DR. MIN XIE: Good afternoon everyone. We're finally here. This is the third session for the UMD NCVS Research Forum. And I believe my colleague Dr. Jim Lynch will be in this meeting as well. And we're also honored to have Dr. Heather Brotsos from the BJS. She's the Unit Chief for the BJS Victimization Statistics Unit who's supervising all the activities related to the NCVS. And we're happy to have Heather to start the meeting for us. So thank you, Heather.

HEATHER BROTSOS: All right. Well, thank you, Min. Welcome everyone. And thank you for joining us for the third and final session of the NCVS Research Forum. We're so excited to be here today. As Min said, my name is Heather Brotsos. I'm the Chief of Victimization Statistics at the Bureau of Justice Statistics. I'd like to first thank Min and Jim for inviting us to this session today and organizing all of these sessions. It's been a really great forum and we're excited to share some information today about the NCVS and how to work with these data. So before we get started, I think we can flip to the next slide. Thanks, Rachel.

Just a couple of technical tips here on the screen, if you're having audio difficulties, you can switch over to the phone. There is live captioning. There's a button on the bottom that you can use to turn that on. And then we have a chat box on the side. And what we're asking is that throughout the presentations if you have questions, please go ahead and type those questions into the chat box and send those to everybody, so that everybody participating today can see those questions. We'll see which ones we can address within the chat and then also be kind of saving them up, so that we'll have some time at the end for Q&A and to answer questions. And make sure that we get all of those answers. Also, as you may have noticed, as you came in this session is being recorded. We do plan to post this session as well as the slides and the transcript on the BJS website in the future, once we get everything all organized and ready to go. I will also use this moment to put in a plug for other webinars that we have on the BJS website. I'll put a link to that in the chat. For anybody that's interested, we do have another workshop that today's session really builds on. So if you're interested in learning some more about some of the fundamentals, I encourage you to check that one out as well.

And so on the next slide, we will just go through some very quick introductions. We've put together an all-star team to guide you through the NCVS today. We have some real live BJS statisticians here to talk about the NCVS. These are the guys that are day in and day out working with these data.

So very quickly, I will just, run through we've got Susannah Tapp, who's going to go through some of the—we can maybe go to the agenda slide here. She's going to go

through some of the fundamental background on the NCVS. And then we are going to have Jenna Truman walk through some more information about the data. Grace Kenna is going to provide a demonstration of our brand new tool, the N-DASH and how to create your own visualizations using that tool. Rachel Morgan is going to talk about some of the weighting variables and all of the things that are involved in analyzing data and things you need to keep in mind related to that. And then we will hand it over to Min who is going to do some data analysis demonstrations, talk a little bit about some of her research, and some of the great work that she's been doing using these NCVS data. Erika Harrell and Lexi Thompson will be monitoring the chat and they will be answering questions as they come up and then we'll help to facilitate some of those questions at the end. So that brings us through the two-hour program today. We got a lot to cover. So without further ado, I will hand it over to Susannah to kick us off with the material.

SUSANNAH TAPP: Thank you, Heather. So I'm going to talk a bit about the background of the NCVS. So next slide, please.

The NCVS is currently sponsored and directed by the Bureau of Justice Statistics, BJS. It first started in 1972. And it was then the National Crime Survey and it was redesigned in 1992 and renamed the National Crime Victimization Survey or NCVS. The NCVS is one of the nation's two primary sources of information on criminal victimization at the national level, along with the FBI's Uniform Crime Reporting or UCR Program. Next slide, please.

The NCVS has several goals. First, it measures crime that is both reported and not reported to police. A lot of other measures, including the UCR, only measure crime that is reported to police. Now, this means that the NCVS provides an independent calibration of the UCR program. And the NCVS and UCR can produce different findings and together they present sort of a more balanced picture of criminal victimization. The NCVS also provides a measure of victim risk and it serves as an index of changes and reporting to the police over time. Next slide, please.

The NCVS is administered by the Census to a nationally representative sample of persons aged 12 and older. Individual households are selected and they complete seven interviews once every six months which means they remain in the sample for about three and a half years. The first interview is generally done in person and the remaining interviews can be done in person or over the telephone. Surveys are self-report, meaning, that the person in the survey are asked about criminal victimization that they experienced in the past six months and specific information is collected about each victimization, NCVS. Next slide, please.

The NCVS collect data on violent crimes, including rape and sexual assault, robbery, simple and aggravated assault, personal larceny such as pickpocketing or purse snatching, and property crimes including burglary and motor vehicle theft. Next slide, please.

In addition to towards the core NCVS versus those general crime types, there are short topical surveys or supplements that are administered at the end of the NCVS to eligible respondents. In general supplements are in the field for six months and that's either from January to June or July to December, supplements allows the BJS to capture the changing of landscape of crime. And in recent years, five supplements have been run. The School Crime Supplement is actually in the field right now. And the Police Public Contact Survey will be going into the field starting in July. Next slide, please.

If you would like to find out more about the NCVS, you can visit the BJS website. And you can also now see an earlier workshop that our unit presented in December. Heather's going to put that link in the chat for you all. The codebooks and public use copies of the data are archived at the National Archives of Criminal Justice Data at the link on your screen. You should have received workshop materials, including the technical documentation, the users guide, and a copy of the most recent bulletin criminal victimization 2020. And now, I'm going to turn it over to, Jenna.

JENNA TRUMAN: Great. Thank you, Susannah. I'm going to be talking about the survey instruments and accessing the data. So next slide, please. Thank you, Rachel.

So the flow of the instruments, the NCVS has a couple of instruments to collect data. First is the NCS Control Card, and this is the basic record for each sample unit or each household that are sampled. So we collect household roster and this contains information—demographic information about each individual living in that household. And this includes collecting information on age, race, origin, sex, marital status, and education—and educational attainment.

Secondly is the NCVS-1 or the Victimization Screener, and those contains questions that are designed to determine whether any crimes were committed against the household as a whole. So talking about property victimization or crimes against individual household members ages 12 or older, and we asked about this during that six month reference periods whether or not they experienced any victimization within the past previous six months. Household responders, answer questions about property crime for the household as a whole. And then all respondents age 12 or older answer questions about personal crimes. And the questions are written in what we call a short queue format. And the interviewer reads the stem of the question and then ask the

respondents whether or not they've experienced certain types of incidents and give us examples of short queues. So something—if something was stolen, something from your, you know, your purse, your wallet, your—you know, your locker, and et cetera.

So given, kind of, just judging—or excuse me, jogging respondents' memories. If a respondent did experience a crime, then we would ask the NCVS-2 or the Crime Incident Report, and the Crime Incident Report or the CIR for short is used to gather detailed information about crimes reported in the Screener. One Crime Incident Report is completed for each crime incident reported in the Screener. And so the length of the interview kind of depends on how many crimes someone reported, and how they're answering the questions. The CIR is attribute or incident-based, so questions are focused on the details regarding the criminal incident. And these questions include things like location, injury, reporting to police, offender characteristics, and emotional consequences.

So one important thing to note here is that no crime estimates are actually coming from the Screener questions, so the Screener is designed to use—to trigger recalled events and then the crime estimates that we produced are actually coming from the CIR. So after confirming that an incident was committed, and the details are collected there, those are where the estimates are coming from. Once the—assuming someone experienced a crime, they answer to CIR. And from there, they go into a series of demographic questions at the end. If they hadn't experienced any crimes, the non-victims, they go straight into the demographic questions. And then from there, as Savannah just talked about, if there's a supplement in the field after they're completed with their NCVS court interview, they would then be asked supplements. So for example, right now they're being asked, assuming their students, they're being asked to school [INDISTINCT] and then next in July, they would be asked the PPCS. Next slide.

