ITC Mexico Survey
Weights Construction for Waves 1-5

By: **Mary Thompson and Mi Yan**
Date: October 14, 2011

Suggested Citation: ITC Project. (2011, October). ITC Mexico Weights Construction Document Waves 1-5 (2006-2011). University of Waterloo, Waterloo, Ontario, Canada; Instituto Nacional de Salud Publica, Cuernavaca, Mexico; and University of South Carolina, Columbia, South Carolina, United States.

In Wave 1, the sample sizes (numbers of adult smokers) from the original four cities were as follows: N=1079

Mexico City (Distrito Federal):	263
Tijuana (Baja):	273
Guadalajara (Jalisco):	280
Juarez (Chihuahua):	263
Total:	1079

WAVE 1 SAMPLING DESIGN

In the Wave 1 sampling design, twenty AGEBs (census divisions) were selected from each city with probability proportional to number of households according to the 2000 census. Typically two manzanas (block groups) were selected from each selected AGEB, again with probability proportional to the number of households according to the 2000 census. Households in the selected manzanas were listed or mapped. Then the households were ordered randomly and visited in turn. Each household at which contact was made was enumerated, and the smoking status of the adult members determined. If a household contained at least one adult male smoker, the adult male smoker with the next birthday was to be selected to be interviewed. If the household contained at least one adult female smoker, the adult female smoker with the next birthday was to be selected to be interviewed. (In 20 households, it appeared that there was more than one interviewed person of the same gender in a household, in violation of the intended protocol. In 10 of these, the second person appeared to be a duplicate of the first, and was deleted from the data set. In the other 10 households, both respondents, though of the same gender, were kept in the sample, and they were assigned new personIDs in accordance with those assigned in the household enumeration dataset.)

Approximately 7 adult smokers were selected in each manzana.

In Wave 1 the decision was made to sample from both male and female smokers in a selected household, in order to give women smokers a larger chance of being in the sample, and to allow for the possibility of investigating the relationship of responses of members of the same household. The Wave 1 sample contained 426 women and 653 men, and there were 376 individuals (188 households) cases where two members of a household were interviewed.

The Wave 1 sample of smokers is spatially clustered into AGEBs and manzanas; however, since selected manzanas contain on average 30-60 households, depending on city, it is reasonable to consider the smoker sample fairly widely dispersed.

WAVE 1 WEIGHTS

We first constructed a household weight for each enumerated household. (By enumerated household, we mean a household which has been contacted and listed.) Following this we constructed a second household weight for each household with an interviewed individual. Finally, we constructed an individual weight for each individual within his/her household. The product of household weight and individual within-household weight were the final individual weights. They were not calibrated to official estimates of smoker numbers by gender and age group, since such estimates were not available. The weights were rescaled to sum to sample sizes within cities for some pooled analyses.

Computation of enumerated household weights EHWT

Step H1: For each enumerated household, a cluster (manzana) level weight $HW1$ was computed:

$$HW1 = H_{ma} / h_{ema}$$

where H_{ma} is the number of households in the manzana of the household in question, and h_{ema} is the number of households with composition enumerated in that same manzana.

Step H2: For each enumerated household, an AGEB level weight $HW2$ was computed. This is the approximate number of households in the same AGEB represented by the enumerated household.

$$HW2 = H_{AG} \times HW1 / (m_{AG} \times H_{ma}) = H_{AG} / (m_{AG} \times h_{ema})$$

where H_{AG} is the number of households in the AGEB, and m_{AG} is the number of manzanas chosen in the AGEB by probability proportional to size.

Step H3: For each enumerated household, a city level weight $EHWT$ was computed. This is the approximate number of households in the same city represented by the enumerated household.

$$EHWT = H_{city} \times HW2 / (a_{city} \times H_{AG}) = H_{city} / (a_{city} \times m_{AG} \times h_{ema})$$

where $H_{city} =$ number of households in city, $a_{city} =$ number of AGEBs sampled in city.

Prevalence estimates

We were able to use the EHWT weights to estimate the prevalence of smoking in the city, by gender.

For example,

$$\hat{P}_{sm,male} = (\sum_j EHWT_j MALESM_j)/(\sum_j EHWT_j MALE_j)$$

where the sums are over enumerated households j, and $MALE_j$ and $MALESM_j$ are respectively the numbers of male adults and male adult smokers in household *j*.

Computation of interview household weights IHWT

Step H4: For each household in which there is an interview, a city level weight *IHWT* was computed. It is interpreted as the number of smoker households in the city represented by that household. We can think of this as being 0 for any enumerated household without an interview. The *EHWT* values for smoker households without an interview (perhaps because of refusal) are effectively redistributed to households with an interview. For a household with an interview

$$IHWT = EHWT \times \frac{h_{esma}}{h_{ima}},$$

where h_{ima} is the number of households in the manzana with an interview, and h_{esma} is the number of enumerated smoker households in the manzana. (The ratio should be close to 1.)

If we sum *EHWT* over all households in the sample, we should get H_{city}. If we sum *IHWT* over all interview households in the sample, we should get an estimate of the number of smoker households in the city.

