161	636
Hispanic	254738	1123
Other	93211	96

Table 24: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	78

Age.Group	Population.Count	Wave6.Count
18-19	89690	390
20-24	283787	924
25-29	294845	563
30-34	252185	346
35-40	128291	179
40+	29058	52

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 25: Summary of untrimmed national child survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.513	28.609	104.227	446.804	379.139	17059.25

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile, and summarize below:

Table 26: Summary of trimmed national child survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.666	31.418	114.632	446.804	415.819	8133.695

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

National weights (excluding city X)

There was 3083 cases who both had a baseline national (excluding city x) weight and were eligible for the Year 15 wave. Of these cases, 2776 cases were located. Of those who were located, 2284 cases completed the child interview. Weights were only calculated for cases where the child interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 natx weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the

population marginal distributions for each of these variables, where the population is the set of all large cities excluding city X. We estimate the population marginal distribution using the weighted count (wave 1 mother natx weights) of these variables at baseline.

Table 27: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	564
Unmarried	450591	1753

Table 28: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	80
Some HS	211988	628
HS or equiv	338497	733
Some College	214319	604
College +	253467	272

Table 29: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	582
Black, Non-hispanic	430161	579
Hispanic	254738	1066
Other	93211	90

Table 30: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	78
18-19	89690	354
20-24	283787	840
25-29	294845	513
30-34	252185	317
35-40	128291	165
40+	29058	50

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex

survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 31: Summary of untrimmed national (exluding city X) child survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.614	35.076	125.538	488.264	426.874	17381.64

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile, and summarize below:

Table 32: Summary of trimmed national (excluding city X) child survey weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.771	38.482	137.506	488.264	464.668	8305.497

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

City Weights

There was 4736 cases who both had a baseline city weight and were eligible for the Year 15 wave. Of these cases, 4197 cases were located. Of those who were located, 3444 cases completed the child interview. Weights were only calculated for cases where the child interview was completed.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 city weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. Unlike the national weights, we rake to the marginal distributions of these four variables within the cities. This meant that we had finer grain cells to adjust to, which (due to the smaller sample size) led to difficulties with convergence. As a compromise, we pooled some levels of age group, ethnicity and education to increase the size of the cells within city.

For data protection reasons, we do not summarize the marginal distributions within city. However in the following tables we compare estimated baseline combined city marginal distributions to the combined city sample distributions in wave 6 for the pooled variables.

Table 33: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	181631	864
Unmarried	165606	2651

Table 34: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade or Some HS	99971.91	1119
HS or equiv	91402.71	925
Some College	81541.68	949
College +	74321.83	504

Table 35: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	102381.7	762
Black, Non-hispanic	120995.1	1754
Other	123861.3	981

Table 36: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<20	44953.32	622
20-24	88552.02	1283
25-34	166342.71	1274
34+	47390.08	318

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "citypsu" as the primary sampling unit and "citystratum" as the stratum structure. The distribution of the weights is summarized below:

Table 37: Summary of untrimmed child city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.816	16.206	35.612	99.296	72.07	4966.773

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile in each city, and summarize below:

Table 38: Summary of trimmed child survey city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.847	16.461	36.378	99.296	72.462	3931.796

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

home visit weights

Weights for the cases who completed the home visit were named with the following convention. Weights are available for all cases where a) a baseline mother weight was available and b) the case was eligible, selected for the home visit, located and completed the home visit.

Table 39: Naming convention for the home visit wave 6/Year 15 weights

Population	Base.Weight	Replicate.Weight
National	h6natwt	h6natwt_rep1 - h6natwt_rep10
National excl. X	h6natwtx	h6natwtx_rep1 - h6natwtx_rep10
City	h6citywt	h6citywt_rep1 - h6city_rep10

Additional propensity adjustments

The primary caregiver and child weights adjusted for whether the survey administrator could locate the case, and given the case was located, whether the participant chose to respond. The home visit weights have two additional stages to the sampling process that we need to account for.

The survey was fielded by two distinct groups, Westat (n=3620) and CPRC group run by Kathy Neckerman (n=1116). Only cases allocated to be fielded by Westat were allocated to in the home visit sample. The CPRC group fielded cases that had been particularly difficult to contact in the past, so it is likely there were some particular differences between those two groups. We account for this by calculating the propensity of being in the Weststat group using a logistic regression, predicting the propensity to be in this Weststat sampling frame, grouping by quintiles and then adjusting.

Secondly, only a portion of the Westat group were allocated to be part of the home visit sample (n=1533). This allocation was completed before any attempt to locate cases. These individuals were probabilistically selected based upon city, with some cities sampled at a rate of 1 in 2, while others sampled at a rate of 1 in 3.