Thank you. Next slide. So in terms of how to access the NCVS data and statistics from the NCVS there are kind of three main ways to do this. Our BJS statistical reports are relatively new online data tools, NCVS Dashboard or N-DASH for short, as well as our data files. Next slide.

So our BJS statistical report includes those annual reports on the core NCVS, our supplements, as well as some special report topics. So for example, and Susannah mentioned this as well, the criminal victimization bulletin is generally the—is the first release of NCVS data each year. And that's typically released in the fall. And we also do reports on special topics, so thinking like looking at hate crime victimizations, or victimizations against persons with disabilities. And then for our supplement data we

also put out report covering those supplements, like most recently on stalking victimization. Next slide.

So these are just showing a few examples from the N-DASH, our new online data tool. As Heather mentioned, Grace is going to be doing a demonstration later. So I'm not going to go into details here. But the N-DASH shows a dynamic analysis tool that we're very excited about. And it allows you to examine NCVS data both on personal and property victimization by certain victim household and incident characteristics. Next slide.

So NCVS data are also archived at the National Archives of Criminal Justice Data or NACJD for short at ICPSR, and you can access this through the webpage here that we have in the slides. And there's also a link on the NCVS data collection webpage as well. Next slide.

So public—excuse me, the public use data files are released on an annual basis. And the supplement data files are released after those data have been collected. Typically, data are released within a year or so of collection depending on data processing. And the easiest way, once you go to the NACJD website is to find the most recent files is sorting by that released updated column. That way you'll get access to the most recent available there. So when you order by that, then you'll see the 2020 data files from there. Next slide.

The NCVS data are also available at the U.S. Census Bureau's Federal Statistical Research Data Center—Data Centers or FSRDC for short. Researchers can access restricted use micro data through the secure environment there. And these restricted use data have more specific geographic information on where the sample household is located. There is an application process for accessing these data through the RDCs as well. Next slide.

Currently, the NCVS data are available from 2005 to 2017, and the RDC as well as supplement data which includes Identity Theft Supplement, the PPCS or Police Public Contact Supplement, Stalking Supplement or SEM, School Crime, and then the Fraud Supplement. They're all available in the FSRDC. And now, I'm going to turn it over to Rachel to talk about the NCVS data file structure.

RACHEL MORGAN: Thank you, Jenna. All right. Let me make sure—all right. So the NCVS data file structure, there's a lot of information that I'm about to present in this section. So please go back review the slides we'll make those available to you, listen to the recording, send us questions via email if you have questions. But in some of our

recent webinars, we've received some feedback that people want to know more about the data file structure, more about the weighting—the weights that are included on the file. So that's what this section is hoping to teach you. All right.

So there are two options for NCVS data that you can download from NACJD. So we have annual data files or concatenated data files.

So annual data files include an individual year of NCVS data, the codebooks include a lot of detailed information on that individual NCVS year of data. And they can be downloaded in a variety of software format, SAS, SPSS, Stata, et cetera. And they include five files when you download the study. So NACJD calls any data collection within a particular year of study. So the first data file within this study is called the Address Record Type File, this contains information about the household is reported by the respondent, the household respondent, and then also includes some characteristics of the surrounding area that are computed by the Census Bureau or data collector. There are not a lot of variables on this file. We don't use this file a lot, but it is there. The household file is the second file, and this contains more detailed information about the sampled household. The primary use of this file for us is to estimate the number of households in the United States using the household weights, which I'll talk about in detail later. The person file is the third file, and this contains information about each household member that's 12 or older, so ages 12 or older. And the primary use of this file is to generate weighted populations of persons ages 12 or older that's representative of the United States. The fourth file is the incident record file, this contains information from the incident report. So when folks that are sampled report a crime incident, then they would be included in this incident record type file, and includes a lot of information about the incident. And then the fifth file is the incident-level-extract file, so this is similar to the number four incident file, but also includes household demographic characteristics and person demographic characteristics, so that you can look at, you know, a person that experienced an incident and their race or their age, or something like that. So the highlighted in blue, the number two, the number three, and the number five are the most important files and the ones that we use all the time in order to generate NCVS estimates, NCVS rates per person or per household.

So the concatenated files or cumulative files as some folks call them include all years of NCVS data from 1992 through the most recent year, so right now that's 2020. The codebooks include fewer details because there's so many years of data in there. So if you have questions about a specific year, definitely go back to the annual files and download the codebook from that specific year to get more information. The concatenated files are also available in SAS, SPSS, Stata, et cetera, for whatever your programming language is. And then there are three concatenated files that are included

in these data studies. The concatenated household file, which is similar to the household record-type file and the annual file just includes all the household information and household respondent information. A concatenated person file, and then the concatenated incident extract file, which includes the demographic characteristics that I was talking about, along with the incident level variables.

So the concatenated files are what we use in-house a lot, because we're typically in our reports, if you've ever looked at our reports, we're trying to look at crime over time or trends in crime over time. And, of course, it would be much easier to look at all those years at once, run all that analysis at once instead of going back and downloading annual data files. All right.

So thanks to Min for this great graphic. I took this from her so you may see it again later. But it really helps to explain the structure of the NCVS in these different data files that we were talking about. So first, at the top, you have an address. So the NCVS is an address space sample. So we are going to 123 Main Street, we're going to that address, and then we are going to talk to the household that's living in that address. So say my last name is Morgan, let's say the Morgan family lives in this household. There are three people ages 12 or older that are living in this household, and the first person reports two crime incidents within the last six month reference period. Then let's say my family moves out and the Smith family moves into 123 Main Street, then we have household number two, because we're going to go back and interview that new household that moved in. There's one person ages 12 or older living in that household and they report one incident within the six month time period. So I think the biggest thing to know is that, you know, we're going back every six months for three and a half years. So there are seven waves, but we're going back to that same address, even if the household leaves. All right.

So a little more detail about the three—let's say the three most important files within an NCVS study. The household-level file contains information about the household and the household respondent, so we asked each household respondent to be the most knowledgeable person about what's going on in that household and to be 18 years of age or older. If you open up an NCVS household file, you'll see that all of the variables are named V2XXX, so we call them the 2,000-level variables. So that you know if the 2 is there, that means it's a household-level variable. Then we have the person-level file which contains information about each person in the household that's ages 12 or older, these variables are named the 3,000-level variables. So again, you'll know if you see a 3,000-level variable, that that's a person-level variable. And then finally the incident-level extract file contains information about each criminal incident reported by respondents, so if someone did not report a crime incident, they're not going to be

included in this incident file. They would be included in the person-level file or the household-level file. So this is only people that are reporting a crime incident. So property crimes are reported by a household respondent. We consider property crimes to be a household crime. So say someone stole your car, it's going to impact everyone in the household. So we asked the household-level person to give us information on that crime and then just them so that we're not over counting those crimes that are occurring at the household-level. And then personal crimes are reported by all persons ages 12 or older. This includes violent crimes and then personal larceny crimes which are purse snatching and pocket-picking crimes.

So incident-level variables are named V4000 level. And so you would know when you open up—if you just opened up the incident file, you would just see the 4,000-level variables. But if you open up that incident-level extract file which is what I recommend using, you're going to see the 2,000-level household variables on there, the 3000-level person variables, and then also the 4,000-level incident level variables. And that's important so you can analyze victim demographic characteristics, you can look at the household characteristics for property crimes.