Computation of individual weights

Step I1: Each interviewed individual was given a household level weight W1. This is interpreted as the number of people in the same household in the same sampling category represented by the respondent.

In Mexico:

- for an adult male smoker, W1 is the number of adult male smokers in the same household
- for an adult female smoker, W1 is the number of adult female smokers in the same household.

Exceptions: If there were two adult male respondents in the household who appeared to be different people, each was given W1=1/2, and similarly if there were two adult female respondents in the household who appeared to be different people, each was given W1=1/2.

The value of W1 was capped at 2.

Step I2: Each interviewed individual was given a preliminary city level weight W4 which is thought of as the number of people in the same city represented by that individual.
The weight W4 is given by

$$W4 = IHWT \times W1.$$

If we sum W4 over all individuals interviewed, we should get an estimate of the number of smokers in the city.

The final weights W6 were set equal to W4. W6 is variable aDE51915v in the data set.

Rescaling

Finally, the weights in the four cities may be rescaled within each sampling category to sum to city sample sizes, for use in pooled analyses. The rescaled weight is aDE51919v in the data set.

The formula used for each city is as follows:

Rescaled weight $RWT = n_C \times W6 / (\sum_C W6)$,

where n_C is the actual (i.e. unweighted) size of the city subsample, and $\sum_C W6$ denotes a sum over that subsample of the original weights. RWT is renamed to aDE51919v in the final dataset.

WAVE 2 SAMPLING DESIGN

In Wave 2, in each city, efforts were made to recontact and interview all respondents from Wave 1, whether they were still smoking or had quit. On average 70.7% of respondents were retained.

In terms of numbers, 756 respondents were successfully recontacted:

Mexico City (Distrito Federal):	208
Tijuana (Baja):	185
Guadalajara (Jalisco):	220

Juarez (Chihuahua):	143
Total:	756

This means, on average, of the 7 adult smokers from Wave 1 in a manzana, we retained 5 or 6, and lost 1 or 2. Within an AGEB, we therefore on average lost 2 to 4 people, and overall, per city, we lost from 55 to 120 people.

It was expected that if a household member dropped out, that person would not be replaced by someone in the same household. However, in 56 cases, this part of the protocol was violated, and a dropout was apparently replaced by someone in the same household. There were also 5 cases of new recruits in Wave 1 households who were replacements of dropouts in other Wave 1 households. In these cases, in the personIDs of new recruits in Wave 1 households, 1 was replaced by 8 and 2 was replaced by 9; the householdID numbers were not changed.

It was deemed appropriate to replenish the sample lost within each AGEB. The manzanas used at Wave 1 were listed at that time, and not all listed households were contacted and enumerated. It was expected to be possible to replenish the sample from the listed manzanas, though not necessarily from the same ones that had lost respondents.

In cases where the sample loss was sufficiently great that the listed manzanas did not have enough households for the replenishment, it was recommended that another manzana be selected within the AGEB, using probability proportional to size, and then the replenishment would ideally be allocated so that approximately 6 or 7 respondents would come from the new manzana. The new manzana would have to be listed and randomly ordered.

If there were several cases in a city where another manzana would have to be selected within an AGEB, adding AGEBs instead could be considered.

In the replenishment it was again the practice to sample up to 2 in a household, one male and one female adult smoker.

The final sample size is 1045, including both recontact and replenishment respondents:

Mexico City (Distrito Federal):	261
Tijuana (Baja):	271
Guadalajara (Jalisco):	275
Juarez (Chihuahua):	238
Total:	1045

WAVE 2 WEIGHTS

For households and respondents present at Waves 1 and 2 we constructed longitudinal Wave 1-Wave 2 household and individual weights. For all Wave 2 respondents we constructed a cross-sectional weight.

Longitudinal Wave 1 – Wave 2 weights

For the longitudinal weights, we first considered the interviewed household weights IHWT from Wave 1. For those households which were still interview households in Wave 2, we rescaled IHWT to sum to the total of the IHWTs at Wave 1 within each AGEB (psu). This produced for those households a Wave 1-Wave 2 weight labelled IHWT12.

Where it was clear that a dropout was being replaced by someone in the same household, we kept the household only if there was someone else in the household who was interviewed in both waves.

For each Wave 1 respondent still present in Wave 2 we multiplied IHWT12 by the within household weight W1 from Wave 1, producing a preliminary longitudinal weight W12WTT. We then rescaled these W12WTT weights to sum to the Wave 1 cross sectional weight (W1XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W12WT for individuals. W12WT is variable bDE51921v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable bDE51951v on the data set.

Wave 2 cross-sectional weights

Household weights

We first constructed for each interviewed household an interviewed household weight IHWT2. For any interviewed household in a Wave 1 manzana, whether new or old, this is the same as the manzana (common) value of IHWT from Wave 1, multiplied by the number h_{imaW1} of households interviewed in the manzana in Wave 1, divided by the number h_{imaW2} of households interviewed in the manzana in Wave 2; then multiplied by the number m_{AGW1} of manzanas selected in the AGEB in Wave 1, divided by the number m_{AGW2} of manzanas used in the AGEB in Wave 2; then multiplied by the number a_{cityW1} of AGEBs sampled in the city in Wave 1, and divided by the number a_{cityW2} of AGEBs sampled in the city in Wave 2:

$$IHWT2 = \frac{a_{cityW1}}{a_{cityW2}} \frac{m_{AGW1}}{m_{AGW2}} \frac{h_{imaW1}}{h_{imaW2}} IHWT_{man}.$$

Note: We considered creating a new household ID for Wave 2 recruits who were recruited from an old Wave 1 household. However, this did not make much of a difference to the cross-sectional weights, and thus we did not carry out this artificial split.

