

**National Cancer Institute
Tobacco Use Supplement to the Current Population Survey (TUS-CPS)
2021 Data User Webinar – Linkage of the TUS-CPS to Other CPS Supplements**

Webinar Transcript

MS. NALINI CORCY: Hello, everyone, and welcome. My name is Nalini Corcy, and I would like to thank you for joining us today for the third installment in the National Cancer Institute TUS-CPS Data User Webinar Series. Today's webinar is going to focus on linkages between the Tobacco Use Supplement, or the TUS, and other CPS supplements.

Just a couple of reminders and things that I want to go over before we get started today. First of all, all participants are on mute. If at any time you have questions about the presentation, or you need technical support, feel free to type into the chat box. You may see another panel on your platform, depending on how it is set up. Please ignore that panel and use the chat box only. For any technical support requests, we will try to get to them as soon as possible. Regarding questions about the material being presented, we will be saving them and fielding them during designated question breaks throughout the webinar. So, thank you so much for your patience with that. We also have a closed captioning service available today, and I'm dropping a link to that in the chat box. So, if you need it, it's available. You can open that link in a separate window and follow along with the captions. The webinar today is being recorded. All of the materials including the recording and the slides, as well as the transcript, will be posted online in approximately three weeks time. Once they are available online, you will receive email notification on where and how you can access materials.

We have two speakers joining us today. The first is Dr. Margaret "Maggie" Mayer. She is a Cancer Research Training Award Fellow with the National Cancer Institute Tobacco Control Research Branch. Also joining us today is Dr. Carolyn Reyes Guzman, who is a Program Director with the NCI at the Tobacco Control Research Branch. So, I want to thank them both for being here today as well as other members of the NCI TUS-CPS team and our colleagues. So, without further ado, I'm going to stop sharing and pass the screen onto you, Maggie. Okay, Maggie, you should have control whenever you're ready. Please go ahead.

DR. MARGARET "MAGGIE" MAYER: All right. Can you see and hear me?

MS. CORCY: Yes, to both.

DR. MAYER: Alright. Okay. I'm going to leave my video off for the presentation, but I will turn it back on again when I get to the end. So, thank you everyone for joining us today to talk a little bit about linkages using TUS-CPS. As Nalini said, my name is Margaret Mayer. I am a Cancer Research Training Award Fellow here at NCI. In this portion of the webinar, I will be presenting a case study on planning and linking the Tobacco Use Supplement to the Current Population Survey or TUS-CPS to the American Time Use Survey or ATUS. The goal here will be to provide an overview of the ... Did I hear something?

DR. CAROLYN REYES-GUZMAN: I think, sorry. Sorry. Yes, yes, go ahead Nalini. I was going to say something but go ahead.

DR. MAYER: Oh, thank you.

MS. CORCY: Sorry, I just wanted to say, you may want to change your display settings. So, we see a fuller screen.

DR. MAYER: I've done this like four times.

MS. CORCY: Not a problem. Yeah, the first one should change it. There you go, perfect. Thank you.

DR. MAYER: Alright, apologies for that. Okay, so yeah, the goal is really to present more of a roadmap than a research talk per se. So, with that said up front, I will give the standard disclaimer, the views and opinions expressed are the presenter's and they do not necessarily represent the 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.

Okay. So, first, I will start with some background on the Current Population Survey, which is the parent survey for both TUS and ATUS. Some of this may be familiar or for those of you who have used CPS or TUS them before, but just so everyone is on the same page, there will be a decent amount of background here.

So, CPS is a monthly labor force survey that's conducted by the Census Bureau in about 56,000 interviewed households across the U.S. Each household is interviewed monthly for four consecutive months for one year and then again for the corresponding time period a year later. We refer to this as the "4-8-4" design. So, to demonstrate the "4-8-4" design, we look at this chart. In this example, a household enters CPS in January of year t, and it does months one through four of CPS between January, and April. And then it's eight months off between May and December, and then completes the final four months during January through April the following year, and then it exits a panel after month eight. So, the month on, we refer to these as the months in sample or "MIS" it's sometimes done in shorthand. So, months in sample one through four come before the eight months off and then month in sample five through eight come after the eighth month off.

Okay. So, this study design allows a reliable month-to-month and year-to-year comparison. This is because so much of the sample is consistent from one month to the next and then from one year to the next. In addition to labor force data, the secondary purpose of CPS is to collect information on the demographic characteristics of participants; and then questions on additional topics are included as what are called supplements and that's where we will spend the majority of our time today.

One such supplement is the Tobacco Use Supplement. TUS is the largest nationally representative survey of tobacco use among U.S. adults 18 and older. TUS has been administered as a supplement to CPS every three to four years since 1992 and survey waves are listed here and range from to '92 to '93 and most recently in 2018-2019. For each of these survey waves, data are typically collected during three time points from the month. Except for in 2000, when only two time points were used. So, this means that everyone who is in the CPS panel, will be you know, invited to take the TUS so regardless of that 4-8-4 design, it does not matter what month in sample you are, you can be invited to take the TUS and that is a fact we will just put in our back pocket for later.