So the household and person-level files, as I said earlier, but I'm going to reiterate this, contain records or rows of data for all persons and all households in the sample, even if they didn't report any victimization within the last six months, because we want to know the denominator of all persons or all households, how many of them experienced a crime once we look at that incident-level file. And then again, the incident file just includes the people that have reported a crime. So the—we call the public-use file, the PUF. This is the structure. So annual files include collection year data from January 1st to December 31st of a given year, the NCVS is always in the field unless there is a hurricane, or fire, or some, you know, force of nature, we are always in the field. And the collection year data, so what you would download from NACJD, those files are based on the date of the interview, so not the date of the incident. So say in February 2020, you were administered an NCVS interview, but the reference period is six months, so you could report a crime that occurred in November of 2019. But because you were interviewed in February 2020, you would be included in the 2020 data file. Most sampled persons or households are included twice in a person in a household-level file. So as we've said, there—the reference periods are six months and we're talking to folks every six months for three and a half years. So typically, their first interview, let's say is in January through June of 2020 and then six months later, we come back, you know, through—in July through December at some point and interview them again. So one-seventh of the sample is phasing in during July through December because another is phasing out.

So we always have just like a rotating sample based on what number of interview that household is currently at. And then a little bit about a sample weight, I am not going to cover everything about weighting here, so refer to your Google and your textbooks from graduate school and undergrad. But just a little bit of an overview to help set the context for what we're going to talk about in Min's demonstrations later. So sample weights are adjustments that are applied to sample data to make the data representative of a specific population of interest. So for the NCVS, sample weights are applied to make a sample data representative of the U.S. population of households or persons ages 12 or older. Sample data that are not weighted, so unweighted data, can be biased as a result of who completed and didn't complete the survey. So those unweighted data are not representative and then cannot be used to generalize about specific populations of interest. So we get this question a lot. BJS does not report out on unweighted NCVS data in our reports. We do examine unweighted data to understand what the sample looks like and assess the sample sizes to make sure the estimates are reliable that the sample sizes aren't too small. But all NCVS statistical estimates that you're going to see in BJS reports are weighted to be representative of the U.S. population of persons, of a particular subgroup of persons, or of households within the United States.

So the NCVS has four types of weights, so we have household weights, person weights, incident weights, and victimization weights. And knowing when to use each type of weight is critical for your analysis.

So NCVS household weights. Something to tell you before just to set up how to give you some background on how the files are set up. So each NCVS annual data file is comprised of two six-month data files. And because we're collecting information, a six-month reference period and a six-month interview cycle, let's say. So when NACJD processes the NCVS data files, they have to combine those two six-month files into one annual-level file. So—but because these six-month files are standalone and technically you could use—we could put out six-month files, but we decide not to do that. We decide to use annual files. The weights, the household, and person weights included on these files are adjusted to be representative of the population for that time period. So in order to put the two six-month files together, we have to adjust the weights so that we're not doubling the population. So V2116 is the household weight—the six-month household weight. But what NACJD does when they get the data files and process it for us, they adjust this weight and basically divide it in half, so that we can use this weight to generate the annual count of households when calculating crime rates. So all this will make sense when we do demonstrations. But if we didn't divide it in half, we would have two six-month periods of the full number of households in the country and it would be doubled.

So here's a quick example from an SPSS data file from 2020. We can see V2116, for this case, the weight is 2178, and then they divided in half this weight HHCY is what we call this one. So it's weight, HH household, and then collection year, that's how it's named. So—and see that it's half of what 2116 is and you can see it for all of these cases. So they simply just divided them in half. So when you use the NCVS household weight, you want to use this adjusted household weight.

And then for persons, similar case, B3080 is the six-month person weight. So what NACJD did is adjusted that and made weight PERCY we say, so weight, person, collection year. And they divided it in half so that we can use this weight to generate the annual count of U.S. persons ages 12 or older that's necessary when calculating violent crime rates or personal larceny rates. So again, we can see here the 26,355, this is representing 26,000 people in the country. But then when we combine the two files together, we have to divide it in half, so now it's representing 13,177 people. All right.

And then getting into the victimization and incident weights and this is on the incident data file. So, first I want to talk about the differences between incidents and victimizations.

So an incident is a number of specific criminal acts that involve one or more victims, so, for example, there was a robbery and it was you and somebody else that someone threatened and then took your purse, then it would be one incident, but there were two victims involved. So what the victimization is, is it would be one incident and then two victimizations. So if every incident had one victim, the number of incidents would be the same as the number of victimizations. But we know that's not the case. There are some incidents that involve more than one victim. So if there was more than one victim, the incident estimate is adjusted to compensate for the fact that this incident could be reported by both victims or all of the victims, and we don't want to over count them. So for victimization, this is the total number of times that people or households were victimized by crime, for violence and personal larceny, the number of victimizations is the number of victims of that crime. And then for household crimes, so property crimes, it's considered as having a single victim because, remember, I said, we're just asking the household respondents to report on property crime, so it would just be one household is affected by that crime.

So again, to reiterate, only respondents that report an incident are included in the incident data file. And, you know, it's just depends on your research questions, you just need to think about if you're more interested in learning about incidents or more interested in learning about victimizations. But whatever you pick, you need to use that specific weight in order to get the accurate number of crimes that occurred.

So the NCVS victimization weights, there are a couple of victimization weights on the incident file. There's, remember weight VICCY we call this one. So weight VIC victimization collection year, so this provides the total count of victimizations. But there's no adjustment for series crimes. And series crimes that are crimes that are similar and tight to one another, but occur with frequency that a victim is unable to recall each events in detail specifically and separate them out. So things like domestic violence that occur, unfortunately, many, many times and the victim can't differentiate between different victimizations that occurred. So what we do is we've created a weight called series weight to account for this because there are a few cases each year of these series crimes that occur where there are a high number of victimizations or incidents that occurred, and they're being—they're outliers in the data.

So this week, I have series victimizations is the actual number of victimizations that occurred up to a maximum of 10. So we—our estimates include an adjustment for series weights. Again, it depends on your research question if you don't want to adjust for that. And you want to see and have those outliers be included in your data and see what, you know, kinds of effects they're having then you would use weight VICCY, if not use the SERIES_WEIGHT. So again, down here is an example from the 2020 data, we can see a non-series crime. So this is just one crime that occurred to this person, one victimization, their weight VICCY and their series weight would be equal to each other. But for a series crime, it's not showing here, but when I looked at it, it was 12 victimizations that occurred to this person and so their weight VICCY is a 750, but we capped it at 10, so we just multiply the 750 times 10 and you get 7506 as their series weight. All right, next. All right.

And then incident weights. So again, incidents are looking at the number or not looking at the number of victims involved. We're just looking at the number of criminal acts that occurred. So 4527 the weight provides a total count of incidents, again, there's no adjustment for series crimes here, but we do have a SERIES_IWEIGHT, so series incident weight, the one before was just called series weight, that accounts for the high frequency repeat incidents. And again, it's capping at 10. So we can see an example here, 4527, the weight is 585, they're going to be equal because it's a non-series. But when you're looking at the series 375 times 10, you get the 3753. So again, this is the number of people are represented by these weights. So we're—this one case is representing thirty-five—or thirty-seven hundred people and this one is representing 500, based on their demographic characteristics and a variety of other things that we're not going to talk about today. But—all right.

So in 2020, the household weight ranged from 61 to about 10,000, the person weight ranged from 72 to 17,000, and then the victimization weight—the series victimization weight ranged from 151 to about 77,000. So as I said, in other words, these weighted counts—this weighted count of about 77,000 is 77,000 victimizations could be based on one person, could be based on hundreds of people, it really depends on the characteristics of the victim. And that—well, that's what the weights are trying to do. They're trying to adjust to make it representative of the population. So it's always important to check the unweighted counts just to see what the sample sizes look like, but to report out on the weighted counts, because you want it to be representative of the population.

So we always make sure that the published estimates we're including in our reports meet data quality standards. In certain cases, an estimate may be flagged or suppressed if it doesn't meet a minimum sample size or we may combine it. So you may see in our reports, at times we combine racial categories together and that's because the sample sizes are so small that we can't report out on those individual racial categories, we have to—we have to combine them together in order to report that. All right.

