

**National Cancer Institute  
Tobacco Use Supplement to the Current Population Survey (TUS-CPS)  
2021 Data User Webinar - Using SAS to Analyze TUS-CPS Data**

**Webinar Transcript**

MS. NALINI CORCY: Hello, everyone. I think we are ready to get started. Again, welcome and thank you for joining us today for the National Cancer Institute data user webinar. This is the second webinar in our webinar series. And we are excited to have you here with us. The topic today is “Using SAS to Analyze the TUS-CPS Data.”

I just want to go over a couple of housekeeping items before we get started with the presentation today. So, first of all, please note that all participants are on mute. At any time during the webinar, feel free to use the chat box to ask any questions or request technical support. We’ll try to get to any technical support issues as soon as possible. Regarding questions, questions about the content, or for the presenter, we’ll be holding onto them and then addressing them during designated Q&A breaks throughout the webinar. So, thank you for your patience with that. And again, just a reminder to please use the chat box for any questions or comments. You may notice another panel, the Q&A panel. Please do not use that. Use the chat box. Thank you.

We also have a closed captioning service available today if you are interested. I will be dropping a link to that in the chat box shortly. Finally, we are recording the webinar today. And all materials, including the slides and the recording, will be posted online in approximately three weeks’ time. All registrants for the webinar will be receiving an email once the materials are available online.

Our speaker for today is Todd Gibson. Todd is a senior programmer and analyst with Information Management Services. And he is a key member of the NCI TUS-CPS Management Team. His contact information is noted here and again will be available once the materials are posted online. So, with that, Todd, I want to turn it over to you so you can start sharing your slides.

MR. TODD GIBSON: OK, thank you, Nalini. And welcome once again, everyone, to the second webinar in the series for the TUS-CPS. I’m going to start sharing my screen here. Give me one second. To alleviate any problems with bandwidth, I’m going to shut my video off during my presentation and will bring it back on during question-and-answer time.

OK, as I said, this is the second in a series of webinars. This one was “Using SAS to Analyze TUS-CPS Data.” Just a disclaimer before we get started. The views and opinions are my own and do not necessarily represent the views, official policy, or position of the U.S. government, U.S. Department of Health and Human Services, or any of its affiliated institutions or agencies. The agenda for today’s webinar is, I’m going to start off by going through how to obtain the TUS-CPS data. We’ll then move into creating SAS datasets for the 2018-2019 survey wave, generate a table of current cigarette smoking status, and then add the replicate weights for calculating standard errors in confidence intervals. We will move from there into you creating the harmonized dataset, the newly released harmonized dataset that has data through 2019, starting at 1992. We’ll do a couple examples with the harmonized dataset, one including merging the replicate weights to the dataset. And at the end, we’ll have some additional links and contact information and time for questions.

As Nalini mentioned, we’re going to have to put your questions in the chat box. During the presentation, enter any questions into the chat box. At the end of each session, I’m going to take a

few minutes, maybe five minutes, for answering questions related to the current section and then move onto the next session. At the end of the presentation, we'll use the remaining time to answer any questions not covered after each session. And after the webinar, if you have any questions, feel free to email myself or Carolyn at the email addresses.

So, to obtain the TUS-CPS data, you can go to – and this is the main web address for our main webpage. And it shows some introductions, it has the news, stuff that has been updated. And within the website, you can click on and get to our question and data file section. And shortly I will jump on the Web and show you this firsthand, but just to review some of the things that are available on the questionnaires and data page, the questionnaires for the 2001 through 2019 survey waves are available. Technical documentation for the individual survey waves, so starting with 1992, '93, going the whole way up to the 2018-2019 data. The user's guide for the 2018-2019 data. The actual data files, the public-use data files for the individual survey waves and the newly released data file for the 1992-2019 harmonized file. SAS programs to read the public use files and create SAS datasets. In the past we've had this for just a few. Recently we've expanded, and now we have code to read all survey waves. And there's links on the page to the replicate weights available from Census CPS FTP site, and additional tables and reports.