The TUS is not the only supplement. There are many other supplements and that is a lot of what we want to talk about today. We, it is important to note here that the supplements, in addition to covering lots of different topics, they cover a lot of different time frames too. So, as I said, TUS is

administered every three to four years. The American Time Use Survey just looks at how Americans spend their time, and it's what I will be talking about, is administered continuously because they want to see how time is spent across days of the week, weeks of the month, months of the year. The Annual Social and Economic Supplement, which Carolyn will talk about, is administered in March. The Food Security Supplement is typically, I believe, administered in December. And then, some of these other supplements are examples of things that either have been added later on or are discontinued or are no longer being offered. So, just important to note that there's a lot of information out there that can be linked to CPS, because it has been administered as a supplement, and that brings us to the next point, which is, what do we mean when we say linkage?

To address various research questions, data can be linked from one month of CPS to another. For example, if you want to track employment over time, it can be linked from a month of CPS to that month's supplement or supplements. In the case of TUS, we do this for you when you download a public use file of TUS from our website. It contains not only the tobacco use data, but also the core CPS labor force and demographic variables that were administered that month as well. And then, the third point here is that you can link from one month supplement to another month supplement. Using that month's CPS data. And that is where we will be spending the bulk of our time today. So, when we say linkage, we are typically referring to the matching of the participant TUS data to their record from another month supplement. Specifically, in our cases, ATUS, and then for Carolyn, ASEC.

Okay. So, moving into the case study here, again, this is meant to be more of a teaching example. What sorts of considerations you may want to think about if you're planning for a linkage using TUS and another supplement. So, if we are linking TUS, which has a wealth of information about tobacco use, with another data set that contains time use data, we can think about how time is spent engaging in health behaviors and how that might differ across tobacco user groups. For example, we might want to consider how patterns of time spent exercising or eating, acquiring, and preparing food, sleeping, or timed that using the computer or Internet, via the proxy of media exposure, might differ between tobacco user groups.

So, linkage could be a powerful tool, but it requires some careful planning. Thankfully, there is a CPS supplement that looks at time use. And that is the American Time use Survey or ATUS. So, ATUS, in their own words, provides nationally representative estimates of how, where, and with whom Americans spend their time. It covers a wide range of activities from work and childcare, to volunteering and recreation and data are collected on most days of the year. With the exception of holidays, weather, government closing, things like that. This is again to try and get a robust estimate of how individuals spend their time. So, the data will come from every day of the week. Most days of the month, and most days of the year. Interviews are conducted over the telephone, using CATI and they are done by a combination of structural, structured interviews. Sorry, that was a typo, and conversational interviewing.

Again, this is supplemental to CPS, so it shares some methodology, but it does differ in certain ways. In terms of differences, CPS participants are eligible to be sampled for ATUS two to five months after leaving the CPS panel. So again, after they complete their four months on, eight months off, and four months on again, they become eligible two months later for ATUS and remain there for five month post completion. And ATUS is conducted continuously. Unlike TUS, which is administered in very particular months that are planned in advance, ATUS is done consistently and continuously. ATUS focuses on national-level estimates, so the sampling is proportion to state size. This is different from CPS and TUS which are interested in getting state-level estimates. One eligible person is selected for

interview per sampled household. So, that's just an extra layer of sampling that is added in there, compared to what would be in CPS or TUS. Participants are interviewed only once about what they did the previous day, where they were, and who they were with. Which means that this is, of all the data sets, a very complicated one. So, right off the bat, there are a couple of things we might want to look out for, as we plan for this linkage, given the characteristics of ATUS that we've just described.

So, first, the ATUS sample is much smaller, it is a subsample of the CPS sample, it includes only one eligible participant per sample household, and because of it a very intensive survey, the participation rates can be lower. So, this reduces the number of participants that can be matched between a given month of TUS and ATUS. The plus side here is that TUS is actually quite large. So, we are starting from a very good point. And there's also the option to pull multiple waves of TUS if you are concerned about sample size. So, these are the questions to ask toward the beginning of the planning process. Before you move forward.

So, next, ATUS eligibility follows exit from the CPS sample. So, this means that time use data is collected anywhere between two and 20 months after the tobacco use data are collected. This variable could impact the interpretability of your results. So, depending on the research question, this may not be an issue. If you're looking at very stable behaviors, for example, or if you are planning to do activity analyses that are stratified by time, or time has elapsed between the surveys, you may not have an issue with this, or you may be able to circumvent it, but it is something to consider early on.

So, let's say that we are not particularly worried about those concerns. We are planning to use separate waves of data, planning to look at a really common group, we are going to look at behaviors that we think are really stable. We continue to do our homework by digging deep into the documentation and understanding the data and how to use them. We learn that there are several different data files. You see, this list is very busy. But this is more to demonstrate how complicated some of these supplements can be. ATUS in particular has a very complicated structure and that is reflected in the number of files that are available to the user. When working with ATUS alone, it is enough to link these files using their own unique identifier, which they label TUSCASEID but it's not enough to link the files of TUS. So, for that, we need the CPS unique identifiers which can only be found in the last file down here – the ATUS-CPS file, which contains information from the participants last administration of CPS. So again, careful planning is key – understanding what is available to the user, and what's included where.