And then unit of analysis, this is something else that we've received feedback on in some previous workshops that we wanted to talk about. So this is just how the crimes are classified and then counted in the NCVS. So property crimes are reported at the household-level. These are burglary, trespassing, motor vehicle theft, and then other types of household theft. So this is from—theft from the housing unit or property, so something, like, someone taking your bike from your front yard, someone taking your mailbox, something like that. And then violent crimes reported at the person-level because these crimes are happening to the person, rape or sexual assault, robbery, aggravated assault, and simple assault. And then personal larceny is reported at the person-level because it's theft from the person's body. So purse snatching, pocket-picking, someone taking something off of you, that is different from having something taken or stolen from your house or your property.

So household estimates are based on counting households affected by the crime, so for property victimizations in our annual bulletin, they're based on the SERIES_WEIGHT, so that's series victimization weight. There are—here's an example of ten households, and there are five property crimes among these 10 households.

So five victimizations, so one experienced four, one experienced one, and the rest of them didn't experience anything.

And then for the victimization estimates, they're based on counting victimizations, so this is each person that was involved in a criminal incident using the series weight. So again, let's say we have ten people and among these ten people, two have experienced victimizations, but they've experienced seven victimizations among these ten people. So we're counting how many times they've experienced something. And then the incident estimates based on the SERIES_IWEIGHT. So say there was one robbery incident, but it involved two victims, that would be one incident. It would count as two victimizations, but one incident. And then to make it even more confusing, we're going to throw in another measure called prevalence that we've been including in our criminal victimization reports for about the last 10 years. And this is based on counting victims. So it's really looking at did someone experience a crime, were they a victim, or were they not? And it doesn't matter how many times they experienced crimes, just whether zero or one, did they experience, did they not experience?

We are not going to get into details on how to generate prevalence estimates, but if that is something you're interested in, please reach out and talk to us later, email, whatever, because it's pretty complicated way to structure the file. The files are not set up to do this currently. So for example, we have ten people here again, it doesn't matter how many crimes these people have experienced, but they've experienced at least one, so they're considered two victims. All right.

And I do think that it's time for a break, so we'll take about a 10-minute break here. In the Zoom invitation we asked you to download workshop files from Dropbox and then to download if you want to follow along and do your own analysis with the ICPSR data, you're welcome to do that. You'll need to download from this website. Just put up this in the search box, 38090. Download those files or you can just walk through what we're going to do, come back and listen to the recording again and run it. But are we okay on time with doing about a 10-minute break right now?

DR. MIN XIE: Yeah, I think so. So we designed this break also just to make sure if you need time to download the dataset, you can do it now. And then we'll be back in a few minutes to start.

RACHEL MORGAN: Okay.

DR. MIN XIE: Okay.

RACHEL MORGAN: So about 1:53 come back, let's say 10 minutes?

DR. MIN XIE: Sure, sounds good.

RACHEL MORGAN: Perfect. Thank you.

DR. MIN XIE: All right. Since we are here, and I'll turn this over to Grace. And Grace, could you do the N-DASH, do you need to share the screen?

GRACE KENNA: Yup, I'm about to do that.

DR. MIN XIE: Okay.

GRACE KENNA: Thank you. Can everyone see my screen?

DR. MIN XIE: Yeah.

GRACE KENNA: Okay. So this is our data tool, the N-DASH. It offers a convenient way to do some simple analyses of select NCVS core data and it gives the option to download tables and images. So for those who knew our predecessor tool, the N-DASH maintains core aspects of that tool. So the two substantive parts of the tool are the quick and custom graphics in these tabs over here on the right, but in addition to support people with using the tool, we have a tool overview which I'll go over some more in a minute, as well as some other supporting information including a user's guide terms and definitions and supporting documents. So we recommend that new users especially review these resources. I won't be able to go over them in-depth today but know that they're there for reference. So, that I am not having my head down as I talk to you. I'm going to stop my video as I do the demo.

So walking through the tool overview, this page covers what a user will find in the N-DASH including the types of analyses that are possible, so here are just the key ones displayed here, multi-year trend, single-year comparison, and a year-to-year comparison. The main crime category is that Susannah and Jenna touched on earlier as well as our main units of measurement including number, rate, and percentage.

So scrolling down a bit further we have two comparison type views, so in terms of views you can compare values as shown on the left here across different types of person or property crime characteristics or you can look at across different values within a single victimization characteristic as shown here on the right. So the two main types of characteristics are incident and demographic characteristic. So incident characteristics include those such as reporting to police or whether the victim was injured and demographic characteristics include those for victims and households such as the race ethnicity of the victim or the household income. So I will turn to quick graphics next. So

quick graphics are a preset graphics that can't be manipulated and we'll look here at two main examples in just a minute. So you see here this row here sort of a table of contents, you can click across depending on what you're interested in. I'm going to look at property victimization. So scrolling down you see this high-level graphic showing the rate of property victimization over time. So you see that as I move my mouse across the screen you can kind of hover over different data points that show some of the key information. So here, the metric is rate, so you can see that. You can see the confidence interval for this estimate as well as the standard error for the estimate. And these thin lines here at the top also just visually display the confidence intervals.

The notes here at the bottom just explain more about things to be aware of as you're looking at the data. And then, we also have for each graphic these features here that you can use. I won't download anything but, so you know, those appear in separate windows as you—as you do that. But just for illustration purposes I'll show the table so all of the data that were featured in the graphic are here in this table for you to just look at if you want to just, you know, have a visual comparison. And then you can also just pivot back and go back to the chart. And again, those both the data table and the image are downloadable.

So looking down here, these are the components of property crimes, so burglary, trespassing, motor vehicle theft, and other household theft. And here again, you can use this hover feature where you can see data for whatever part of the sub-graphic that you're looking at as well as the estimates for the other parts of the graphic for ease of comparison. So those are a few examples under property crime.

Another quick graphic example that I will show is Victim Service Use. So this graphic looks at for select crime types, the percentage of victimizations where victim services were used. So you can see for rape and sexual assault for example these confidence intervals are noticeably larger than the once we were just looking at and we also note that some of the estimates for both rape and sexual assault and robbery are flagged. So those again are just things to be mindful of when you're looking at the data. Okay.

So turning to custom graphics and you can access custom graphics either up here at the top or down here at the bottom. You can go to that part of the tool. So I am going to—for illustration purposes you note that there are six main types of views that you can look at using the tool. And I'm going to start with a year to year crime type comparison of data. So note that there's a default setting for each of these, so the default is the rate of violent victimizations and that's just set up here. So I'm going to change some of these filters. I'm going to change to property up here and you noticed that begins to populate in real time and then I'm also going to change my unit to number from rate. I'll

change my years to 2000 as my first year and 2020 as my second and here for crime type you can either select all or select a subset and I'm going to unselect the ones that I am not looking at here. So I'm going to look only at burglary and other household theft for this illustration.

So you can also filter the data and here I'm going to look at population size, so I'll select that and then within that you can select one of the categories within that variable, so I'm going to look at places with under a hundred thousand people. So, here's my results down here. So the dot for each of these represents whatever the base year is and then the arrow is the comparison year that you're looking at. So note that these are just numerical increases we did not include significance testing in the tool, so that's just something to note. But again here, you can hover and you can see the key components of this comparison here. And then all the other features including showing and downloading the table and image are still there. So I'll just make one change to this to look at some larger areas, and we'll select one million or more and then you can see that the graphic updates with those selections as well.