For an interviewed household in a manzana newly drawn in Wave 2, we proceeded as in Wave 1, recognizing however that the household composition data was available only for interview households.
That is, we let

$$HW1 = H_{ma} / h_{imaW2},$$

where h_{imaW2} is the number of households interviewed in the new manzana, and H_{ma} is the number of households in the manzana. Then we let

$$HW2 = H_{AG} \times HW1 / (m_{AGW2} \times H_{ma}) = H_{AG} / (m_{AGW2} \times h_{imaW2})$$

where H_{AG} is the number of households in the AGEB, and m_{AGW2} is the number of manzanas used in the AGEB in Wave 2. Then we let

$$IHWT2 = H_{city,sm} \times HW2 / (a_{cityW2} \times H_{AG}) = H_{city,sm} / (a_{cityW2} \times m_{AGW2} \times h_{imaW2})$$

where a_{cityW2} is the number of AGEBs sampled in the city in Wave 2, and $H_{city,sm}$ is the estimated number of households in the city, which can be estimated from the sum of the IHWT in Wave 1.

Individual weights

Each newly interviewed individual was given a household level weight W1. This is interpreted as the number of people in the same household in the same sampling category:

- for an adult male smoker, W1 is the number of adult male smokers in the same household
- for an adult female smoker, W1 is the number of adult female smokers in the same household.

For the 27 individuals for whom this information was not available, we set W1=1.

Exception: The value of W1 was capped at 2.

Then each interviewed individual was given a preliminary city level weight W42 which will be thought of as the number of people in the same city represented by that individual.

The weight W4 is given by

$$W42 = IHWT2 \times W1.$$

If we sum W42 over all individuals interviewed, we should get an estimate of the number of smokers in the city.

The final weights W62 were the values of W42. W62 is variable bDE51915v on the data file.

The sums of W62 were checked to see if they were close to the sums of the W6 from Wave 1 in each city. They were seen to be quite close, as in the following table:

City	W6 sum over wave 1	W62 over all the Wave 2 manzanas
Baja	212760.41	206874.84
Chihuahua	199855.36	201005.04
Distrito	1135542.57	1151763.04
Jalisco	297701.47	282271.26

Rescaling

Finally, the weights in the four cities may be rescaled within each sampling category to sum to city sample sizes, for use in pooled analyses.

The formula used for each city is as follows:

Rescaled weight $RWT2 = n_{CW2} \times W62/(\sum_C W62)$,

where n_{CW2} is the actual (i.e. unweighted) size of the Wave 2 city subsample, and $\sum_C W62$ denotes a sum over that subsample of the original weights. The rescaled weight is variable bDE51919v on the file.

WAVE 3 SAMPLING DESIGN

In Wave 3, in each of the original 4 cities, efforts were made to recontact and interview all respondents from Wave 2, whether they were still smoking or had quit. On average 71% of respondents were retained.

In terms of numbers, 742 Wave 2 respondents were successfully recontacted:

Mexico City (Distrito Federal):	205
Tijuana (Baja):	149
Guadalajara (Jalisco):	241
Juarez (Chihuahua):	147

Total:

In addition, 19 Wave 1 households (20 individuals) lost in Wave 2 were recontacted in Wave 3:

Distrito Federal:	3
Baja:	3
Jalisco:	7 (one male + one female)
Chihuahua:	6
Total:	19

There were some issues with linking of the Wave 1-Wave 2 data files with the Wave 3 data files. In many cases gender or age or both did not match. Most of these cases were resolved with reference to the paper data files.

It was again deemed appropriate to replenish the sample lost within each AGEB. The manzanas used at Waves 1 and 2 were considered first.

In cases where the sample loss was sufficiently great that the listed manzanas did not have enough households for the replenishment, it was recommended that another manzana be selected within the AGEB, using probability proportional to size, and then the replenishment would ideally be allocated so that approximately 6 or 7 respondents would come from the new manzana. The new manzana would have to be listed and randomly ordered.

If there were several cases in a city where another manzana would have to be selected within an AGEB, adding one or more AGEBs was the next option.

In the replenishment it was again the practice to sample up to 2 in a household, one male and one female adult smoker.

Beyond replenishment, the sample in Distrito Federale was increased by another 136, selecting 10 new AGEBs, and within them, a total of 21 new manzanas.
The final sample sizes in the original cities:

Mexico City (Distrito Federal):	397
Tijuana (Baja):	252
Guadalajara (Jalisco):	298
Juarez (Chihuahua):	250
Total:	1197

In addition, 813 respondents were newly recruited in 3 other cities, using the sampling design of Wave 1:

Monterrey:	277
Merida:	265
Puebla:	271

Total: 813

Thus the sample size overall was 2010.