In terms of what we have for TUS, I won't spend too long here because the other webinars and resources on our website cover the structure of TUS data, where to find it, and how to use it, but I will highlight a couple of things for the purposes of this example. We opted to try and link the 2014-2015 TUS-CPS to ATUS, and this is because the '14-'15 survey wave is the most recent TUS wave with, I will say "complete follow-up" in the publicly available data set. The 2018 -2019 TUS the people who are interviewed in the last month of the 2019 TUS who are very early on in their CPS rotation may not yet be eligible for ATUS in 2020. And so, that - using the most recent wave won't have the best set up in terms of follow-up. So, we are going to use '14-'15, 2014 to 2015. It's comprised of three survey months. July '14, January '15, and May '15. The data files for each of these surveys are on the TUS website on the data and questionnaires page, along with the back code to read in the file. The replicate weight files and the SAS code to read in are available on the Census website. And those are pretty easy to find. But the one other thing I want to note here is that some of you may be familiar with the harmonized TUS-CPS data set. And they have combined all of the survey waves and they've harmonized the variables to the extent that they can. And while that is very convenient, and very useful for a lot of

things, the harmonized data set does not have the necessary variables to be linked with any other CPS supplements. So, if you are planning to link TUS to another CPS supplement, you must use the individual survey waves that are available on the NCI website. Okay.

So, now that we thought about or looked into the study design of the supplements and familiarized ourselves with the data files that we will actually be linking, let's shift gears a little bit and think about the timing of the data that we are hoping to link. There are a few reasons why we might want to do this. If we are using several reasons, TUS, if we are using many variables, or working from home during the pandemic and we're using the VPN, [and it's] running slowly, not that that has ever happened to me; we may improve our runtime if we narrow down the ATUS files from 2003 to 2020 as they are available on the website to something that is a little bit of a shorter range. The other factor here is that there can be some temporal factors that might influence the match rate. So, it's really helpful to know – when are we drawing the data from? What does that coincide with in terms of the CPS structure overall? And how might that affect our match rate? But I will return to that in a few moments.

So, when might we find ATUS responses that link what 2014 and 2015 participants if they exist? If you are like me, you approach this using something visual. Here we have a little timetable. Each of the columns represents a calendar year. Remember that CPS participants are eligible for TUS regardless of the month in sample. That means that in July 2014 TUS, we have participants in every month in sample, first through eighth. CPS participants are eligible for ATUS between two and five months after exiting the CPS sample or completing month in sample eight. If someone is in their eighth month in sample during the first month of the 2014-2015 TUS-CPS, they will have the earliest ATUS eligibility window. The earliest we may find an ATUS record that matches with the 2014-2015 TUS record will be September of 2014, which is highlighted here with that black box.

Let's look at the other end of the range. The latest month of the '14-'15 TUS-CPS is May of 2015. Participants have the earliest month in sample and the last to become eligible for ATUS. So, participants in the first month in sample in May of 2015, will reach their eighth month in sample during August of 2016, which is denoted by the asterisk making them eligible between October '16 and January '17. So, the latest that we may find an ATUS record that matches with the '14-'15 TUS record is January of 2017. So, in total, the possible range of time for our linked data is July 2014, which is the earliest time that somebody could complete TUS, to January 2017, which is the latest time that someone could complete ATUS. However, records can be found for any time during the period. Depending on what month they were, that they took TUS – whether it was July '14, January '15, May '15, depending on what month in sample they were at the time they took TUS. These are just the outer bounds if you're trying to narrow down the ATUS data to only the month and years that you might find matches – if they do exist. This is the range that you want to subset your files to.

And, as promised, here's the additional reasons we are looking at timing here. At some point, roughly halfway through a decade, a Census begins to phase in the new CPS sample, which is based on the most recent Census, this means that some households are released from the sample before they complete all eight surveys, and some households are added in the middle of their 4-8-4 rotation. Over 16 months, the sample is entirely replaced. This means that if you're trying to link data during the timeframe, you may have lower match rates. In this case, the July '14 to January '17 timeframe that we just calculated falls within the phase in period with the 2000 sample to the 2010 sample, which, the 16 month phase in period would be April '14 to July '15. So, I'm kind of expecting a lower match rate because ATUS is generally smaller. We may see in even lower match rate because the

CPS sample is changing as we go along. We also want to be careful about false matches because of the recycling of household identifiers so that we can weed out matches with a few checks later on.

Okay. So, now we are ready to work with the data, we've done our homework. We'll start by preparing the TUS data. I'm not going to go too far into depth on this section if you like to learn more about working with TUS data including the adjustment of weight, you can check out - I believe it was the most recent SAS webinar with Todd Gibson and there are also some resources on the TUS website. A couple of notes, actually, first we are going to merge the data with the replicate weights. We are going to subset the data sets by selecting only the self-responding adult civilians, and we are going to keep only the necessary variables. This is going to improve run time immensely; TUS is quite large. We are going to stack all the survey waves. Now again the survey waves have the weight already. We are going to adjust the self-response survey and replicate weights, in this case dividing by three because we are using three months of data. So, we'll get a weighted population size that matches the U.S. population.