So, to go into a little bit about what we need for today's examples, the data that we're going to be using is the 2018-2019 survey data, which is the July 2018 to January 2019 to May 2019 surveys. The self-response replicate weights for the 2018-2019. As I mentioned, they're available from Census. The 1992 to 2019 harmonized dataset. And the replicate weights for the harmonized dataset. Also available will be the technical documentation and the SAS code.

So, at this time, I'm going to jump up onto our website. And this is the main page. I mentioned there's a little bit of information on the main page, some news about recent happenings. The fact sheet important for today's seminar is down at the bottom. You can click on "questionnaires and data files." This will bring us to our data files page. And there's information on registration for our webinar series. And if you go down about halfway through the webpage, you get to the section on the 2018-2019 data.

So, at this point, I am going to go ahead and -- I know that a document was sent out with links to data that was needed for today, but I am going to go ahead and download the necessary data and SAS code and work through it. I'm going to do it fairly quickly. So, starting with the July 2018 data, I've created a folder on my computer, TUS-CPS, and then within that folder I have a data folder. So, I'm downloading the July data, the January data, and the May data. In addition to the data files, there are two SAS programs. One SAS program is used for the July and January file. I'm going to save that link. I'm going to make one small change. It has a .txt extension. I'm going to change it to SAS, and I'm going to put it in my main folder. And then I'm going to repeat the same thing for the May. Additional things that are helpful to have are the technical documentation for the three surveys, and this is the guide to using the 2018-2019 data. There's data tables, other reports. But I'm going to move up the webpage a little bit to where this harmonized dataset section is. While I'm here on the website, I'm going to download the harmonized data and put it in my data folder. At the same point, I'm going to download the replicate weights that we'll need later on, so for -- I'll go into this more later, but there's three replicate files, one for the 1992-93 data, another for '95 through 2003, and a third for 2006 through 2019. The webpage also has a SAS code for reading in and merging those to the harmonized file. I'll do that within my examples, so I'm not going to download that at this point.

Now that we have the data that we need from the TUS-CPS website, I'm going to jump over to Census's website so that I can download replicate weights for 2018 and 2019. So, on their website, they have a tab, toggle to it, select 2018, and we'll need the self-response replicate weights, so I'm going to download it and put it in my data folder. So, that's for 2018, July 2018. For 2019, there's a section for self-response replicate weights, and one for January and another one for May. So, at this point, if I look into the data folder, we have three files which are the public use files, three files of replicate weights for the public use files, the harmonized dataset, and three files for the replicate weights for the harmonized dataset.

I'm going to jump back to my slides here. So, now that we have downloaded the data, the first thing I'm going to do is create SAS datasets for the July 2018 through the January-May 2019 public use files. The files are text files, so – and we have downloaded the data for 2018-2019. The steps that we've taken is downloading the data, we will unzip the data files, open the SAS program that creates the SAS datasets for July 2018 and January 2019. We may have to modify Filename and Libname statements to match where the data's been stored. We'll run [00:16:17 inaudible] code to create the actual SAS datasets. And we'll repeat it for the May 2019 data.

So, I'm going to go back to my folder that has the data that we've downloaded and, starting with the July 2018 data, I'm going to extract and do the same thing for January and then for May. And since we're going to be using the replicate weights also an example, too, I'm going to do that now, save us a step later on. So, the July replicate weights. We'll go to January. And for some reason my January didn't download. I'm going to go ahead and do the May right now. Unfortunately, the 2019 data didn't download. I'm going to jump back on the website and redownload. For some reason, I'm unable to download it, so I'm going to go to – I have the files stored on my drive, elsewhere on my drive, so I'm going to pull them from there. I'd like to open it up while I'm doing this, if there's any questions, we've received any questions maybe in the chat box we can handle while I'm taking care of this.

