ITC Mexico (MEX) Project
New Cohort Survey Weights

User's Guide

December 17, 2025
With Guidance from Mary Thompson and Yingchen Fan

Background

Sampling design and protocol for Waves 1 through 8

The data were collected over 8 waves, approximately 4 months apart, beginning in November 2018 and ending in March 2021.

At Wave 1, efforts were made to obtain responses from approximately 900 cigarette smokers who were not using NVPs and approximately 600 cigarette smokers who were using NVPs.

Smoking Status

A respondent was classified as an exclusive smoker (**smk_status=1**), if

fr51316=1 past 30-day smoked cigarettes coded as 1=Yes

and

nc51305=2 past 30-day use of e-cig/vaping device coded as 2=No

A respondent was classified as a dual user of cigarettes and NVPs (**smk_status=2**), if

fr51316=1 past 30-day smoked cigarettes coded as 1=Yes

and

nc51305=1 past 30-day use of e-cigarette or vaping device coded as 1=Yes

Otherwise, we set **smk_status=7** implying that the respondent was not eligible at recruitment.

At later waves, efforts were made to recontact all respondents from previous waves. New respondents were added to replace dropouts and to keep the numbers having **smk_status=1** and **smk_status=2** approximately the same for each wave. Those who had quit smoking were retained, and respondents who were not smoking at a wave were given **smk_status=7** for that wave.

It was possible for a respondent to skip one or more wave and rejoin at a later wave.

Upon reviewing data, there were cases for which the same respondent ID was recorded with age and/or sex discrepancies across waves. These cases were dealt with individually in the following ways:

- age and/or sex being "corrected" if the discrepancy was determined to be an error
- the respondent being split into two respondents if it was determined that the original respondent had been replaced by another legitimate respondent
- the respondent being dropped if a reasonable explanation could not be determined.

Benchmark survey used for pseudo-weights construction

The calibration figures for the pseudo-weights are from the *2018 National Survey on Nutrition and Health (ENSANUT)* in Mexico. These figures are in terms of estimated population proportions (rather than estimated population numbers) of exclusive smokers and dual users in 24 cells defined by sex (2 levels) crossed with age category (4 levels) crossed with education level (3 levels). Without estimated population numbers it was not possible to compute calibrated inflation weights in the usual sense, and the calibrated weights for our survey are given as rescaled weights. That is, what would be obtained if each set of calibrated inflation weights were rescaled to sum to the corresponding wave sample size.

The two categories used for sex were:
Category 1: Female
Category 2: Male

The four categories used for age were:
Category 1: 18 – 29
Category 2: 30 – 39
Category 3: 40 – 49
Category 4: 50 – 65

The three categories used initially for education were:
Category 1: levels 1 and 2 (none or primary)
Category 2: level 3 (secondary)
Category 3: levels 4 – 8 (some post-secondary)

At the conclusion of data collection, there were relatively few smokers in education category 1 and relatively few dual users in education categories 1 and 2 combined. Although respondents in all education categories were kept in the sample, for weight calibration purposes, the exclusive smokers in education category 1 were taken to belong to category 2, and the dual users in education categories 1 and 2 were taken to belong to category 3.

Cross-sectional weights for Waves 1 through 8, all smokers

For each wave, the sample of smokers (exclusive cigarette smokers or dual users) was given cross-sectional (pseudo-) weights, making them representative in terms of smoking status crossed with sex crossed with age category crossed with their attributed education category. The cross-sectional weights at each wave are rescaled to sum to sample size; that is, to have an average value 1.

The formula used for a Wave *t* respondent with smoking status ss, sex sx, age category ac and attributed education category ed *before rescaling* was:

Weight(t, ss, sx, ac, ed) = P(ss, sx, ac, ed)/n(t, ss, sx, ac, ed)

where P(ss, sx, ag, ed) is the corresponding population proportion from the *2018 ENSANUT* and n(t, ss, sx, ag, ed) is the corresponding sample size from Wave *t*.

The wave *t* cross-sectional weight is zero for all respondents of **smk_status = 7** in Wave *t*.

WTS51101v is the variable name of the rescaled cross-sectional weights. An initial letter (**a** to **h**) is used as a prefix to designate Waves 1 to 8.

Cross-sectional weights for exclusive smokers and dual users separately

We have also created rescaled cross-sectional weights (rescaled to sum to sample size) for exclusive smokers and for dual users separately.

Exclusive smokers

wts51201v is the variable name of the cross-sectional weights for exclusive smokers. It is equal to zero for all dual users.

Dual Users

wts51401v is the variable name of the cross-sectional weights for dual users. It is equal to zero for all exclusive smokers.