And a couple of notes here. Pay attention to the variable names and the consistency across the years. The consistency across years is really only necessary to worry about if you're using multiple waves of TUS data. I ran into this a couple of variables when I was working on this with multiple instances of TUS data. So, paying attention to how things change will save you some trouble in the long run. Also, it takes a lot of forethought to plan ahead and identify, not only the variables that you need for analysis, but also what you will need for linkage. But hopefully, this example will help with the latter and again you may find that your particular project calls for something different and that is okay. Hopefully, this gives you a rough idea of what you will need to plan for. So, then we are going to prepare our ATUS data. We are going to merge the respondent file the activity summary file and the ATUS-CPS file. These come from that slide earlier. The respondent file has date on the reference day. The Activity summary file has the time use data. The ATUS-CPS file has data from the respondent eighth month in sample with their last month of CPS including the unique identifying variable that we are going to need to link ATUS to TUS. These files merged with each other using a supplement specific identifier, that TUSCASEID that I mentioned earlier, which is different from the unique item part of the variable used by TUS-CPS overall. Again, one thing to be aware of, it's important to be careful of many overlapping variables, especially using the ATUS-CPS file. There are variables in that file that overlap in terms of names with what is in the TUS-CPS. So, paying attention to what those are and renaming them if you need to, dropping them if you don't need them will be very helpful. I will get back to this in a minute. But, demographic variables, from the ATUS-CPS file and the TUS-CPS can be used to confirm the matches, so you want to keep those for later. You'll Just want to make sure you keep track of what they are.

Okay. So, final step in preparing the data for linkage. We're going to create a composite of unique identifier variables. To do this, we'll use HRHH ID and HRHH ID2 which are the unique household identifiers, and then the TUI number which is the person identifier. We are going to create a variable called LINKID, but you can call it whatever you want. And for each supplement in the staff data step, I included the following code that shows the combination of the unique identifier variable.

So, then we are going to merge the data set by the LINKID, we are going to create a link identifier variable to identify the cases that meet the following criteria. We want to have an exact age match, or we want the ATUS age to be no more than two years older than the TUS age. That's really the extent of what we could expect to see change in that two to 20 month time period. And then we want to see an exact match for sex and race. Finally, we are going to tabulate the percentage of TUS

participants who have successfully been matched that meet the above criteria, and we are going to check the coverage across subgroups and the weight if necessary.

Okay. So here, we have some code for the match, linking the prepared data set with the prepared TUS data set with the prepared ATUS data set by the LINKID that we created a couple of slides ago. If the record is found in both data sets. Then, we will want to check to see whether or not it's a false match or a true match. And this is where the renamed demographic variable coming in handy. So, any of these demographic variables that came from the ATUS file, I had appended them with underscore A [_A] so I knew which ones they were. And I'm looking to see whether or not we have an exact match on sex, an exact match on age, and then a little window of wiggle room. Sorry, the race, and then a little window of wiggle room for age just in case someone did have a lag in time between their TUS and ATUS survey.

So then, finally, we will create a flag to identify cases that successfully match. So, there are records in both the TUS and the ATUS files, and then their exact matches on sex, race, and close enough on age, and that includes equals one that are going to be our linked sample.

From there, we can determine what percentage of cases were successfully linked. So, included equals one. We can run a quick proc freq for the included variable among the TUS records, and we find that 12,531 records were successfully matched with an ATUS record, which I think is a match rate of 7.64%. Coverage here is very low. But, as with many things, the benefit of using TUS is that there's a very large sample size to start with. So, even with 7.64% match rate, we are working with 12,531 records to analyze, which will support a lot of different analyses. Again though, it will depend on the research question, so, considering things in advance is helpful.

So, in a separate example, Carolyn will talk a little bit more about coverage among subgroups and checking the need to reweight, but there it is, a case study in preparing for and executing the linkage between TUS-CPS and ATUS. I will reiterate the goal here is not to necessarily make a case for this particular linkage but offer guidelines and consideration as sort of a roadmap for any potential linkages that may suit your research needs. I've tried to compile those considerations here, although every case will be different. Mostly, they come down to this – careful planning is key. Researchers should think about the expected match rate, and the subsequent sample size. Will this be enough to support the analyses that you're planning to do? Think about the timing of the supplements being linked. Will the timing effect the interpretability of the data? Think about the data sets that you've chosen and the relationship to the phase-in and phaseout period for CPS samples. Is this going to be a problem for your match rate and sample size? Plan ahead to check for overlapping variables, variable names, you know, have you identified the necessary variables for linking and matching in advance and renamed them as needed to avoid confusion there?

Finally, I just want to stress that documentation is your friend. It is through very careful review of documentation that you can develop an effective plan to link your data, and there are so many resources available through Census, through NCI, through other agencies that sponsor CPS supplements. If you're struggling though, please reach out. We want to help. I have included an email address here on the bottom and then my contact information will appear, and in a few slides and – my colleagues and I are really, really happy to troubleshoot and discuss any issues or questions that you may have about TUS or about linkages in general, because we do want to make sure that we are supporting the extramural research community in the best way we can. So, with that, I will thank my colleagues for their excellent work. And here is my contact information, and I will wrap up. Nalini, should we take questions now? Or?