MS. CORCY: We do not have any questions so far, Todd.

MR. GIBSON: OK.

MS. CORCY: If anyone does have any questions, though, feel free to enter them into the chat box.

MR. GIBSON: I apologize for the delay. I'm working as quickly as I can.

MS. CORCY: Todd, I am sorry to interrupt. We do have one question. Do you want to address it while you're pulling it up?

MR. GIBSON: Sure.

MS. CORCY: So, we have a question, can the replicate weights that are being demonstrated for use with the harmonized dataset also be used when we create our own datasets from the raw files across 1992 to 2019?

MR. GIBSON: No. The reason being – well, it depends on what sample you're using. If you're just using the self-respondents, you can. But it's better to use the individual files from Census. OK. I apologize for the delay. For some reason, my browser isn't downloading things like it did an hour ago. But now I've unzipped the public use files, the three public use files, which are our survey files, and then I also have the replicate weights. So, sorry for the sidetrack.

So, we were going to create SAS datasets from the public use files, so I'll start off with SAS. We have our July '18–January '19 SAS program. I'm going to open it up in SAS. And now I have it open. As mentioned before, we'll have to go in and change – you may have to modify your Filename and Libname statements to where you stored the files, which I have mine on my C drive, TUS-CPS. So, I've updated the path for the July '18 to January '19 in the library where it's going to be stored. What the program does is it reads in every variable from the file and adds labels. And also, several variables on the file have implied decimal places, so there's code there to divide by 10 or 100 or 10,000 depending on the variable.

So, at this point, I'm going to go ahead and run the code. It will take a few seconds to run. Once it's run, we should see two replicate weight files, or two files with the SAS datasets, one for July '18, one for January of '19. At this point I'm going to go ahead and open the second program, which was for the May 2019 data. And similar to the previous program, go in and change the path to the path where I have the data stored, and run the code to create the SAS dataset for the May 2019 data. And there's a typo. So, after this program runs, we'll have SAS datasets for the three surveys in the 2018-2019 survey wave. Now that the SAS datasets have been created, we'll start with the examples and give a minute or two if there's any questions that have come up. And if there's no questions I'll move on.

So, the first thing that we're going to do is we're going to create a table of current cigarette smoking status. To generate the table, we'll read in the July 2018-2019, May 2019 SAS datasets that were created. We'll make a few selections. One selection is on adult civilian records, which is PrPerTyp=2. We'll want the records that are the interviews for the supplement, which is the Intrview, so "interview" without an "e," =1. And we'll be looking at self-respondents. We'll keep only the variables needed for the analysis, so we'll need year of survey, month of survey, region, age, Hispanic origin, race, sex, unique household identifiers, the smoker recode, which is the current cigarette smoking status recode variable, then the self-response weight. Since we're working with three survey data files—each of the survey data files are weighted to the population, the total population—so since we are using all three, we'll need to divide by three. We'll construct a couple of variables, age group variable and a race/ethnicity variable, and generate a table of percentages and counts by sex, region, race/ethnicity, and age group using proc tabulate..

This next slide is the code that we'll be using. It may be hard to see on here, so I'm going to go ahead and open up the example in SAS. And these are – I'm going to close out our other programs that created the SAS datasets and open example 1. These were the examples that were sent out at the end of last week. Expand this a little bit. So, the code has the library where our SAS datasets are stored. There's a few formats here for age, region, sex, race/ethnicity, and smoking status. This first dataset reads the three files and concatenates them and makes the selections. The next data step here, as I mentioned, we need to divide by the number of surveys, creating our age group variable. So, we have age groups 18 to 24, 25 to 44, 45 to 64, and 65-plus. And I'm creating a race/ethnicity variable using the Hispanic variable on the core and the race variable. Labels for those variables and formats for the variables that will be used in the table.