WAVE 3 WEIGHTS

For households and respondents present at Waves 1, 2 and 3 we constructed longitudinal Wave 1-Wave 2-Wave 3 household and individual weights. For all Wave 3 respondents we constructed a cross-sectional weight.

Longitudinal Wave 1 – Wave 2 – Wave 3 weights

For the longitudinal weights, we first considered the interviewed household weights IHWT from Wave 1. For those households which were still interview households in Wave 3, we rescaled IHWT to sum to the total of the IHWTs at Wave 1 within each AGEB (psu). This produced for those households a Wave 1-Wave 2-Wave 3 weight labelled internally IHWT123.

For each Wave 1 respondent still present in Wave 3 we multiplied IHWT123 by the within household weight W1 from Wave 1, producing a preliminary longitudinal weight W123WTT. We then rescaled these W123WTT weights to sum to the Wave 1 cross sectional weight (W1XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W123WT for individuals. W123WT is variable cDE51921v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable cDE51951v on the data set.

Wave 3 cross-sectional weights
Household weights

We first constructed for each interviewed household an interviewed household weight IHWT3. For any interviewed household in a Wave 2 manzana, whether the household is new or old, this is the same as the manzana (common) value of IHWT2 from Wave 2, multiplied by the number h_{imaW2} of households interviewed in the manzana in Wave 2, divided by the number h_{imaW3} of households interviewed in the manzana in Wave 3; then multiplied by the number m_{AGW2} of manzanas used in the AGEB in Wave 2, divided by the number m_{AGW3} of manzanas used in the AGEB in Wave 3; then multiplied by the number a_{cityW2} of AGEBs used in the city in Wave 2, and divided by the number a_{cityW3} of AGEBs used in the city in Wave 3:

$$IHWT3 = \frac{a_{cityW2}}{a_{cityW3}} \frac{m_{AGW2}}{m_{AGW3}} \frac{h_{imaW2}}{h_{imaW3}} IHWT2_{man}.$$

For an interviewed household in a manzana newly drawn in Wave 3 (in old cities or new), we proceeded as in Wave 1.

Step H1: For each enumerated household, a cluster (manzana) level weight $HW1$ was computed:

$$HW1 = H_{ma} / h_{ema}$$

where H_{ma} is the number of households in the manzana of the household in question, and h_{ema} is the number of households with composition enumerated in that same manzana.

Step H2: For each enumerated household, an AGEB level weight $HW2$ was computed. This is the approximate number of households in the same AGEB represented by the enumerated household.

$$HW2 = H_{AG} \times HW1/(m_{AGW3} \times H_{ma}) = H_{AG} /(m_{AGW3} \times h_{ema})$$

where H_{AG} is the number of households in the AGEB, and m_{AGW3} is the number of manzanas used in the AGEB in Wave 3.

Step H3: For each enumerated household, a city level weight $EHWT$ was computed. This is the approximate number of households in the same city represented by the enumerated household.

$$EHWT = H_{city} \times HW2/(a_{cityW3} \times H_{AG}) = H_{city} /(a_{cityW3} \times m_{AGW3} \times h_{ema})$$

where $H_{city} =$ number of households in city, a_{cityW3} = number of AGEBs used in the city in Wave 3.

Step H4: For each household in which there was an interview, a city level weight $IHWT3$ was computed. It is interpreted as the number of smoker households in the city represented by that household. We can think of this as being 0 for any enumerated household without an interview. For a household with an interview

$$IHWT3 = EHWT \times \frac{h_{esma}}{h_{ima}},$$

where h_{ima} is the number of households in the manzana with an interview, and h_{esma} is the number of enumerated smoker households in the manzana. (The ratio should be close to 1.)

For the new cities, if we sum *EHWT* over all households in the sample, we should get H_{city}.   If we sum *IHWT3* over all interview households in the sample, we should get an estimate of the number of smoker households in the city.

Individual weights

Each newly interviewed individual was given a household level weight W1.   This is interpreted as the number of people in the same household in the same sampling category:

- for an adult male smoker, W1 is the number of adult male smokers in the same household
- for an adult female smoker, W1 is the number of adult female smokers in the same household.

For recontact individuals, the value of W1 was carried over from Wave 2, when available. For the 73 recontact individuals and 88 replenishment individuals for whom the required information was not available, we set W1=1. There was one male smoker in a household for whom 0 male smokers had been recorded, and for that person W1 was also set as 1 for this calculation.

Exception: The value of W1 was capped at 2.

Then each interviewed individual was given a preliminary city level weight W43 which will be thought of as the number of people in the same city represented by that individual. The weight W4 is given by

$$W43 = IHWT3 \times W1.$$

If we sum W43 over all individuals interviewed, we should get an estimate of the number of smokers in the city.

The final weights W63 were the values of W43. W63 is variable cDE51915v on the data file.