MS. CORCY: Yes. I don't see any questions in the chat box just yet, but we can give everyone two minutes if they do have any questions. Again, please type them into the chat box.

DR. MAYER: Nalini, I can answer stuff later too if it comes up between now and then.

MS. CORCY: That's fine, yeah. I don't see any questions right now. So, we can continue, and I believe next we have Dr. Carolyn Reyes-Guzman. So, I'm going to pass presenter privilege to you Kristen so you can go ahead to share your slides.

DR. REYES-GUZMAN: Actually, Nalini if you want to pass it on to me that would be better...Okay, great. Let me share my screen...Okay.

Alright, thank you, Maggie, for an excellent presentation. That was I think really interesting and sharing the challenges and intricacies of, you know, the American Time Use Survey, the potential for a really interesting research questions, and the linkage to ATUS, so thank you for that. So, I'm going to be talking about another example of a linkage. Can you all hear me okay? Sorry, I'm a little bit soft spoken. Nalini, can you hear me okay?

MS. CORCY: Yes, we hear you.

DR. REYES-GUZMAN: Great, okay wonderful. All right. So, again, I'm going to be...good afternoon, everyone, I'm going to be talking about the linkage of the TUS to the ASEC CPS supplement as Maggie very nicely described all the different supplements. And again, my disclaimer similar to Maggie's. Here's what I plan to cover. So, first I want to walk you through how TUS and ASEC fit into the context of CPS, I'm going to talk about the weighting issue of TUS and ASEC, and I'm going to walk you through the specifics, with some SAS code on how to link the ASEC and TUS supplements. And then I'm going to walk through an applied example of this linkage based on an analysis that I've been working on with collaborators from NCI and CDC on smoking in vaping and policy attitudes among residents in multiunit housing. So obviously the house in question comes from the ASEC as you will see soon. Just also, call out if you have any questions, please feel free to type them into the chat box, as I'm going through the presentation, and Nalini will ask those questions at the end.

Okay, so let's begin, sort with the basics of how TUS and ASEC fit within CPS. So, a very brief recap, I know Maggie talked about this quite a bit, but I just want to get us all on the same page again. Again, TUS is another CPS supplement, the survey is fielded every three to four years, and each cycle has three time points. And so, as Maggie mentioned, the TUS has the fielding in July of the first year and then in January and May of the second year. And here's some important things that – I think you may have seen this slide in the previous webinar, about the importance of TUS, but mainly, to track progress and assess actual trends and looking at tobacco-related health disparities, evaluate tobacco control programs, and gather some interesting data on national, state, and county level on many of these tobacco-linked patterns.

So, just to get you more familiar with ASEC, ASEC stands for the for the Annual Social and Economic Supplement of the CPS, it's also known as the March supplement, mainly because all data is gathered in the month of March. Some respondents do participate in ASEC during February and April, but most of the data is gathered in March. The universe of ASEC includes noninstitutionalized U.S. populations and also members of the Armed Forces living on post or their families on post where there is at least one, at least one civilian adult living in the same household. Whereas the TUS only includes the civilian noninstitutionalized U.S. population, not any Armed Forces. For ASEC housing units are

selected using probability samples similar to CPS and TUS, and states, regions, and divisions in ASEC are identified in their entirety, and there is also some kind of level data for metropolitan areas similar to what you can find. One of the unique aspects of ASEC data is that it captures detailed employment information among persons aged 15 and over. So, this little table is just meant to [review] some of these sampling designs – difference between CPS and ASEC. So, the general CPS, again as Maggie said, is a monthly sample. Approximately 60,000 occupied households each month. Whereas, the ASEC has about 98,000 households and it does include an oversample of Hispanic households. The general CPS design is a multi-stratified sample, again, following the 4-8-4 panel design whereas the ASEC, it's also multistage sample housing units, but it has a hierarchical file structure with three record types of possible household family in person. The CPS, by having the monthly survey, it allows for a constant replenishment of the sample, with a panel design without excessive respondent burden again, whereas the ASEC is only one month, it's an annual supplement, and this annual supplement does allow for annual timeseries comparisons. So, just some reference of technical documents again, highlighting Maggie's important comments about how useful these documents can be.

So, why would we want to use the ASEC supplement? So, as we said the ASEC provides very important economic and data that can be useful, especially when looking at tobacco use patterns. So, the ASEC gives you monthly labor force data and data on work experience, income, non-cash benefits, housing, which is the example we are looking at, and migration. And there's also detailed questions on work experience, including employment status, occupation, and industry. There are questions on weeks and hours per week worked, reasons for not working full time, total income, and then also some supplemental income questions. Non-cash benefits include food stamps, school lunch programs, employer-provided health insurance and other types of health insurance, as well as energy assistance. And just a note that the employment and income questions referred to the previous calendar year of when they survey was taken.