So, we'll be doing our table with a proc tabulate, a couple titles. This statement here will exclude respondents who have indeterminate for their cigarette-smoking status. Typically, when we do analysis we exclude these. They're people that have answered – had no response or refused or don't know to the two questions that go into making the variable, and those questions are A-1, which is have you ever smoked 100 cigarettes, and A-3, which is current smoking status. So, if you answered yes, you're asked, do you currently smoke every day, some days, or not at all. So, if a respondent had

answered either of those questions “don't know,” they end up falling into this indeterminate smoking status. So, for the table, we're going to exclude those, and we'll do the table by overall, sex, region, race/ethnicity, age group, so I'm going to go ahead and run it. Once this runs, we get a table, and our rows are, as I said, total and then by sex, by region, race/ethnicity, and age group. We have current cigarette smoking status percentages for never, every day, some days, and former, the population, and the sample size.

So, that's the first example. Once again, I'll open it up for any questions for the first example. If there's no questions, I'll move on to the second example, which is, it works off the first example, but we're going to add the replicate weights and calculate the standard errors and confidence intervals. To do this, we'll use code from the first example, but the new stuff is we'll read the replicate weights for July 2018, January 2019, and May 2019. We'll concatenate the replicate weights data into a single dataset and divide the self-response base and replicate weights by the number of surveys. We'll sort the main survey data and the replicate weight data by year, month, and unique household identifier. The identifiers are QstNum and OccurNum. We'll merge the data, the main data and the replicate weights by the sorted variables, and it will do a proc freq to check to be sure that the data merge properly and that every record in the main file has a replicate weight record. And then we'll generate percentages, standard errors, and 95% confidence intervals using the replicate weight in the Proc SurveyFreq.

So, our next slide is the code. It's the additional code from example 1, but I'm going to pull it up, go ahead and pull it up in SAS. Close example 1. Open example 2 up. As I mentioned before, the beginning of the program is the same as what example 1 was, so the formats concatenate in the three survey files together in creating the recode variables is all the same. So, the new piece starts right here, so these are our three replicate weight files. I'm going to update the path so that it knows where the data's at. The next piece of code is a small macro that reads the replicate weight files in. Instead of repeating the code three times, the macro can be used to read the July, the January, and the May. It also ties a couple variables, the month and surveys – month and year of survey is needed to merge, so I passed 2018, 2019, 7, 1, and 5 for the surveys. The next step is to concatenate those three together. This piece of code divides all the weights by 3. And then there's two sorts, one for the replicate weights, one for the main file, and a section of code that merges everything together. Proc Freq to check that everything merged properly. And because we're using the SurveyFreq, I set SmokStat equal to missing if it's indeterminate, and then the SurveyFreq.

Let me go ahead and run that. One of the replicate weight files still isn't working properly. The files on the Census site, the May file has a folder structure that, when I unzipped it, it went into that folder instead of the entire data folder, so it wasn't found, so I'm going to rerun the code. And the SurveyFreq procedure takes a minute, so I'm going to jump back over to my slides and show the results from there. These is the results from the SurveyFreq. We have our current cigarette smoking prevalence, percentages here, our standard errors, and our 95% confidence intervals. One of the reasons I use this as an example is, on our website there's tables for some of the surveys and – for some of the survey waves. As you can see, this table was actually pulled from the website, so shows the every day, someday, former, and never; prevalence with the confidence intervals, the population size, and the sample size; overall by gender and by region. So, it was just to check the estimates that I was getting from the SurveyFreq matched this.

And one final slide for this section. It's been a few years now, but SAS, you couldn't use replicate weights in SAS, so we didn't have SurveyFreq and some of the other procedures that allowed

replicate weights. So, we would use SUDAAN. And this is just the matching code in SUDAAN that we would have used. Some things to point out in the SAS procedure, Fay=0.5 corresponds to what we used to use, ADJFay=4, the Design is the same, BRR. SubPopX is something that's newer in SUDAAN that allows you to select off records.