So looking at a different example of a custom graphic I'm going to go to multi-year and I'm going to choose a characteristic for my comparison. So this time I'll leave most of the default settings in place. Note here that you can also turn your confidence intervals on and off just to show that feature if you don't want them on, leave them off, if you do, turn them on. And then, I will also turn on reference lines, so note that for things that are not applicable you can't do it, so because I don't have any sub-categories of data our reference line isn't relevant. That only comes into play when there's sub-categories within the variable. So for this example I'm going to choose victim age as my comparison characteristic and see now I can turn on the reference lines. So the reference line will just be the overall data. So overall violent crime. You just have that line to show as a reference point what that looks like against the various ages. And all of the other same features apply. So I will also use some more of the data filters here on the right. So I'll select reporting to police as an example, and then I will select for my value "No." So people who did not report to police. So, all of that updates, here as well. So these are just a couple of illustrations to give a flavor of the types of features of the tool. You can also—we don't recommend that you turn the estimate flags off, but depending on your purposes, perhaps that will be of use. But you know, that's another thing that you can manipulate. So all of these, you know, just depend on what you're interested in looking at. So I will turn quickly to the user's guide just to show you that, so that you can scroll down if you like to the different sections. There's also this table of contents, sort of feature that goes into more detail on different aspects of the tool. Just as an example, I'll go down to data consideration. So this is a part of the user's guide that we definitely recommend people review just to see some known kind of challenges

with some of the data years that we've tried to explain here and outline. So those are the main components of the tool. We also have terms and definitions that go over some of the key terms that we use in the tool, just to help make sure that that's clear for everyone. And then we have supporting materials. So we have a link to—a link back to some of these, the user's guide and terms and definitions. And then we have links to the—our main NCVS page on the BJS website, as well as questionnaire. So things that you can kind of access in multiple places, but we include it here just for ease of reference, including our technical documentation and then population counts through 2019 for 2020. Moving forward, we're going to be hosting those directly on the BJS site.

So that was an overview of the tool. So just a couple of things to note about the tool in general. The N-DASH provides the type of core NCVS data that would typically be found in our annual bulletins. And so high-level data for which the data can be supported—the estimates can be supported for a single year of data. So, as Jenna mentioned, some crime types kind of require aggregating multiple years. So that's not the type of thing that you're going to find in the tool. So for things like that, you're definitely going to want to use the public and restricted use data based on your needs. So—also, as mentioned, the tool is relatively new so we're very much eager to hear feedback about how people find it and we appreciate those, such as Dr. Lofton who's already been sharing the tool in some of his classes. And so if you have any feedback for the tool, anything that you recommend or like, any of that feedback is great. Please contact us at askbjs@ usdoj.gov. We'll put that in the chat and let us know of any feedback. So I will turn it back over to Min.

DR. MIN XIE: Thank you, Grace. So today we're at the data analysis demonstration part of the presentation. What I tried to do today, I think, is make sure that we see some of the examples of applying the knowledge we just learned from the several presenters today and to try to replicate some estimates you can get from BJS reports. And, of course, I understand many of you here will be interested in using the data not only just to provide summary statistics from the NCVS but also use that data to do, say for example, like modeling or regression analysis. And what I believe is that in order to do that, you really have to understand the data structure. So what I'm trying to accomplish today is to demonstrate the data structure and how to code the data, how to produce those point estimates, including the standard error for the point estimates. Once you know that, then really it's just another step for you to go from these types of analysis to say multiple regression. And that's what I'll try to do for today. And then, as I mentioned, do make sure that you download the NCVS 2020 data from ICPSR. If you try to use the syntax file that we prepared for you for today, and we ask you to download the data in SPSS format. But, of course, those of you who would like to use other software, such as Stata or SAS, and—you are more than welcome to do that. It's just very easy to switch

going from one program to another. And that's really sort of your choice. And I think Rachel mentioned this important idea is before you try to analyze NCVS data, it's very important that you get familiar with the concept that the data are hierarchical, right?

So we have the address-level data, then you have household-level information and personal-level information, and also incident level information. So often my student will be asking, "Which one should I use?" That really depends on your research question. So, for example, if you have access to the address information, such as—like which county, which state, which census track it's located in, if you have access to the restricted use data and the Census Bureau, then you can study issues such as—like what is the characteristics of the communities and how's that related to crime. So, for example, here I listed some of the reports, and also by—other media statisticians produced a report, so on, for example, the first one is a report that I worked with Mike Planty. There was a BJS statistician to look at violent victimization in areas that has a lot of growth in Hispanic populations, right? So that's an example of studying the community characteristics with the crime. And then if you look on those households, like the example of Smith family replacing the Morgan family, right? So the—those turnover of households within the same address, then you can study, for example, issues such as residential mobility. Like will a crime make people want to leave, and then will the turnover of these households at the same address, will that cause more crime? So issues like that that are also very important, you can look at using NCVS data. And then because we also have information about potentially multiple incidents within the same person, then you can study issues such as the repeat victimization. There is a BJS report published in 2017 by Oudekerk and also Truman. They wrote about repeat victimization. And also another publication by Dr. Jim Lynch and I, we wrote this paper about intimate partner violence, and the victim's decision to call the police, and whether the police came and arrested the offender, and then whether the victims are contacted victim service agencies, and how is that going to affect their likelihood of experiencing future victimizations by the same partner. So these are important issues have clearly theoretical and practical implications. And so I listed these reports in this PowerPoint site, and you can refer to these reports later if you are interested in any issues. And there are many, many other questions, and I can give you a lot of information about—people have published using NCVS data. After you understand the hierarchical linear structure of the data, then it's you can come up with very creative research questions to study.

So, today, the first thing, of course, is to help you understand this very complex level of this kind of data structure, and therefore I'm going to show you what the data should look like. So when you download the data from ICPSR, as we mentioned, ICPSR has this very nice format. It tells you in this particular folder, if you unzip the data, you just

download it from ICPSR and you can save it somewhere in a—in a folder on your computer, right? So it will say ICPSR and then the study number. And once you open up the folder, it will have all the data files and also the codebook to go with the data. And now you should be very familiar with this idea.

Now, we have the 1,000-level variables, 2,000, 3,000, 4,000, and 5,000. As we just mentioned, one thousand will be the address level, and two will be the household level, and three is the person-level files. And then four and the five are incidents, except where the five would have all those household and personal level variables being attached to the incident-level file. And that's the reason Rachel mentioned that we would have two and three and five are the most frequently-used data files to be used. So get used to this concept for the NCVS. And now I'll show you—what I'm going to demonstrate to you is how to make sure you understand the identifiers that would allow you to identify the address, the household, and also person. What this means then—once you understand these identifiers, then you could link the data any way you want depending on your research question.

And so why don't we start by opening up the syntax file that you downloaded. So let me see. You are asked to download the workshop files, right? So when you open up the workshop file, there is a SPSS syntax file called UMD NCVS Workshops Syntax. Open that file. Then you will see this syntax file. Some of you, when you use SPSS, you may be familiar with the click and point. But I like to use syntax, but you can do the exact—the same thing use—using SPSS. It's just a way to organize the files. I already wrote down the notes. Later when you get a chance, you can read these notes very carefully to see what I'm trying to do, but today I will explain the first part of this demonstration is to open up the data files and then you can see what kind of variables are there. And then that will help you to understand how you're going to use these files. So why don't we start by opening up the address file. You can see immediately I had—I used the syntax and get the files from this director and this particular data set. Obviously, you may be saving the data at a different folder. So, in that case, just write down your own directory to that folder that you can use. So I'm just going to use this automatically because I already set it up. So I'm going to run this file. So now you can see I have the address-level file open as mentioned. You don't have a whole lot of variables here because this is public use file. So information such as which census tract this house—this address, it's okay that you would not have. But this is an important file. In a sense, it does give you some information about how these ID variables are generated. When you see the variable label—this is why I like ICPSR. They provide all these detailed information, right?