Okay, so now, I want to take you through some of the weighting considerations if you plan to try to link TUS and ASEC data. So, briefly I want to talk about – I think you've heard quite a bit about the TUS survey data and with respect to weighting and replicate weights. But I want to highlight some components of ASEC. So, as of 2019 the ASEC data is available on the Census website as either a single ASCII file, text file also known as .dat, or you have the option to download it as the CSV or SAS files that are split into one file each for – you remember the hierarchical file structure, right? Household, person, and family members. ASEC weights account for file structure, there's a supplemental house weight, a supplemental family weight, and a supplemental person weight. And so, the final weight is used to provide population-based estimates for the various items covered in the monthly CPS, but there's differences in the questionnaire, the sample, and data uses for the CPS ASEC supplement that are factored into the calculation of the ASEC weights.

So, when dealing with TUS weights, and again, this was something that I have covered when I talked about the harmonization webinar, but it still applies in this case to using single year data. So, for this 2018-2019 data set, on the TUS website, the TUS data on the website would also include a full sample weight for self-respondents, the variable name is the sample age. And again, just to define this, the sample weights are created to compensate for the differential selection probabilities, nonresponse and under coverage for the entire population of U.S. adults. But there's also the second type of weight in TUS that we have to think about depending on the analysis of the research question that you're interested in, and these are replicate weights. So, these can accommodate various types of analysis that you may want to carry out, and they're created to more accurately estimate standard

errors by accounting the complex survey design. And I've included a hyperlink to the technical document for 2018 2019, the user guide on how to use the replicate weights.

So, reminder here that for this example that I'm going to be talking about, which is using the TUS 2018 - 2019 data, there's 160 replica weights. You may recall from the harmonization webinar that I talked about the prior years have a smaller number of replicate weights that have changed over times, but it's part of the more recent TUS data, we have to use the 160 replicate weights. And so, to run the analyses you first need to merge the TUS respondent, survey data, basically with the corresponding race by the ID variable which I will show you shortly. And so, this ID variable, similar to what Maggie discussed, it's unique within the survey year and the survey month. So, I called up the specific slides that I will cover on this, and we will show, you know, that there is a similar ID variable data that needs to be created for the replicates. And then, we combine the survey data with the replica weights by the ID, which also factors in survey year and survey month. Next question is, do we need to use the ASEC weights? And the answer to that, I think depends on your research question. Here, pointing again to what Maggie talked about, the factor to determine the question should be what you as the investigator can find in terms of the overlap, right? So here I'm going to talk about, with the considerations in mind that, that I've made for the analysis that I'm going to show you, you know, I had to examine what the proportion of overlap was between the TUS and the ASEC with respect to the core general CPS variables, as Maggie showed, also these were sort of the key. The key demographic variables, gender, age, and race, right? So, the consideration then is depending on, back to the point of it depends on your research question. If the proportion is matching in the TUS ASEC samples is proportional, meaning that the overlap proportion is quite small, and you think that might be a problem for your research question, then, users should recalculate a new weight based on using both the TUS and the ASEC weights. And that would be more complicated process that I'm not going to go into and there's also be consideration of whether you need to readjust any replicate weights. But my goal for today is to keep things simpler, and so, I'm going to be talking about shortly what happens when the overlap and the linkage is actually quite good.

Just one more point to consider that when conducting various estimation for the TUS replicate weights, the rates, Fay's method is applied, and Fay's method is a variation of balance repeated replication or BRR. So, depending on which software program you're using, if you're using SAS, then your Fay adjustment factor should be set to .75. If you're using SUDAAN, it should be set to 16.

Okay, so now I'm going to walk through in detail step-by-step how to actually link the TUS with ASEC. I'm going to talk about it first and then I'm going to show you in SAS how I actually get that. So, again, back to the assumption that I'm making, with what example and going to present to you is that in the case of the research question that I was interested in, right? These were tobacco use questions within the TUS with respect to the housing questions in the ASEC. So, the first step for me was to examine proportion of the overlap sample between TUS and ASEC on the key demographic variables that I was talked about to determine whether weight recalculation was more.

So, these are the steps that I followed in order to get to that decision, right? The assumption that I made that the overlap was strong. So, I first merged the January and May 2019 TUS survey data. I was not interested in the July '18 TUS, because of the longer timeline window of the 2019 March supplement, which basically would not capture overlap of respondents. Right? Then, merging the January to May 2019 survey data to the replicate weight files. Combining, right? The survey data to the replicate weight. The next step was to gather the ASEC data, downloaded from the Census website, and merge the household in person files for ASEC using the household id variable. Then, following that,