And that's the end of example 2. So, if we have any questions – and I can't see the chat box if there's any question, so can one of the panelists let me know?

MS. CORCY: Yeah. There is one question, Todd, that's come in. I don't know if you can revisit it or answer it right now, but the question is about explaining the first data steps about the macros.

MR. GIBSON: About the macros?

MS. CORCY: Yes.

MR. GIBSON: OK. Let me pull the code back up. Before I go into that, now the code has run, and we see that there is the frequency table, so there was a match. For every record in May, there was a record in the reps file and then the output. So, I'll go back to the code. The question had to do with the macro here, I think. So, what this macro does is that, instead of having to repeat the code three times, once for the July '18, once for the January, and once for the May, this short macro just – what it sends in is the filename for where it's stored. And each of the files have the QstNum, the OccurNum, the base weight, and then the replicate weights. There's 160 in the '18-'19 data. The way that the replicate weights files are set up is, there's a record in the file even for proxy respondents and respondents to the CPS that didn't get the supplement. So, all that this does, this piece of code right here, it excludes those records. The next two pieces is just HRYear4 gets set to whatever this second value is here, and HRMonth gets set to the third value. does that answer the question?

MS. CORCY: Yes, I think that was what the person who asked the question was looking for. Thank you, Todd. And then there was another question about showing the results, I think the confidence intervals table, which you were just showing before.

MR. GIBSON: OK.

MS. CORCY: There was no question, just an ask to see the table.

MR. GIBSON: OK. Yeah, it's right here, and it's also in the slides, so it will be available when we put it up on the Web.

MS. CORCY: I don't see any questions in the chat box, but again, feel free to enter them at any time, and we will get to them during the question breaks.

MR. GIBSON: OK, I'll move on to the third example. And prior to the third example, we're going to go ahead and create a SAS dataset of the harmonized dataset or file this time. And just a couple notes on the harmonized dataset. Included, as I mentioned earlier, is – included with the data file are two SAS programs. One is the main SAS program that reads the data and creates the SAS dataset. And the second one is just the formats that are used. And this second program is called by the first one. As I mentioned when I was downloading the data, There's several other useful things for the harmonized file: the technical documentation, which gives an overview of the CPS, the TUS, and the harmonized data; the Proc Contents of the data file; there's unweighted frequency tables for all the

variables, which can be helpful when you're working with the data and you're not familiar with it, just to make sure that you're reading the data properly and coming up with the same kind of results. And another helpful thing is an Excel table that lists each variable in the file and the source of that variable by survey wave. Over time, question numbers change, variable names change, so this is a nice document to have just to see how the harmonized variables were developed. So, we'll go ahead and create the 1992-2019 harmonized dataset. I'll download the data. We've already downloaded data, so now we'll unzip the data files, open a SAS program, and create the SAS dataset, and modify the Filename and Libname statements similar to what we've done for the other programs, run the code to create the dataset.

So, I'm going to go ahead back into my data folder. I'm going to extract the harmonized file and then each of the replicate weight files so that we have – once we get those all extracted, I'll go ahead and start telling the code to create the – because the two pieces of code used for creating the SAS dataset were in the zip file, I'm going to move them to my root files so they're with the rest of my SAS code. As mentioned, there's two programs, that's the main programs and the formats program. So, I'm going to go ahead and open that, the main program in SAS and close our other example. There's our main programs. Again, changing the path for the data. That's where I have it stored. And also the path for the second program that has the formats. As you can see, that's the variables. And there's code to take care of implied decimal places and labels and formats for all of the variables in the RMS file. And this piece of code also has a Proc Contents at the end. I am going to go ahead and run that. This will take a minute to run. It's 1.5 million records, 1.7 million records. There's Proc Contents, 288 variables in the file. Variables are variables from the CPS core and the variables from the tobacco use supplement.