So you can sort of see very quickly there is NCVS ID for households. Interestingly, this is a variable that's generated by ICPSR which is actually a combination of several variables. And the reason I do want to mention how this ID variable is generated is because it does help you later when you actually try to create your own ID that will suit your research purpose. So you can sort of see in the—as I mentioned, in the syntax file we can read later, there are four variables that are very important. The first one is this sample number. So V1004. And then you have this very important Scrambled Control Number, which is generated by Census Bureau. It cannot give you the original control number because that's confidential information, so they created this Scrambled Number—Control Number to allow you to identify the households without having those restricted use variables. And then another one is the Household Number, and the fourth variable will be the panel and the rotation group, right?

So I just mentioned, this Household ID Number is a combination of those four variables I just mentioned. And to show you why this is the case, if you go back to the data view, you can quickly see this ID Household Number generated by ICPSR. It's very—like a straight variable that's very long which is a combination of the four variables that I just mentioned. So you link all these frames together. That's the identifiers for the household. So what this means is you can actually use these variables to create two IDs. One is the ID for the address and then add the Household Number then you get the identifiers for the household. So this is those useful variables in the address-level file.

And then now let's go back to the syntax file. So now why don't we open up the household-level file, which is the DS2, the 2,000-level files. Let me open up that data set. Okay. You can see now there are a lot more variables and they all have this V2 as a very clear label to tell you, hey, you're using the household- level file right now, right? So it would have those important identifier variables we just mentioned. I mean, they changed the name from V1 to V2 but they are exactly the same values. And you will see it has this ID for the household is exactly the same, like we just mentioned, but now you get a lot more information about tthese particular households. For example, do they own the unit and do they have a phone? Where is the location of the phone? What is the household income? And then also for those who maybe missing information about their income, they have these allocated income information. So this is very important because when they—before they add these values, you might get a lot of missing information. So they provided information so that you can actually—using the income information to study victimization status, for example. So very nice to have these variables in the data set.

Again, even though the ICPSR just gave you the NCVS ID for households, you can use it to create an address by minus the Household Number. That will be the address

information. So that's the household level file. And then more information will be available for the person level file.

Now, why don't we start opening up the person-level file. So let me put it up. Because this is a person-level file, right, and so you should know what the ID for the household we just mentioned, and then ICPSR also created a identifier for persons in that household. And this is essentially would be those important variables we just mentioned, plus another variable called Person Line Number. So in addition to the Household Number, you add a Person Line Number that will help you to identify the specific person. Okay. Now you should make sure you understand what the ID is for the address, what the ID is for the household, and then you have ID for the person. And what this means is now we can sort of see how you can link all these different levels of files together. And, of course, for these household members interviewed, then you would get all those important information such as age variable and marital status. Knowing that, all of these variables would have original value and then the allocated value, and you should use the allocated values for your analysis just because those are verified by data coders and the Census Bureaus and they are more accurate. And so you have—you can sort of see there's a lot of variables about these individual information and then you can choose to find those variables suitable for your own research purposes, okay?

And then the last data file would be, of course, is to open up the incident-level file, we'll do next. But this is the first part of the demonstration is to make sure you understand the correct IDs for all those different levels of files and that would allow you to understand how to analyze the data. I do want to make sure you understand, after you link the records. For example, you linked person records across time, do make sure that you actually check the quality of the matching. What happens is logically speaking when you link the same person across time that you have, for example, the same gender, they should have the same race ethnicity information and also their age information should make sense, it's a three-year period. They should, you know, grow older over time. If you see some very strange patterns about age coming, you know, up and down, that might be some indication about potential errors in the matching. In the most recent NCVS data, the percentage of records that could be matched, the error rate would be very low. But this issue should—could be, you know, the percentage, it could be slightly bigger in older data. And also because you are using the Scrambled Control Number, if the Census Bureau changes the way they generate the Scrambled Control Number, meaning that then you couldn't link data over time anymore. So there are certain time periods, the routine changed. You need to understand that, and that's why it's also very important for you to check the consistency of the records over time and don't just use

these Scrambled Numbers blindly. You do want to make sure that the data you use are good data, okay? That's very, very important. Okay.

So we just did—demonstrate how to understand the ID variables. Let's move on to the next topic. And, today, I'm going to show you two examples of data analysis. Essentially, is that we are trying to, first, calculate victimization rate using—the survey terminology would be generating a point estimate for a very important indicator here, right? But then because this estimate is based on the sample, you also need to calculate the standard error, so you know what's the [INDISTINCT] into what you have for this point estimate, right? And so you can get—for today's example, the first example, we will use the public use level data is to see what would be the rate of violent victimization for Hispanics in year 2020, okay? And if you download the BJS report of—entitled “Criminal Victimization 2020” and if you go check the report, in table six, you can actually see that the rate is calculated as 15.9 per 1,000. And then the standard error, you could get it from the N-DASH or from the appendix table in the report. So the standard error here is calculated as—I highlighted it here, 1.89.

For those of you who are new to the NCVS data, I always encourage you to do what we do today here, that is—you may download the 2020 public use data and then be able to produce the same result. And if you could, that means you really understand the data. Okay. So why don't we see how we're actually going to use the data for this purpose. So go back to the syntax file and then the first analysis example is try to produce that point estimate. And in order to do that, you really have to follow several steps, right? We are familiar with the concept of victimization rates, right? So in order to calculate the rate, you need the two numbers. The first number is to use the incident-level file and calculate what's the total number of violent victimization for Hispanics. And for that, you want to use the weighted statistics but also look at the unweighted sample case to make sure you have enough cases to ensure the quality of the data, right? And so—then you need the denominator, which would be the total number of Hispanic population, which will be a weighted number, right? So when you have those two numbers, you can easily calculate the rate. And then we'll show you how to get the standard error. And all those files would be included in the files to be downloaded for the file that I'm going to show you very quickly.

So now why don't we go back to the SPSS syntax file. Here, let's calculate what's the—first step is open up the incident-level file and then trying to calculate the number of violent victimization for Hispanics, right? So the first step, of course, is to open up the incident file. So once you open it up, you will see a lot of variables. You have the ID for the household. You have ID for person. And then—but, most importantly, because this is a incident-level file, right, so you can see there are a lot of characteristics about all the

incident characteristics associated with these crimes that are reported to the NCVS interviewers. And so I think that's the reason we have repeatedly said this, you really need to read the survey instrument. You need to understand the questions. You need to understand the variables that are available. But for our purposes, I'm going to show you which variables are important for answering that question.

The first one—think about this, okay? In order to calculate the rate of violent victimization, of course, I need to know whether a incident is a violent crime or not, which means then the first thing I have to do is to code the canvas indicator to say whether or not this crime is a violent crime, right? And in order to do that, you—we'll use a variable in the NCVS incident-level file which is called type of crime is V4. So now you're familiar with this variable name now, V4 Incident. And 529, that's the type of crime variable. I'm going to code it and using the recode and say, you know, which crime is it. If it's rape, sexual assault, I'm going to code it as violent—as rape and sexual assault, for example. And then there will be violence and so on. So these are SPSS syntax that actually had been shared very nicely by both BJS and also other researchers to show you how you could code whether or not the incident is a violent crime, is a personal larceny, is a property crime, as Rachel had mentioned.

And so we can run this syntax, okay? I'm going to just run it very quickly here to show you what we're going to get. So out of the 8,000 number of cases at the incident-level file, you will see that we have about 1,600 or so cases that are violence and then the rest of them would be personal larceny or property crime. So—which means now you have this indicator you can use. And then, of course, there's another very important characteristics about the NCVS data is some of the crimes may not necessarily happen inside the United States. And in order for you to calculate the violent victimization rates, you will only count those crimes that happened in the United States. So for that reason, I'm going to use a variable. So NCVS has a indicator v4022 to tell you where the crime happened. And then, I'm going to use that to code another variable to say whether or not this incident happened in the United States. So if you run the syntax I just showed you, then you can see just, you know, the majority of them. Of course, what happened inside the—but there are some number of cases which will be dropped from the analysis, okay? So that's another component.