merging the ASEC and TUS files. And then, as Maggie showed us, creating the set of match variables, right? Using the demographic variables that were the most critical, you know? Gender, age, and race. For both TUS and ASEC to examine what the overlap proportion of those samples was. And finally, I took a look at those results. I created a matched data set, with the final data steps that I was interested in. And that of course, those will be specific to the investigator, right? So, okay. So, let's kind of began with step one that I've highlighted on the top right. So, this is the step one that we, that I showed you on the previous slide at first, we want to link the two TUS time points the survey data, right? So, we begin with a very top line, and sorry this is a bit more in the weeds so I will try to walk through this slowly. But the first line are just some SAS options that you can include so not to get the error message and formats. And you know, how long you want your page to be, and so, the format is just to apply a format to the age variable with the TUS, but importantly, you need to obviously change the library name TUS data to whatever path it is you are following the TUS data two. Obviously, that's my highlight footnote here at the bottom. Before you begin to run this step, you need to download and read in meaning in SAS click on the programs, right? To run the January and May 2019 TUS data. So, once you have the TUS survey data saved to whatever path you're going to store it under, you're running that library name, you're also adding a call out to the ASEC data that you will also download from the Census bureau, which I have included that link farther down on a subsequent slide. And so then, your first step really is to combine to stack the January and May TUS data. And, to subset the records to only give the adult civilians who self-respond, we are not interested in proxy responses here, and to have certain criteria that you are only interested in the IDs for those who have the overlap data. So that's what the next few lines here are going to be, they are going to match with the March data. And the last time here is going to be the ID that is a combination of these other ID variables, plus survey year and survey month. Okay? And so, once we run this step, we can now sort our TUS survey data before we move onto the next step. And I just highlighted a couple of variables that you will see coming back on the next slide.

Okay. So now, we have also downloaded, right, previously the replicate weights for 2019 January, [and] 2019 May. So, you're just pointing wherever you're saving them, they should be this January '19 and the same for me, right? And this macro here is just going to run that data and it puts it sort of the right column headers to make sure that your data is correct. Right? And, once we run this macro, we end up two temporary data sets for January and May of 2019. Which, we can then stack and list the next data step here. And, again, we are creating the same ID variable that we did with the survey data so that we can then combine everything, and everything lines up.

So, this is now what we are doing, we are linking the TUS survey data TUS-CPS '19 with the replicate weights '19, right? And this proc freq here will just give us a nice table to be sure that we don't have any missing records of TUS versus replicate right? Alright, so at this point, we're done with getting all the TUS data in order, and ready for our merge with ASEC so now remember we previously downloaded our ASEC data. We would have included that file name call out on our first line in our program and now we are going to call out the file name with this in file statement. And the input statements are just basically going to assign all variable names to the right column positions. Just a note, here, that I used for simplicity, I ended up using the ASCII text data. And so, this program points exactly to the call if you download the documentation for ASEC you'll see that these variable names and column pointers represent where are those variables are located within the data files, if you don't have the documentation, you will see that these variable names and column pointers represents, you know, represent four those variables are located within the data file. So, if you want this, it should work and should give you the data set with the variables in the right places so that everything reads in correct. Alright, so now, we have, remember, I said the file hierarchy of ASEC household, person, and family, right? And we are not interested in the family, we are only interested in the household data because it's

going to bring the information from the household that we need and the person file which obviously has the response that we want on housing questions, right? So, this is what first set of data steps is going to do first. We are taking the household data that is the HH, right? And we are running back that data step. So, to pull the ID variable that we need, we are going to do the same thing for this person's file, sort both of these data sets, and then merge both to get a final ASEC file. And now we are at the point where we have everything that we need in terms of data, we are going to sort our ASEC 2019 and are TUS January and May 2019. These variables, these are the ID variables that we created early on, right? So now, we are at the final step of linkage, where we have everything ready, and we can download our TUS and ASEC files and select what you saw in Maggie's presentation we can create the match variable to check what are overlap sample ends up being. All that was for was to get us here. And, once we take a look at that output, our last sort of data step is to output a final data set that only includes records where there is a match, right? Between the record in TUS and ASEC for these demographic variables that are key, and we can run proc tabulate at the end to also check those numbers. Sorry, this is just to check you know that we haven't lost any observations with respect to age and year and the person weight. Ans we can now check also on our matched variables for sex age and race to see what the crossover is right?

And we now check also our matched variables for sex, age, and range to see what the crossover is, right? How many people end up in both samples, so as we see we have a pretty good match, right? Zeros are the people who were not matched, so we have .14% of those with respect to age were not matched, and .04% with respect to sex were not matched and similarly for race. So, the missings are just referenced in the TUS that does not fit into the months that we needed, right? So, we only brought in, those are the references from ASEC, not to TUS.

So, our final step in making the decision on how well our overlap sample is, is to actually look at distribution, right? Because this code that I just showed you is just showing you how good the match is, it's not actually giving you any proportions on these key variables of interest. So, what we can do is run some weighted steps on these same variables. Note: These are not the matched variables. These are the actual variables for sex age and race. And we can actually do that for the overlap sample of TUS and ASEC and for the TUS only data, and how well they are, or how different our distributions are for specific variables. So, I'm going to use age as the example, right? And so, this is the, sorry the formatting did not come through, but these are the age groups that we will see similar in the next step that you'll find. But this is the TUS only data set, right? We have 45,000 respondents in January and May 2019 for the TUS only data and for the first age group, the proportion I believe that's 18 to 34-year-olds, so that is 29.8% in the TUS only. And they are 18 to 34 when we output the same variables for the emerging data set, right our sample size has dropped to about 41,000 that's because there's people in the overlap samples the tobacco use used data, right? From the TUS only with an ASEC. But in terms of our matched sample, the distribution still looks the weighted, the weighted distribution still looks good, right? Previous version 29.8, this one is 28 .4. So, still pretty good, pretty close numbers. So that gives us confidence that we are making a correct assumption of this to say that we can use only the TUS weight for the whole reason of doing these checks, to give us validation that we don't have to run off and do a whole recalculation for TUS and ASEC because we are confident that there is enough overlap between TUS and ASEC that we can just use the weight from TUS, since those are our main variables of interest.