Adding to this there is an additional complexity where some cases (n=35) were in the Weststat sample, were not allocated to be in the home visit sample, but where still invited to participate in the home visit. After much discussion, it was decided that this was most likely a random process,

so we exclude these cases from the calculation of propensity, but predict the propensity for being sampled and roll these cases into the home visit sample.

National weights

Focusing just on the cases who were included as part of the national sample, 2645 cases were fielded by Westat and while 759 were fielded by the CPRC group. We account for this by calculating the propensity of being in the Weststat group using a regularized logistic regression with membership of the Weststat group is the outcome variable and the covariates listed in Appendix A as the predictors.

Following this, we further adjust the weights by the probability of being selected in a given city, using the selection rate of the sampling company.

Following these two adjustments, we then adjust for the ability to locate the case and the probability of completion given the case was located in a manner similar to the primary caregiver and child weights. Of the 1129 cases allocated to the home visit sample, 1120 were able to be located. Of these cases invited to participate in the home visit, 789 completed the home visit. Together with the 26 who completed the home visit but were not allocated to the home visit sample but still completed a home visit, the total home visit sample is 815.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the population marginal distributions for each of these variables. We estimate the population marginal distribution using the weighted count (wave 1 mother national weights) of these variables at baseline.

Table 40: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	188
Unmarried	450591	627

Table 41: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	33
Some HS	211988	215
HS or equiv	338497	262
Some College	214319	211
College +	253467	94

Table 42: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	243
Black, Non-hispanic	430161	200
Hispanic	254738	336
Other	93211	36

Table 43: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	23
18-19	89690	125
20-24	283787	299
25-29	294845	184
30-34	252185	112
35-40	128291	54
40+	29058	18

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 44: Summary of untrimmed national home visit weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
5.15	66.91	325.83	1388.11	1093.02	39964.4

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile, and summarize below. These weights still have some quite extreme values, but trimming at 95% only reduced the variability marginally so we decided keep the trim at 97.5%.

Table 45: Summary of trimmed national home visit weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
6.155	80.041	383.349	1388.108	1216.254	19686.08

Replicate weights were created in the method described above, and the previously described

checks conducted to test the weights.

National weights (excluding city X)

Focusing on the national survey excluding city X, 2409 cases were fielded by Westat and while 674 were fielded by the CPRC group. We account for this by calculating the propensity of being in the Weststat group using a regularized logistic regression with membership of the Weststat group is the outcome variable and the covariates listed in Appendix A as the predictors.

Following this, we further adjust the weights by the probability of being selected in a given city, using the selection rate of the sampling company.

Following these two adjustments, we then adjust for the ability to locate the case and the probability of completion given the case was located in a manner similar to the primary caregiver and child weights. Of the 1052 cases allocated to the home visit sample, 1044 were able to be located. Of these cases invited to participate in the home visit, 742 completed the home visit. Together with the 26 who completed the home visit but were not allocated to the home visit sample but still completed a home visit, the total home visit sample is 768.

As described above, the first stage of creating the wave 6 weights is to adjust for systematic non-response, both in the likelihood of being located and the probability of responding in the survey. Using two logistic regression models (one for propensity to locate, the other for propensity to respond), we calculate the predicted propensity for each completed case. These propensity scores are multiplied by the original wave 1 natx weights to adjust the current wave.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. The following tables compare the number of completed cases in wave 6 to the population marginal distributions for each of these variables, where the population is the set of all large cities excluding city X. We estimate the population marginal distribution using the weighted count (wave 1 mother natx weights) of these variables at baseline.

Table 46: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	680818	177
Unmarried	450591	591

Table 47: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade	113128	29
Some HS	211988	198
HS or equiv	338497	251
Some College	214319	201
College +	253467	89

Table 48: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	353198	223
Black, Non-hispanic	430161	190
Hispanic	254738	321
Other	93211	34

Table 49: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<18	53450	23
18-19	89690	115
20-24	283787	276
25-29	294845	175
30-34	252185	110
35-40	128291	51
40+	29058	18

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "natpsu" as the primary sampling unit and "natstratum" as the stratum structure. The distribution of the weights is summarized below:

Table 50: Summary of untrimmed national (exlc. city X) home visit weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
4.36	72.85	354.08	1473.06	1233.3	38947.8

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile, and summarize below.

Table 51: Summary of trimmed national (excluding city X) home visit weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
5.202	85.535	415.618	1473.058	1435.578	20075.15

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

City Weights

Focusing on the city weights, 3620 cases were fielded by Westat and while 1116 were fielded by the CPRC group. We account for this by calculating the propensity of being in the Weststat group using a regularized logistic regression with membership of the Weststat group is the outcome variable and the covariates listed in Appendix A as the predictors.

Following this, we further adjust the weights by the probability of being selected in a given city, using the selection rate of the sampling company.