In each Wave 1 city the sums of W63 were checked to see if they were close to the sums of the W6 from Wave 1 and of the W62 from Wave 2.   They were seen to be quite close, as in the following table:

City	W6 summed over Wave 1	W62 summed over all the Wave 2 manzanas	W63 summed over all the Wave 3 manzanas
Baja	212760.41	206874.84	259112.51
Chihuahua	199855.36	201005.04	229892.86
Distrito Federal	1135542.57	1151763.04	1439981.73
Jalisco	297701.47	282271.26	290918.01

Rescaling

Finally, the weights in the seven cities may be rescaled within each sampling category to sum to city sample sizes, for use in pooled analyses.

The formula used for each city is as follows:

Rescaled weight $RWT3 = n_{CW3} \times W63 / (\sum_C W63)$,

where n_{CW3} is the actual (i.e. unweighted) size of the Wave 3 city subsample, and $\sum_C W63$ denotes a sum over that subsample of the original weights. The rescaled weight is variable cDE51919v on the file.

WAVE 4 SAMPLING DESIGN

In Wave 4, in 6 of the 7 cities, all except Juarez, which was dropped for safety reasons, efforts were made to recontact and interview all respondents from Wave 3, whether they were still smoking or had quit. On average 1304/2010=64.9% of respondents were retained. (In addition, 5 of the Wave 3 replenishment respondents had been excluded from Wave 3 due to missing age. They were recontacted in Wave 4. These respondents received the recontact survey in Wave 4, but were not given Wave 3-4 longitudinal weights, and were treated as Wave 4 replenishment respondents for purposes of weight construction.)

In terms of numbers, 1304 Wave 3 respondents were successfully recontacted:

Mexico City (Distrito Federal):	310
Tijuana (Baja):	171
Guadalajara (Jalisco):	268
Juarez (Chihuahua):	0
Monterrey:	204
Puebla:	175
Merida:	176
Total:	1304

No households recruited in Wave 1 or Wave 2 but lost in Wave 3 were recontacted in Wave 4.

There were some issues with linking of the Wave 1-Wave 3 data files with the Wave 4 data files. In 10 cases gender or age or both did not match. These were cases deemed by the ITC Mexico team to have been recorded incorrectly at Wave 3. Thus the Wave 3 data set was revised accordingly.

It was again deemed appropriate to replenish the sample lost within each AGEB.

In cases where the sample loss was sufficiently great that the already listed manzanas did not have enough households for the replenishment, it was recommended that another manzana be selected within the AGEB, using probability proportional to size, and then the replenishment would ideally be allocated so that approximately 6 or 7 respondents would come from the new manzana. The new manzana would have to be listed and randomly ordered. In Mexico City, all replenishment respondents were from new manzanas.

If there were several cases in a city where another manzana would have to be selected within an AGEB, adding one or more AGEBs was the next option. This turned out not to be necessary in Wave 4.

In the replenishment it was again the practice to sample up to 2 in a household, one male and one female adult smoker.

Beyond replenishment, the sample in Distrito Federale was increased by another 32 (in Wave 3 N_$_{DF}$ =397; in Wave 4 N_$_{DF}$=429). For replenishment and this additional sample, no new AGEBS, but a total of 19 new manzanas, with 116 respondents, were selected.

Juarez was replaced by a new city, Léon.

The final sample sizes in the cities:

Mexico City (Distrito Federal):	429
Tijuana (Baja):	294
Guadalajara (Jalisco):	282
Juarez (Chihuahua):	0
Monterrey:	278
Merida:	278
Puebla:	279
Léon:	288
Total:	2128

WAVE 4 WEIGHTS

For households and respondents present at Waves 1, 2, 3 and 4 we constructed longitudinal Wave 1-Wave 2-Wave 3-Wave 4 household and individual weights. Because of the design change at Wave 3, for all respondents present at Waves 3 and 4, we constructed longitudinal Wave 3-Wave 4 household and individual weights.   For all Wave 4 respondents we constructed a cross-sectional weight.

Longitudinal Wave 1 – Wave 2 – Wave 3 – Wave 4 weights

For the longitudinal weights, we first considered the interviewed household weights IHWT from Wave 1. For those households which were still interview households in Wave 4, we rescaled IHWT to sum to the total of the IHWTs at Wave 1 within each AGEB (psu).   This produced for those households a Wave 1-Wave 2-Wave 3-Wave 4 weight labelled internally IHWT1234.

For each Wave 1 respondent still present in Wave 4 we multiplied IHWT1234 by the within household weight W1 from Wave 1, producing a preliminary longitudinal weight W1234WTT. We then rescaled these W1234WTT weights to sum to the Wave 1 cross sectional weight (W1XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W1234WT for individuals. W1234WT is variable dDE51921v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable dDE51951v on the data set.

Longitudinal Wave 3 – Wave 4 weights

For the Wave 3 – Wave 4 longitudinal weights, we first considered the interviewed household weights IHWT3 from Wave 3. For those households which were still interview households in Wave 4, we rescaled IHWT3 to sum to the total of the IHWT3 at Wave 3 within each AGEB (psu).   This produced for those households a Wave 3-Wave 4 weight labelled internally IHWT34.

For each Wave 3 respondent still present in Wave 4 we multiplied IHWT34 by the within household weight W1 from Wave 3, producing a preliminary longitudinal weight W34WTT. We then rescaled these W34WTT weights to sum to the Wave 3 cross sectional weight (W3XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W34WT for individuals. W34WT is variable dDE51925v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable dDE51955v on the data set.