GEE weights and their use

A GEE weight has been provided for repeated measures analyses (using Generalized Estimating Equations) that take into account that the data are longitudinal and that an individual's responses will tend to be associated across waves if their situation has not undergone a substantial change. The GEE weight is constant for each respondent over all waves beginning with the first in which that respondent is eligible (as usually is the case at recruitment). It is the cross-sectional weight at first eligibility, rescaled to sum to sample size among all others of first eligibility at the same wave. For respondents recruited and eligible at Wave 1, it is their Wave 1 rescaled weight.

WTS51971v is the variable name for GEE weights in the dataset.

Two-wave longitudinal weights

Longitudinal weights for Wave *t* to Wave *t+1* have been provided for *t* =1, …., 7. Such a weight is positive for a respondent who is present in both Wave *t* and Wave *t+1* and is smoking in Wave *t*.   The longitudinal weights for Wave *t* to Wave *t+1* compensate for dropout between Wave *t* and Wave *t+1*. The weights of those Wave *t* respondents who have remained in the sample at *t+1* are raised to sum to Wave *t* sample size within each cell defined by sex crossed with age category crossed with education.

The variable names of the longitudinal Wave *t* to Wave *t+1* weights in the data set are **bWTS51951v**, **cWTS51953v**, **dWTS51955v**, **eWTS51957v**, **fWTS51959v**, **gWTS51961v**, and **hWTS51963v**.

The initial letters **b** to **h** designate the values of *t+1*, taking values 2 to 8.

Advice on the use of the weights

(a) If you want to calculate descriptive quantities such as a mean value or a prevalence estimation for the population of smokers (regardless of whether they use e-cigarettes) *in a single wave*

Examples:

- To calculate the proportion of all smokers at the time of Wave 6 who smoke manufactured cigarettes

  You should use cross-sectional weights (**WTS51101v**) from Wave 6

- To calculate the proportion of all smokers at the time of Wave 1, working outside the home, whose workplace has a smoking ban

  You should use cross-sectional weights (**WTS51101v**) from Wave 1

(b) If you want to calculate descriptive quantities such as a mean value or prevalence estimation for the population of smokers (regardless of whether they use e-cigarettes) *in more than one wave*

Example:

- To calculate the proportion of all smokers at the time of Waves 6 and 7 who purchase at a pharmacy

  It is important to account for the fact that many of the respondents will have been present in both waves, and would be responding twice.

  If the responses of Wave 6 respondents at Wave 7 are left out, essentially being replaced by those of new recruits at Wave 7, there are no specific weights to be applied in this context. One solution is to use the cross-sectional weights for the Wave 6 respondents and the GEE weights **WTS51971v** for the new recruits in Wave 7.

  Alternatively, if you are using Stata with svyset, you can use both responses of the Wave 6 respondents who stay in, and fit an intercept-only model to the Wave 6 and Wave 7 responses with xtgee, specifying the GEE weight **WTS51971v** as the pweight (in svyset). The standard errors of GEE would account for the association between the responses of Wave 6 respondents who stay in (there is also a slight adjustment by the GEE estimation to the regression point estimates).

(c) If you have analytical uses such as regression or logistic regression

Example: logistic regression of support for smoke-free law on socio-demographic variables and daily vs non-daily smoking

In analytical uses it is strongly recommended that the weight-determining variables be included in model either as explanatory variables or as covariates. (The weight-determining variables are smoking status, i.e. whether exclusive or dual; sex; age category; education.) This helps to bring the results of the analysis closer to what they would be without weights, while the weights continue to compensate for possible biasing influence of the sampling design. (Skinner and Mason, 2011)

Where the analysis involves only exclusive smokers or only dual users, the *cross-sectional weights* for the appropriate category are appropriate; **wts51201v** for smokers and **wts51401v** for dual users.

When the analysis involves both exclusive smokers and dual users, the cross-sectional weights for all smokers would theoretically be appropriate. However, in the design of the survey, dual users were heavily oversampled due to their importance for the research questions, and their cross-sectional weights from all smokers (**WTS51101v**) are much smaller than those of the exclusive smokers.

If the subset of smokers relevant to the regression/logistic regression is small, using cross sectional weights from all smokers can cause instability in coefficient or odds ratio estimates, in particular an odds ratio for smoking status as an explanatory variable. In such a case, specifying a new weight **specialwt** that is the maximum of the two smoking category-specific weights **wts51201v** and **wts51401v** should give more stable results.

```
gen specialwt = max(wts51201v wts51401v)
```

However, such a weight should not be used in estimation of prevalence or in analyses where smoking status is not an explanatory variable.

If you are estimating a GEE model, note that Stata requires that the weight be constant across all observations for a particular id. Thus, you must use the **WTS51971v** weight in such analyses. Mixed-effects models should use cross-sectional weights as appropriate; either **WTS51101v** for an analysis with all smokers and **wts51201v** for smokers and **wts51401v** for dual users.

More information may be found here:

> Skinner, C. and Mason, B. (2012) Weighting in the regression analysis of survey data with a cross-national application. *The Canadian Journal of Statistics* 40, 697-711.