So, at this point we'll double check that the file was created in our data folder. The harmon SAS dataset is right there, and we'll move into the examples using the harmonized data. So, I'm going to create tables similar to what we've done with the other examples. One is going to be current cigarette smoking status, but I'm also going to add a table for e-cigarette smoking status. To do so, we'll read in the harmonized dataset. Unlike the previous examples, the harmonized data set only has self-respondents, so we don't have to do the selections that we did on example 1 and 2. And we'll keep variables that we need for the analysis or merging the replicate weights. The survey wave, the year, month, region, record I.D., sex, the variable for current cigarette smoking status, the variable for current e-cigarette smoking status, and the self-response weight. The harmonized data has data from all 10 survey waves, and one thing to note is the 2000 survey wave only had two surveys in it. So, for that survey wave, when you're using the data, you divide by 2 whereas all other survey waves, you would divide by 3. We'll generate a table of current cigarette smoking with percentages and counts by survey wave and gender, sex. And then we'll also generate another table of e-cigarette smoking.

This slide's a code for doing that. As you can see, it's a lot less code because things are already coded in the harmonized file, so I'm going to head over into SAS again, and I'm going to open example number 3, look at my library statement and also where the formats are stored, uses the formats. And so, as I mentioned earlier, this will read in the harmonized file, keep variables that we're interested in for these tables. The second dataset divides the 2000 data by 2 weight and all others by 3. The survey wave 4 corresponds to the 2000 data. Proc Tabulate, once again, CigStat=-9, so we're excluding those with indeterminate cigarette status. And for e-cigarette we're doing the same. We're excluding the indeterminate and also only including survey waves 9 and 10, which are the 214, 15, and 218-19 data. I am going to go ahead and run. It will take a minute to run here. I will let this run because we have data from all 10 survey waves in the harmonized file. We can generate our cigarette smoking status percentages for all 10. And for this one I just used sex, so we have the total and by sex for each of

the survey waves. And then for e-cigarette smoking status, we use the '14-'15 and '18-'19 data, and I did it by sex and also by region. These are just slides of the results and that's the end of example number three.

Any questions? OK. So, I'll move on to the final example that I'm going over today. And this example will merge the replicate weights to the harmonized dataset. It shows how to merge them. The replicate weights for the harmonized file are in three files. Over time, the number of replicate weights generated by Census for the TUS change. In '92-'93, it started with 48 replicate weights, and then starting in 1995 and going through 2003, it was 80 replicate weights. And then starting in 2006 to the current, it has 160 replicate weights. As mentioned before, there's SAS code on the TUS-CPS website showing how to read the three files and merge them into a harmonized dataset. To do so, we'll read the harmonized dataset, read the three replicate weight files, concatenate the replicate weight datasets together. We'll need to sort both the harmonized dataset and the replicate weights by survey year, survey month, and record I.D. One note on record I.D., this variable, if you're merging the individual files, say just 1819 to the replicate weights, we used QstNum and a current number. That's the variables that have been used recently. In the past, though, those variables have changed, so one of the things that we did in the harmonized file to make things easier and also in the harmonized replicate weights was create a unique record I.D. that can be used. It's unique within surveys, so you still need survey year and survey month. We'll merge the harmonized dataset and replicate weights by the sorted variables. Once again we'll do a Proc Freq to check that the data merged properly. And for this example we're going to use the survey means to calculate the number of cigarettes per day. So, the code, once again, is in the slides for future reference.