And then the last component you need to know, of course, is because we only care about the violent victimization for Hispanics, right, so you need information about the victims, race information. So in this particular one, noticing that I have a variable called V3, that's person-level characteristics. And that's the reason why you want to actually use 5,000-level variables because if you use the 4,000-level variables, this variable will be missing. So always use 5,000, okay? So frequency, and I'm going to code whether

or not that is a Hispanic victim. And if you run that syntax, you'll see nicely about, you know, 1,100 or so victims are Hispanics, right? So now you have these needed variables for your analysis, violent crime, inside the United States, and Hispanics, and that would allow you to calculate the total number of violent victimization, right? So how do you do this? And I want to, you know, thank Rachel for talking about the weights. And, of course, you wanted to—since I'm interested in victimization, I'm going to use the Series Victimization Weight. So I weight by this variable. And then I will only count those violent crimes inside the United States by Hispanics and I ask the SPSS to tell me how many total number of victimizations there would be. So if you run that part of the syntax, you will get this number, more than 700,000 cases that are the violent victimization. So this is the numerator you will use for the rate you're trying to get, right?

But as I mentioned, before you use that number, you do want to make sure what will be the number if you don't use the weight, right? So put in the weight off and then you rerun the syntax, and this will tell you the actual number of unweighted. Several cases—it's a little bit more than 200, which is higher than the threshold provided by BJS. So sort of like you wanted more than 10, historically. And so you want to have this number is large enough and you can trust the weighted estimate. So that's the first part.

Now, you will know the total number of violent victimization for Hispanics. And, of course, write down this number. And then the next step is for us to calculate the total number of Hispanic population in the U.S. And for that, now you should think, "Okay, which file should I use?" And the answer is the person-level file, the 03 file. So we should open up that file. And then you need to think about the variables that we'll need for the calculation. And because we want to count the number of Hispanics, then—which means that you will need a variable like we did before saying whether or not this person is a Hispanic person. And so use the same variable and recode it into Hispanic. When you do that, you will see—let me see what I did here. I was using weighted. So stop. And then now what I'm going to do is, after you code it the v_Hispanic, and then you can use the weight variable—weight the variable by the weight person which is also a weight that was mentioned by Rachel. And so having that weight variable, and then you say, "I'm only going to count the Hispanics," = 1, and then run that syntax, it will give you these characteristics, about 48 million or so. That will be the variable to tell you the total number of Hispanic population in the U.S. based on the NCVS estimate. And so what this means is you already know what is the total number of violent victimization. You also know what's the total number of Hispanic population in the U.S., and then you can calculate the rate, which will give you the answer of 15.9 per 1,000.

So when you're at home by yourself you can walk through the steps, and when you can get this particular answer, that will reassure you that you actually understand the data

structure. So that's the—how you calculate the rate, right? But as we just mentioned, having the rate by yourself is not sufficient. You also need to calculate the standard error. And in order to do that, you need to understand that NCVS is a complex sampling designthat uses very complex—which involves stratification. It also involves clustering. All those means that the calculation of standard error could be a little bit more complex than a simple random sample. And so there are different methods you could use. Every single BJS report will tell you what's the specific method has been used to calculate the standard error. For today's demonstration, we'll use something called the Generalized Variance Function Parameters. And we choose that just simply because the Census Bureau produced Excel files that can easily be used to calculate the standard errors. But you don't have to use this particular method. You can use something else, which will be explained by the BJS report or other documents produced by other researchers. And in order to facilitate, like help you understand how you can use these different strategies, we also have this file, I think we already uploaded to the—to the work file called Extra NCVS Replication Examples, right? If you opened up that zipped file, inside, there are several files that would like to point out—that help you. So including NCVS GVF Users Guide, right? Also, like NCVS Variants Users Guide, if you use different software like Stata, SPSS, SUDAAN, and SAS. All these are different softwares. You could use these different routines to help you with your analysis. That's why I really recommend you to read this document in order to understand the data. But then I will show you very quickly about how we actually produce these standard errors in our analysis, okay?

So the workshop file you downloaded, because we just calculated the violent victimization rate, right, and so in order to calculate standard error, you can open up this file called Significance Testing Rates and Percentages. You open up— already open. No. So you open up that file. The first one is what I—what I already generated called Crime Rates, right? And this particular Excel file will ask you for some information. And if you put those information correctly in this file, you will be able to calculate the standard error. The first information this file asks you is, "Are you doing rates or are you doing percentages?" We just said we're doing rates, so we put in number one. And then they asked you, "Do you need something called ROS?" These are essentially year to year correlations in the NCVS. Since we're only interested in a single year, we say, "No. We just use a single year." And then the—as the documented file will tell you, there are several numbers, you need to put in this Excel file including a B parameter, a C parameter. Today, we won't have a chance to tell you what these are specifically, but the document I just showed you will tell you exactly these are—for all these different point estimates, whether it's total victimization, whether it's victimization rate, whether it's, you know, percentages of something, we'll have different equations. But these are numbers—parameters that's supplied by the Census Bureau to the BJS to help with

your calculation of standard error. And how do you get those files as this cert file that's in the workshop file. If you open up that one, you can see very quickly, right, if you go to the— enable editing. Okay. So go to the GVF Parameters. In 2020, you have these parameters. And the first thing you need to know is which one to use. And remember that we said we want—we're calculating violent crimes, so it's the person crime. And then because we are only interested in victimization among Hispanics, that's not overall violent victimization. It's just some parts of the total population. So you use this domain estimate, not the overall one, right? So then you'll see B is this particular number and C is this particular number. And so what that means is you could put those B and C into the cells, and then it asks you, what's the base of the rate, which is 48 million, we just mentioned for Hispanics. What's the calculated rate? 15.9. And then I put it in some written notes. You don't have to, but, you know, I say Hispanics Violent Victimization Rate. And then once you do that, right, then you can see these are based on the equations that are already automatically programmed into the files. So you can see what's the standard error, 1.89, which is identical with the one that's generated by N-DASH. So looks very simple and straightforward, almost like magic. But, again, I want you to read those files before you use those Excel files, because, again, it's like—you know, sometimes if you're—you don't understand what you're doing, it could be dangerous, right? So the—this is the steps that took us to calculate the standard error.

And then the last example of demonstration, we also want to do is to calculate the percent of Hispanic violent victimization that has been reported to the police. In order to do this, think about, logically, that is—first, I know how many violent victimizations there were for Hispanics, which is the estimate number, more than 700,000. And then I need to figure out how many have been reported to the police. You can use it to get this number, which is this highlighted number, more than 260,000 violent victimizations that have been reported to the police from N-DASH. And so your homework is to say can you figure this out by using the downloaded public use file? And I have the syntax written down for the—this workshop. Essentially, what you have to do is to open up the incident-level file. The level—the 5,000-level file again. So you open up this file. You see all those instances that happened in year 2020. And then, in addition to just the code, those very important variables, including whether or not this is a violent crime, did they happen inside the U.S., did they happen to a victim that's Hispanic, you also have to have a variable, which is whether or not that incident has been reported to the police, right? So we could very quickly just replicate those. So this is the one part of the syntax that calculate the violent crime. So if we run this part, we'll show you how many of those is violent crime. And then we'll use the second part of the syntax to show whether or not it happened inside the U.S. You get the same result. And then whether or not the victim is Hispanic. You can run that very easily, get the answer. And then the new part of the syntax is whether or not it has been reported to the police. To do that, you need a new

variable called V4399 in the NCVS. And then you say—okay, read the codebook, understand what this variable is and recode it, and then you will get a indicator tell you whether or not the crime was reported to the police, you can see, which is here, right? So many of them are reported to the police. Some of them—but many of them are not. You also have a little bit of don't know. And as a researcher, you can decide how to handle the don't know. But in this particular case, our important goal here is to weight these incidents by the series weight we just mentioned and then calculate what's the number of those crime has been reported to the police. So around this part, you will get the number that's identical to the N-DASH, about—more than 260,000 cases that has been reported to the police. And, remember, you also need to look at the numbers without the weight. And if we run—take—say, weight off, what will be the number? You get this number 94, which is still larger than the minimum threshold. And so what it means is now you could—you get this—both of these two numbers. You can calculate the percent of victimizations reported to the police for Hispanics is this percentage, right? And remember now, next thing is that you will have to calculate the standard error for the percentage. And when you go to the Significance Testing File, there is another person reported. And so this time instead of saying it's a rate estimate, it's actually a percentage. So put in number two still. It's just one year, so say no, you don't need ROS. And then you need the same B parameter, C parameter, and you need to—what's the base? It's a total of more than 7,000 victimization. And the calculated rate of reporting is 3.—33.8. And then write down the notes you want to. And then the equations are programmed into this Excel file. And then you get the standard error, which is 4.16, as I show here.