Okay, so I hope this was clear, the intent of doing all this work is to really check whether we can correct, make a correct assumption in our need to not have to create any of recalculated weight using the two supplements that were interested in.

Okay. So, finally I am going to walk you through the applied example that I wanted to showcase based on some data that we are about to submit for publication. So, these results are not published yet, but hopefully will soon be. And so, this is a project that was looking at TUS questions on smoking and vaping in home rules and the smoke-free policy attitudes among residents of multi-unit housing versus single unit housing. So once again, this was a linkage of the 2019 ASEC public use ASCII data set to the January and May TUS files. And so here, what I just wanted to highlight of note was that in the first set of columns on current smoking among multi-unit housing we found that current smokers they were about 67% of current smokers who lived in privately rented multiunit housing, 17.6% who lived in privately owned multiunit housing, and 15.5% who live in public multiunit housing and you can compare that to cross other smoking status groups and also among residents of single unit housing. This figure here is showing some results with respect to housing tenure, remember the housing tenure question is whether they are in public housing, in private rented, or in private owned so that's the housing tenure question. And so, the two segments I want to highlight with respect to multiunit housing, the first top left figure is saying that among multiunit housing residents, 10.8% of the adults living in public housing reporting, reported that they had allowed smoking in all areas of their home. That's what the home smoking rule refers to. Versus 5.2% of those living in privately rented housing, multiunit housing, versus 3.8% in private owned housing. the bottom left figure is the question that is asking about the respondents support for indoor smoking bans, right? And so, what this figure is showing is that there is 11.6% of multi-unit housing residents who lived in public housing and who supported allowing smoking in all places in sort of the indoor common area. That is what the question is referring to. Versus 9% of those living in privately rented multiunit housing, and 10.6% of privately owned multiunit housing. So just to clarify here, multiunit housing again does not necessarily refer to government housing. Let multiunit housing can be like a condominium. So, just to note that public housing is going to be sort of a government type housing, but privately rented and privately owned is more likely not going to be. And so, that's the first thing to highlight, here. The second thing to highlight here is with respect to smoking status, and once again, looking at the home smoking rules and support for indoor smoking bans in common areas, right? The top left is saying that 26.7% of current smokers in multiunit housing residents have allowed smoking in all areas of their homes, compared to 4.8% of former smokers and 2.4% of never-smokers. And similarly in the bottom left, 28.2% of current smokers have said that they supported having smoking be allowed in all indoor common areas. And then finally, the last slide is with respect to the new questions in the '18-'19 TUS with respect to vaping home rules. And so, this was also part of the analysis, and similarly, looking at it by housing tenure, 11% of vapers supported in public housing, being able to vape in all common areas and similarly smoking status about 29% of current smokers supported being able to vape in all areas in respect to their homes. So, a lot of interesting data, I'm here to highlight the opportunities that, linkage to the ASEC shows, right? Not only the housing data, but the data on employment, health insurance, so a lot of opportunities for very interesting research questions. And again, as Maggie said, you know, we're happy to support your efforts into your investigations of wanting to explore some of these linkages. Please feel free to reach out to us and we will do our best to walk you through that and provide you code or help you walk through your own code. And finally, just thank you to all those have helped with this analysis, and with the webinar. And that's all I have, and I will take questions.

MS. CORCY: There are no questions yet, but again I want to encourage anyone who does have questions to type them into the chat box.

DR. REYES-GUZMAN: Okay, well, we can also just continue to support your efforts as you think about further questions, or whether this is prompting research topics for you to explore. We can

wait a couple of more minutes to see if there are any more questions coming in. But we hope this has been helpful, and in triggering interest in other linages.

MS. CORCY: I also wanted to offer for anyone who is interested in being unmuted if that is easier for you than typing in the chat box. If you go to the participants panel, this should be an icon on the bottom right to raise your hand. If you do that, I can unmute you if you wish to ask a question.

Okay, I just see one question regarding the slides for this presentation and previous presentations, and I can answer that. So, the slides are available online for the previous, for webinar one and for the second webinar in the series, we will be posting them online shortly as well. Once the materials are available for this third webinar that we just conducted today, we will be emailing everyone with the link so you can access those materials.

Okay, I don't see any other questions or any raised hands, so I think we will go ahead and wrap up. Thank you to everyone for joining us today. Thank you for your participation. We would love to hear more from you about your experience with this webinar series and with this specific webinar. I'm going to share a link in the chat box to a feedback survey. It's a very brief survey and we appreciate if you would fill it out and let us know what you think so that we can improve future webinars for you. If you do have any questions regarding today's presentation or about the TUS in general, please visit our website - the link is here. We also have a newsletter, an email newsletter, and you can sign up for that. That link is here as well for the email subscription. And if you have any questions, for our team members, for Maggie or Carolyn who presented today, or for any other team members, please email us at the email address on the bottom. Alright, once again, thank you, everyone. Have a good rest of your day. Thank you.

[Event concluded]