And I'll go to SAS now and bring the example in and run it, bring the example in and run it. So, example 4. It's updating the Filename statements where the data's stored and also where the formats are stored. So, the first dataset, just reads in the harmonized SAS dataset, keeps survey wave, year, month, record I.D., sex, CigStat, current cigarettes per day for daily smokers, and current cigarettes per day for some-day smokers in the self-response wave. The next data step reads in replicate weights for '92-'93. As you can see, there's 1-48 replicates. The next step is for the '95 through 2003 that has 80 replicate weights, and the 2006 through 2019 that has 160 replicate weights. The next step concatenates the replicate weight files together. You sort the harmonized file and sort the replicate weights and merge them by survey year, survey month, and the record I.D.s, add Proc Freq to make sure that everything merged properly, and for this example, we're only going to look at the 2014 data. The code could've been written a little bit differently so that we only looked at only folding in the 2014 replicate weights from this file here. But I wanted to give an example showing how to merge all three of these with the harmonized file. So, we have SURWAVE=9, which is the 2014-'15 data. And we go through, similar to what we've done in the previous examples, dividing all weights by 3, the main weight and the 160 replicates. The next data step, we create a current smoking status variable, which is set to 1, if the respondent is either an every-day or someday smoker. CigStat=2 and 3, or a nonsmoker, which is never or former, which is CigStat=1 or 4, else it sets current smoke to missing. Because our number of cigarettes per day are in two variables in the harmonized file, one for daily smokers, which is the CPDD variable, and one is for some-day smokers, which is the CPDS variable, one cigarette stat equals 2, which is every day. We set our new variable to what is CPDD, and if it's some-day smokers we use the CPDS variables. Then we have a Proc SurveyMeans. Once again, Fay is equal to 0.5, our variable is Cig per day, our domain is current smoking status, our main weight and our replicate weights. And for the example, I'm just going to do this overall, but I have included code to do it by sex.



So, while the code is running, I'll jump back to the slides. I have the results in the slides. This is the code, the second part of the code. And the results, it shows the number observations in some of the weights, various variance estimations, and then a table with number of cigarettes per day and the means, the standard error, and 95% confidence intervals for the mean, and a distribution of the variables, and then also the information by nonsmoker versus current cigarette smoker. I mean, a nonsmoker wouldn't have any cigarettes per day, so – and then a domain analysis. And, as you can see, it takes a while when you're using the harmonized file when you're merging all those replicate weights for the SAS code to run. So, that's the end of example 4. If we've had any questions as we're waiting for this to run.

MS. CORCY: No questions thus far. Again, I would encourage participants to enter any questions that they do have into the chat box. Thank you.

MR. GIBSON: So, that example 4 was my final example for today. One thing that I want to mention is for a more in-depth example of using the replicate weights in the harmonized data file, I would refer to the first webinar that was analysis using the harmonized dataset. And that webinar went into a little bit more detail of using all the replicate weight files and generating an estimate using all 10 survey waves. My final slides are just – I have some links to the main website, the questionnaires and datafiles website or FAQ, and also a link to our workshops. So, you can go to and see what we have presented at previous workshops. And my final slide is contacts, which is the main email address for the project and Carolyn's and my email addresses. And finally, if there's any questions, I'll take any questions about any of the examples. As I mentioned before, if there's no questions for today, feel free to email me. I do apologize for the couple of snafus earlier on with the data file that sort of threw things off. Please email with any questions.

MS. CORCY: Hi, everyone. I also wanted to offer, you know, in case you're having trouble with the chat box, the other option for asking questions, now that we're at the end of the presentation, is to raise your hand. So, I have instructions on screen on how to raise your hand. And if I see your hand raised, I can unmute you so you can ask your question instead of typing it into the chat box. So, please let me know if you would like to by raising your hand.

OK, I do see a couple of questions coming through the chat box. So, the first question is, is it common to analyze the one set of data alone? For example, analyzing January 2019 alone.

MR. GIBSON: We recommend using the three surveys within a wave because it gives you a better estimate. I leave it up to others on the panel if they want to chime in. But we feel, especially if you want to look at, like, things at a state level, to get a big enough sample size we recommend using the three.

MS. ANNE HARTMAN: Can you hear me?

MS. CORCY: Yes, we can hear you.