Following these two adjustments, we then adjust for the ability to locate the case and the probability of completion given the case was located in a manner similar to the primary caregiver and child weights. Of the 1533 cases allocated to the home visit sample, 1517 were able to be located. Of these cases invited to participate in the home visit, 1055 completed the home visit. Together with the 35 who completed the home visit but were not allocated to the home visit sample but still completed a home visit, the total home visit sample is 1090.

Following this adjustment, we then rake to four demographic variables identified in the first wave in the survey by Carlson (2008). These four variables are marital status, education, ethnicity and age group. Unlike the national weights, we rake to the marginal distributions of these four variables within the cities. This meant that we had finer grain cells to adjust to, which (due to the smaller sample size) led to difficulties with convergence. As a compromise, we pooled some levels of age group, ethnicity and education to increase the size of the cells within city. Compared to other city weights, we had to pool age to three levels rather than 4. This was due to decreased sample size.

For data protection reasons, we do not summarize the marginal distributions within city. However in the following tables we compare estimated baseline combined city marginal distributions to the combined city sample distributions in wave 6 for the pooled variables.

Table 52: Baseline population and wave 6 marginal counts for marital status

Marriage.Status	Population.Count	Wave6.Count
Married	181631	253
Unmarried	165606	837

Table 53: Baseline population and wave 6 marginal counts for education level

Education.Level	Population.Count	Wave6.Count
<8 grade or Some HS	99971.91	341
HS or equiv	91402.71	294
Some College	81541.68	295
College +	74321.83	160

Table 54: Baseline population and wave 6 marginal counts for ethnicity

Ethnicity	Population.Count	Wave6.Count
White, non-hispanic	102381.7	234
Black, Non-hispanic	120995.1	518
Other	123861.3	338

Table 55: Baseline population and wave 6 marginal counts for age group

Age.Group	Population.Count	Wave6.Count
<26	153031.59	632
26-34	146816.47	360
34+	47390.08	98

Like Si and Gelman (2014) we use the *survey* package in R (Lumley, 2004). We code the complex survey design into the survey object within this package in a similar fashion, using "citypsu" as the primary sampling unit and "citystratum" as the stratum structure. The distribution of the weights is summarized below:

Table 56: Summary of untrimmed home visit city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.816	16.206	35.612	99.296	72.07	4966.773

However, as noted by Si and Gelman (2014), weights calculated in this way typically require trimming due to very extreme weights. We use a slightly different approach to trimming; trimming and then renorming. We trim at the 97.5% quantile in each city, and summarize below:

Table 57: Summary of trimmed home visit city weights

Min	1st Quantile	Median	Mean	3rd Quantile	Max
1.847	16.461	36.378	99.296	72.462	3931.796

Replicate weights were created in the method described above, and the previously described checks conducted to test the weights.

Appendix A

Table 58: Variables included in the propensity model

cm1age	m1e3c
m1a4	m1s4a
m1a9	m1e4b
m1a11a	m1e4c
m1a11b	m1f2
m1a11c	m1f3
m1a11d	m1f4
m1a13	m1f5
m1a13a	m1f6
m1a15	m1f7
m1b2	m1g1
m1b3	m1g2
m1b8	m1g3
m1b27	m1g4
m1b28	m1g6
m1d1a	m1h3
m1s1b	m1h3a
m1d1c	m1i1
m1d1d	m1i2a
m1d1e	m1i3
m1d1f	m1i11
m1d2a	m1j3
m1d2b	m1j4
m1d2c	m1j5
m1d2d	labor
m1d2e	child_support
m1d2f	welf
m1e3a	hosp_type

References

Carlson, Barbara Lepidus. 2008. "Fragile families & child wellbeing study: Methodology for constructing mother, father, and couple weights for core telephone public survey data waves 1-4.".

Friedman, Jerome, Trevor Hastie and Rob Tibshirani. 2010. "Regularization paths for generalized linear models via coordinate descent." *Journal of Statistical Software* 33(1):1.

Goodrich, Ben, Jonah Gabry, Imad Ali and Sam Brilleman. 2018. "rstanarm: Bayesian applied regression modeling via Stan.". R package version 2.17.4.
URL: *http://mc-stan.org/*

Honaker, James, Gary King and Matthew Blackwell. 2011. "Amelia II: A Program for Missing Data." *Journal of Statistical Software* 45(7):1–47.
URL: *http://www.jstatsoft.org/v45/i07/*

Lumley, Thomas. 2004. "Analysis of Complex Survey Samples." *Journal of Statistical Software* 9(1):1–19. R package verson 2.2.

Piironen, Juho, Aki Vehtari et al. 2017. "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics* 11(2):5018–5051.

Si, Yajuan and Andrew Gelman. 2014. "Methodology for Constructing Mother, Father, and Couple Weights for Core Telephone Survey Wave 5.".

Wolter, KM. 1985. "Introduction to variance estimation.".