Wave 4 cross-sectional weights
Household weights

We first constructed for each interviewed household an interviewed household weight IHWT4. For any interviewed household in a Wave 3 manzana, whether the household is new or old, this is the same as the manzana (common) value of IHWT3 from Wave 3, multiplied by the number h_{imaW3} of households interviewed in the manzana in Wave 3, divided by the number h_{imaW4} of households interviewed in the manzana in Wave 4; then multiplied by the number m_{AGW3} of manzanas used in the AGEB in Wave 3, divided by the number m_{AGW4} of manzanas used in the AGEB in Wave 4; then multiplied by the number a_{cityW3} of AGEBs used in the city in Wave 3, and divided by the number a_{cityW4} of AGEBs used in the city in Wave 4:

$$IHWT4 = \frac{a_{cityW3}}{a_{cityW4}} \frac{m_{AGW3}}{m_{AGW4}} \frac{h_{imaW3}}{h_{imaW4}} IHWT3_{man}.$$

For an interviewed household in a manzana newly drawn in Wave 4, we proceeded as in Wave 1, as follows, with the exception that in Mexico City, h_{ema} was replaced by 2.5* h_{ima} for all new manzanas. This exception was made because it appeared that the protocol for enumerating households whether or not they contained smokers had not been followed in many new manzanas; 2.5 was the approximate ratio of enumerated households to interview households in Wave 3 replenishment in Mexico City.

Step H1: For each enumerated household, a cluster (manzana) level weight $HW1$ was computed:

$$HW1 = H_{ma} / h_{ema}$$

where H_{ma} is the number of households in the manzana of the household in question, and h_{ema} is the number of households with composition enumerated in that same manzana.

Step H2: For each enumerated household, an AGEB level weight $HW2$ was computed. This is the approximate number of households in the same AGEB represented by the enumerated household.

$$HW2 = H_{AG} \times HW1/(m_{AGW4} \times H_{ma}) = H_{AG} /(m_{AGW4} \times h_{ema})$$

where H_{AG} is the number of households in the AGEB, and m_{AGW4} is the number of manzanas used in the AGEB in Wave 4.

Step H3: For each enumerated household, a city level weight $EHWT$ was computed. This is the approximate number of households in the same city represented by the enumerated household.

$$EHWT = H_{city} \times HW2/(a_{cityW4} \times H_{AG}) = H_{city} /(a_{cityW4} \times m_{AGW4} \times h_{ema})$$

where $H_{city} =$ number of households in city, a_{cityW4} = number of AGEBs used in the city in Wave 4.

Step H4: For each household in which there was an interview, a city level weight *IHWT*4 was computed. It is interpreted as the number of smoker households in the city represented by that household. We can think of this as being 0 for any enumerated household without an interview. For a household with an interview

$$IHWT4 = EHWT \times \frac{h_{esma}}{h_{ima}},$$

where h_{ima} is the number of households in the manzana with an interview, and h_{esma} is the number of enumerated smoker households in the manzana. (The ratio should be close to 1.)

Individual weights

Each newly interviewed individual was given a household level weight W1. This is interpreted as the number of people in the same household in the same sampling category:

- for an adult male smoker, W1 is the number of adult male smokers in the same household
- for an adult female smoker, W1 is the number of adult female smokers in the same household.

Exception: The value of W1 was capped at 2.

Then each interviewed individual was given a preliminary city level weight W44 which will be thought of as the number of people in the same city represented by that individual. The weight W44 is given by

$$W44 = IHWT4 \times W1.$$

If we sum W44 over all individuals interviewed, we should get an estimate of the number of smokers in the city.

The final weights W64 were the values of W44. W64 is variable dDE51915v on the data file.

In each city the sums of W64 were checked to see if they were close to the sums of the W6 from Wave 1, the W62 from Wave 2, and the W63 from Wave 3. They were seen to be reasonably close, as in the following table:

City	W6 summed over wave 1	W62 summed over all the Wave 2	W63 summed over all the	W64 summed over all the Wave

		manzanas	Wave 3 manzanas	4 manzanas
Baja	212760.41	206874.84	259112.51156	276421.93
Chihuahua	199855.36	201005.04	229892.86129	0
Distrito Federal	1135542.57	1151763.04	1439981.7339	1402824.99
Jalisco	297701.47	282271.26	290918.00954	292474.91
Monterrey			176229.93	184586.18
Merida			55478.36	53922.91
Puebla			211348.21	207257.22
Léon				166172.79

Rescaling

Finally, the weights in the seven cities may be rescaled within each sampling category to sum to city sample sizes, for use in pooled analyses.

The formula used for each city is as follows:

Rescaled weight $RWT4 = n_{CW4} \times W64 / (\sum_{C} W64)$,

where n_{CW4} is the actual (i.e. unweighted) size of the Wave 4 city subsample, and $\sum_{C} W64$ denotes a sum over that subsample of the original weights. The rescaled weight is variable dDE51919v on the file.