MS. HARTMAN: Hi, this is Anne Hartman. You can use one month of data, depending on what your object is. If you were trying to look at a certain date in terms of evaluation of an intervention, as long as the sample size is large enough [01:11:16-01:11:23 inaudible] nothing you can do. In general, unless there is something about a particular date, it's better to have more data and use all three months of a wave because each month has also their separate replicate weights.

MS. CORCY: OK, does that answer your question? Feel free to type in the chat box or ask to be unmuted if you would like to follow up. OK, it looks like that was a good answer to the question. There's another question. I think someone was running their code, Todd, and they received an error that said, due to nonpositive weights, 163,000 observations were deleted. They're not sure what the issue was. Any suggestions on how to investigate that error?

MR. GIBSON: Which example were they running? Are they saying they were just doing something on their own?

MS. CORCY: No, I don't see details about that. Example 4.

MR. GIBSON: Example 4? To me that sounds like the merge didn't happen properly, but without seeing, it's kind of hard. It would probably be easier to follow up with me via email.

MS. CORCY: OK. Sounds good. And I do see that we have a hand up, so I'm going to unmute that participant. Luis, you should be on unmute now.

LUIS: Hello, can you hear me?

MR. GIBSON: Yeah.

MS. CORCY: Yes, we can hear you.

LUIS: Thank you for your presentation. It was very helpful. My question is about that in U.S. people can respond – some participants can respond more than one time the survey. So, when you merge the datasets, how do you handle that people can participate twice in the survey? Is there or you can exclude these participants, or what's your recommendation about this for descriptive and also for divided models? Do you think that it is correct to exclude, for example, a second wave that the participant respond to the survey, or not?

MR. GIBSON: There is some overlap in survey waves, and maybe Anne can chime in again, but for most survey waves, there isn't an overlap, the way that it's designed.

MS. HARTMAN: Let me see. Yeah, that's true. There's only a couple. Probably '14-'15 and maybe '18-'19 where there are some. It's not a large number. And you probably can – I mean, even if you're using the harmonized file, I believe you can – in any analysis you can still select who you want to be in the analysis. So, you would have IDs, right?

LUIS : Yes.

MS. ANNE HARTMAN: You could identify those people who were in more than once. And also, lots of times they may have been – especially '14-'15, they may have been proxy. And if you are doing a self-analysis, it will be even a smaller number because they'd have to be self.

LUIS: Right. Thank you. Like now, for 2018-19, when you identify these duplicates who can – I think that there are, like, 3 percent are duplicates or something like that. So, have you ever tried to exclude these people, sort of samples or the tables that you present, and is there any change when you exclude these people from the estimates in the U.S., for example table 1 and table 2 that we see today? Are there any exercise that you can see? There are no difference when you exclude these people, or do you know what happened when you exclude the people?

MS. HARTMAN: Well, all I'll just say is that we did do some analysis, but I'm not sure if we did it just on one month or if we had all three. So, I don't know, is Maggie still here? If not, maybe we can have her follow up with you. So, if you can send the question, we could have a further discussion.

MS. MAGGIE MAYER: Yeah, I'm here. I am happy to follow up.

MS. HARTMAN: Oh, Maggie. You are there. OK.

MS. MAYER: Yes. Luis, I'll drop you my email address in the chat.

LUIS: Thank you, thank you.

MS. CORCY: OK. Are there any other questions? OK. If not, thank you, everyone, for joining us today. Again, as Todd said, if you do have any other questions that you think of after the webinar, please feel free to reach out to us. Our email and contact information is listed on this last slide. Feel free to also visit us online, and we also have an email subscription that's available. You can sign up for the latest news and updates on the TUS-CPS. Again, thank you for joining us today. We'd really appreciate it if you'd provide feedback on our webinar by completing a brief survey. I'm dropping the link in the chat box right now. Once again, thank you so much for joining us today, and, as a reminder, all of the materials will be posted in approximately three weeks' time, and we will send you an email when they are available. Thank you, everyone, and bye.

MR. GIBSON: Thank you.

[ event concluded ]