So what I just showed you, essentially, is that you need to download those NCVS data at different levels, understand the identifiers that could be used to link across levels of [INDISTINCT] also or across time. And then in order to calculate all these victimization rates or reporting rates, you just need to code those variables needed for the calculation. And then you can use weight to produce the estimates about the total number of victimization, total number of persons, and then figure out the rate of victimization or the percentage of cases reported to the police and so on. It's a very straightforward process. It does require you to understand how to code data. But, more importantly, is to understand the complex sampling strategies in order to do significance testing as required. And once you understand all of these, you might say, "Okay, what's the next step if I want to do regression analysis?" As I mentioned, the Extra NCVS Replication Examples, for example, like these different softwares, has information about—like how you use direct variance estimation. Those are very useful knowledge you can use when you try to run multiple regression rather than what I just showed you using the GF—GVF parameters, trying to figure out what's going on. So I know this is quite a lot of information for you to digest so hopefully later you can watch the recording

again and help you to understand the data set. I think—and then we should—we only have a few minutes left for questioning. Sorry it was too long. And I'll stop sharing to see if we have questions.

ALEXANDRA THOMPSON: Thank you, Min, for that great demonstration, and thank you to all of the presenters. Do you have any—as Min said, if anyone has any questions, please feel free to put them in the chat. Erika and I had been kind of monitoring questions throughout the presentation, so we have a couple that were asked that we can get started off with while everyone is still thinking of any other questions.

So the first question I'm going to post is to Rachel Morgan. And there was a question asked by Kristen Ravi. The question is, "Is the extract file basically a merged file of those household level, person level, and CIR?

RACHEL MORGAN: Merged file? Yeah. I mean, you could call it that, yeah. That they took the incident level record file and connected the demographics from the household and the person to make one file. But only for crime victims. So if you want to look at folks that were not victims of crime, you would have to look at the household file or the person file.

ALEXANDRA THOMPSON: Thanks, Rachel. And then, Erika, did you want to ask the next question?

DR. ERIKA HARRELL: Yes. It's also for Rachel. "Is there a way to filter a person—is there a way to filter if a person reported any victimization across the years they were in the study?" And that's also from Kristen.

RACHEL MORGAN: Across the years they were in the study? Min, have you looked at that more? Is that repeat victimization you would consider?

DR. MIN XIE: Right. So the—all the files, if you just get, they will be by year and quarter or the information is stored that way. And if you try to figure it out whether a person has been experiencing multiple victimizations across interviews then you have to do your own linking.Once you link using the person ID and check and make sure the records are matched, well, then you could do that.

DR. ERIKA HARRELL: Thank you.

ALEXANDRA THOMPSON: And it looks like Kristen might have posed a follow-up to that question. "I use ID per as the ID-to-merge data from 2017 to 2020. I just want to

make sure this ID does not change over time." I'm not sure if that is a follow-up or a slightly different question, if Min or Rachel could—I mean, Rachel or someone could answer that.

DR. MIN XIE: I'm not aware of any changing of Scrambled routines during that time period. What you need to be careful is when they switch from—say, in older years, you—they switched from 2000 census design to 2010 to—in that time period, they may change the routine. But if you match the data set and check the consistency in terms of race, gender, and age, and they are correct for the majority of the observation, I think it's safe to say they are correctly matched. Isn't that right, Rachel? Consistent with your…

RACHEL MORGAN: Yeah.

DR. MIN XIE: Yeah.

RACHEL MORGAN: I would say that too. And I would say, also look at the codebook because it would tell you if those variables changed over time.

ALEXANDRA THOMPSON: Great. Thank you. There's also another question about what is the best variable to use to determine if the victimization is domestic violence. I can quickly answer that one. At least for NCVS reports, like for the bulletin, the Criminal Victimization 2020, we define domestic violence as if it was committed by a current or former intimate partner and if it was committed by a family member. And so there's offender—victim-offender relationship variables in the data set. And so you can use those to filter for incidents for victimization that would be considered domestic violence. I don't remember the exact ones off the top of my head but a codebook—the codebook would be a good resource for that.

RACHEL MORGAN: One follow-up. We did include a bunch of our common recodes that's in that Extra NCVS Replication Examples file and the victim offender recode is in there. So you would just copy that syntax into SPSS and just run it all, and it'll give you all those victim offender categories.

DR. ERIKA HARRELL: Lexi, did you want me to go to the next question?

ALEXANDRA THOMPSON: Sure.

DR. ERIKA HARRELL: Okay. I have a question about regarding missingness, i.e., household income in the data set. "Do you have any recommendation- or advice to deal

with missingness?" I can sort of answer this with regards to income. We actually impute our income variables now. We've done it—we have imputed data for household income since 2015, in terms of—on household income. But in terms of other variables, I'm not sure that we deal with—we have any real recommendations on how to deal with missingness. They're just—they're in the file. We really kind of don't—we try not to tell people, give them advice on how to analyze their data or things like that, so…

DR. MIN XIE: Yeah, I concur. The household income now is complete information, right, so—because it's the imputation. That's really helpful, because the percentage of missing income could become to—a big problem. But in terms of other variables, NCVS data have actually very small percentages missing. A lot of times people say there are a lot of missing data, it's because they didn't notice the skip patterns. So if they—if they code the data incorrectly, they might say, "Oh, you know, a lot of missing data." But if you code correctly, my experience with NCVS' other variables have very, very small percentages of missing. It's not a huge concern for the NCVS data. Some of the, you know, sensitive information but it's—most of the information is not.

ALEXANDRA THOMPSON: Thank you, Min. We have now hit 3:00. I think there were a couple of questions we did not get to in the chat but we can try and follow up with anyone separately or feel free to email the askBJS email that Grace posted in the chat earlier if you have any further questions for any of our presenters. And maybe, Min, if you want to close it out.

DR. MIN XIE: Oh, I think Heather, because I really appreciate the effort that—by all the staff members in the BJS to help with this effort. So I would give this to Heather if she wants to say something about the workshop.

HEATHER BROTSOS: Sure. So just thank you once again, Min and Jim, for inviting us to this session. I think we really enjoy talking about the NCVS and sharing our expertise, so—thanks to everyone who came out today to learn more. I'll put in one more plug for the webinars that we have been doing. We're going to post this one on the BJS webpage in the same place where I posted the previous ones that we've done. We're also planning to offer a workshop before the ASC Conference this fall. So if you're planning to attend that, keep an eye out for information about that. And also check out our website, askbjs@usdoj.gov if you have more questions. We look forward to talking to you more about NCVS data and excited to see all the ways that you all use this data in your research and excited to hear about your successes with this as well. So thank you for coming out. Thank you again, Min, Jim. Thank you to all our presenters and we hope to see you soon.