WAVE 5 SAMPLING DESIGN

In Wave 5, in all 7 cities, efforts were made to recontact and interview all respondents from Wave 4, whether they were still smoking or had quit. On average 1762/2128=82.8% of respondents were retained.

In terms of numbers, 1762 Wave 4 respondents were successfully recontacted:
Mexico City (Distrito Federal) 09: 357
Tijuana (Baja) 02: 245
Guadalajara (Jalisco) 14: 280
Monterrey 19: 216
Puebla 21: 204
Merida 31: 215
Léon 11: 245
Total: 1762

No households recruited in Wave 1 or Wave 2 or Wave 3 but lost in Wave 4 were recontacted in Wave 5.

There were few issues with linking of the Wave 1-Wave 4 data files with the Wave 5 data files. All the re-contact smokers could be linked; however, some of the respondents provided a different Birth Year than in the last wave.

It was again deemed appropriate to replenish the sample lost within each AGEB.

In cases where the sample loss was sufficiently great that the already listed manzanas did not have enough households for the replenishment, it was recommended that another manzana be selected within the AGEB, using probability proportional to size, and then the replenishment would ideally be allocated so that approximately 6 or 7 respondents would come from the new manzana. The new manzana would have to be listed and randomly ordered.

In terms of numbers, 75 new Manzanas were added in Wave 5:
Mexico City (Distrito Federal) 09: 15
Tijuana (Baja) 02: 8
Guadalajara (Jalisco) 14: 0
Monterrey 19: 11
Puebla 21: 16
Merida 31: 18
Léon 11: 7
Total: 75

If there were several cases in a city where another manzana would have to be selected within an AGEB, adding one or more AGEBs was the next option. This was not required in Wave 5.

In the replenishment it was again the practice to sample up to 2 in a household, one male and one female adult smoker. In Wave 5 replenishment, 305 households have two respondents sampled. Within the 305 households, there are 4 households having two respondents with the same gender.

The final sample sizes in the cities:

Mexico City (Distrito Federal): 433
Tijuana (Baja): 294
Guadalajara (Jalisco): 280
Monterrey:278
Puebla: 280
Merida:280
Léon: 288
Total: 2133

WAVE 5 WEIGHTS

For households and respondents present at Waves 1, 2, 3, 4 and 5 we constructed longitudinal Wave 1-Wave 2-Wave 3-Wave 4-Wave 5 household and individual weights. Because of the design change at Wave 3, for all respondents present at Waves 3, 4 and 5, we constructed longitudinal Wave 3-Wave 4-Wave 5 household and individual weights. Because of the design change at Wave 4, for all respondents present at Waves 4 and 5, we constructed longitudinal Wave 4-Wave 5 weights. For all Wave 5 respondents we constructed a cross-sectional weight.

Longitudinal Wave 1 – Wave 2 – Wave 3 – Wave 4 – Wave 5 weights

For the longitudinal weights, we first considered the interviewed household weights IHWT from Wave 1. For those households which were still interview households in Wave 5, we rescaled IHWT to sum to the total of the IHWTs at Wave 1 within each AGEB (psu). This produced for those households a Wave 1-Wave 2-Wave 3-Wave 4-Wave 5 weight labelled internally IHWT12345.

For each Wave 1 respondent still present in Wave 4 (N=388) we multiplied IHWT12345 by the within household weight W1 from Wave 1, producing a preliminary longitudinal weight W12345WTT. We then rescaled these W12345WTT weights to sum to the Wave 1 cross sectional weight (W1XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W12345WT for individuals. W12345WT is variable eDE51921v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable eDE51951v on the data set.

Longitudinal Wave 3 – Wave 4 – Wave 5 weights

For the Wave 3 – Wave 4 – Wave 5 longitudinal weights, we first considered the interviewed household weights IHWT3 from Wave 3. For those households which were still interview households in Wave 5, we rescaled IHWT3 to sum to the total of the IHWT3 at Wave 3 within each AGEB (psu). This produced for those households a Wave 3-Wave 4 – Wave 5 weight labelled internally IHWT345.

For each Wave 3 respondent still present in Wave 5 (N=1105) we multiplied IHWT345 by the within household weight W1 from Wave 3, producing a preliminary longitudinal weight W345WTT. We then rescaled these W345WTT weights to sum to the Wave 3 cross sectional weight (W3XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W345WT for individuals. W345WT is variable eDE51925v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable eDE51955v on the data set.

Longitudinal Wave 4 – Wave 5 weights

For the Wave 4 – Wave 5 longitudinal weights, we first considered the interviewed household weights IHWT4 from Wave 4. For those households which were still interview households in Wave 5, we rescaled IHWT4 to sum to the total of the IHWT4 at Wave 4 within each AGEB (psu). This produced for those households a Wave 4 – Wave 5 weight labelled internally IHWT45.

For each Wave 4 respondent still present in Wave 5 (N=1762) we multiplied IHWT45 by the within household weight W1 from Wave 4, producing a preliminary longitudinal weight W45WTT. We then rescaled these W45WTT weights to sum to the Wave 4 cross sectional weight (W4XWT) totals for age group (18-24, 25-44, 45-54, 55+) and gender within cities. This produced the longitudinal weights W45WT for individuals. W45WT is variable eDE51927v on the data set.

There is also a version of these rescaled to sum to sample size within cities. This is variable eDE51957v on the data set.

Wave 5 cross-sectional weights

Household weights

We first constructed for each interviewed household an interviewed household weight IHWT5. For any interviewed household in a Wave 4 manzana, whether the household is new or old, this is the same as the manzana (common) value of IHWT4 from Wave 4, multiplied by the number h_{imaW4} of households interviewed in the manzana in Wave 4, divided by the number h_{imaW5} of households interviewed in the manzana in Wave 5; then multiplied by the number m_{AGW4} of manzanas used in the AGEB in Wave 4, divided by the number m_{AGW5} of manzanas used in the AGEB in Wave 5; then multiplied by the number a_{cityW4} of AGEBs used in the city in Wave 4, and divided by the number a_{cityW5} of AGEBs used in the city in Wave 5:

$$IHWT5 = \frac{a_{cityW4}}{a_{cityW5}} \frac{m_{AGW4}}{m_{AGW5}} \frac{h_{imaW4}}{h_{imaW5}} IHWT4_{man}.$$

For an interviewed household in a manzana newly drawn in Wave 5, we proceeded as in Wave 1, as follows, with the exception that in Mexico City, h_{ema} was replaced by 2.5* h_{ima} for all new manzanas. This exception carried forward an exception made in Wave 4. This time the exception was made because 2.5 was the approximate overall ratio of enumerated households to interview households in Wave 3 and in Wave 5 replenishment in Mexico City, while the actual ratios h_{ema} / h_{ima} were highly variable from manzana to manzana.

Step H1: For each enumerated household, a cluster (manzana) level weight $HW1$ was computed:

$$HW1 = H_{ma} / h_{ema}$$

where H_{ma} is the number of households in the manzana of the household in question, and h_{ema} is the number of households with composition enumerated in that same manzana.

Step H2: For each enumerated household, an AGEB level weight $HW2$ was computed. This is the approximate number of households in the same AGEB represented by the enumerated household.

$$HW2 = H_{AG} \times HW1/(m_{AGW5} \times H_{ma}) = H_{AG} /(m_{AGW5} \times h_{ema})$$

where H_{AG} is the number of households in the AGEB, and m_{AGW5} is the number of manzanas used in the AGEB in Wave 5.

Step H3: For each enumerated household, a city level weight $EHWT$ was computed. This is the approximate number of households in the same city represented by the enumerated household.

$$EHWT = H_{city} \times HW2/(a_{cityW5} \times H_{AG}) = H_{city} /(a_{cityW5} \times m_{AGW5} \times h_{ema})$$

where $H_{city} =$ number of households in city, a_{cityW5} = number of AGEBs used in the city in Wave 5.

Step H4: For each household in which there was an interview, a city level weight *IHWT*5 was computed. It is interpreted as the number of smoker households in the city represented by that household. We can think of this as being 0 for any enumerated household without an interview. For a household with an interview

$$IHWT5 = EHWT \times \frac{h_{esma}}{h_{ima}},$$

where h_{ima} is the number of households in the manzana with an interview, and h_{esma} is the number of enumerated smoker households in the manzana. (The ratio should be close to 1.)

Individual weights

Each newly interviewed individual was given a household level weight W1. This is interpreted as the number of people in the same household in the same sampling category:

- for an adult male smoker, W1 is the number of adult male smokers in the same household
- for an adult female smoker, W1 is the number of adult female smokers in the same household.

Exception: The value of W1 was capped at 2.

Then each interviewed individual was given a preliminary city level weight W45 which will be thought of as the number of people in the same city represented by that individual. The weight W45 is given by

$$W45 = IHWT5 \times W1.$$

If we sum W45 over all individuals interviewed, we should get an estimate of the number of smokers in the city.

The final (inflation) weights W65 were the values of W45. W65 is variable eDE51915v on the data file.

In each city the sums of W65 were checked to see if they were close to the sums of the W6 from Wave 1, the W62 from Wave 2, the W63 from Wave 3, and the W64 from Wave 4. They were seen to be acceptably close, as in the following table:

City	W6 summed over wave 1 manzanas	W62 summed over all the Wave 2 manzanas	W63 summed over all the Wave 3 manzanas	W64 summed over all the Wave 4 manzanas	W65 summed over all the Wave 5 manzanas
Baja	212760	206874	259112	276421	279560
Chihuahua	199855	201005	229892	0	0
Distrito Federal	1135542	1151763	1439981	1402824	1317885
Jalisco	297701	282271	290918	292474	291180
Monterrey			176229	184586	173027
Merida			55478	53922	64056
Puebla			211348	207257	204873
Léon				166172	155431

Rescaling

Finally, the weights in the seven cities may be rescaled within each sampling category to sum to city sample sizes, for use in pooled analyses.

The formula used for each city is as follows:

Rescaled weight $RWT5 = n_{CW5} \times W65 / (\sum_C W65),$

where n_{CW5} *is the actual (i.e. unweighted) size of the Wave 5 city subsample, and* $\sum_C W65$ *denotes a sum over that subsample of the original weights. The rescaled weight is variable eDE51919v on the file.*