AMANDA WILMOT: Thank you for standing by. The webinar will now begin and is being recorded. You are in listen-only mode. There will be a Q-and-A session at the end of the presentation. I'd like to give a warm welcome to all those attending the National Center for Health Statistics webinar on the design of survey questions about COVID-19 and their administration during the pandemic. My name is Amanda Wilmot. I'm a survey methodologist in the Collaborating Center for Questionnaire Design and Evaluation Research, CCQDER, which forms part of the Division of Research Methodology at NCHS. It is my pleasure to be hosting this webinar today. The challenge for designing any data collection instrument is to design one that provides accurate data and the ability of the instrument to measure change over time is of particular importance for public-health surveillance. When developing questions on emerging topics, such as COVID, understanding the way in which the general public interpret our questions within the fast-changing context of the COVID-19 pandemic is paramount to helping us effectively operationalize those questions. The use of question evaluation and pretesting methods is now firmly entrenched in research that involves the use of questionnaires with cognitive testing being the dominant method used while increasingly researchers are combining methods to complement and expand on the information obtained about how those questions perform through the use of quantitative survey data and web probing techniques. This webinar covers both findings from cognitive and quantitative question evaluations and addresses the challenges of developing good COVID-19 survey questions. So I'll start today by introducing a stellar lineup of senior staff from CCQDER who will be presenting. First, I'd like to introduce Dr. Kristen Miller. She is CCQDER's director. Dr. Miller is extensively published in the field and co-editor of two survey-methodology books, "Cognitive Interviewing Methodology" and "Question Evaluation Methods." Her writings are focused on question comparability across respondent groups and the improvement of evaluation methods for cross-cultural and cross-national testing studies. Throughout her tenure at NCHS, she has led collaborative international testing projects, including with the World Health Organization and the United Nations. Next, I'd like to introduce Dr. Stephanie Wilson who is also co-editor of the book "Cognitive Interviewing Methodology." Dr. Wilson has focused much of her career on the advancement of cognitive-interviewing methods and the empirical study of question performance. Throughout her tenure, Dr. Wilson has worked on a wide variety of topics, including survey measurement of opioid use and COVID-19 experiences, as well as administrative data collection such as blood and organ donation. And last but not least, Dr. Paul Scanlon is the project lead for the division's research and development survey, RANDS. He has also worked on a range of topics, including the development of vaccine hesitancy questions, which have been adapted for use during the COVID-19 pandemic. Dr. Scanlon's research interests include mixed-method survey evaluation and the development of web-probing methodology, an area of pretesting in which he has become a leading author. So a few bits of housekeeping: All attendees today are in listen-only mode. We are due to finish at 3 p.m. prompt. There will be some time for questions at the end of the presentation. Please type your questions into the Q-and-A and I'll try and ensure that as many are dressed as possible, so just try to keep them clear and concise, where possible. That would be appreciated. That said, without further ado, I'll hand you over to Kristen Miller.

KRISTEN MILLER: Okay. So thank you so much, Amanda. And thank you, everybody, for being here with us. It's really a pleasure to be able to show our work like this. So this is how we've organized ourselves for today. The first thing -- I'm going to give a quick overview of our program and then we're going to dive into our COVID-19 work. Stephanie Wilson is going to be talking about our cognitive-interviewing study. And Paul Scanlon's going to be talking about some work we did with our research and development survey. And then like any good presentation, I guess we'll be talking -- summing it up with lessons learned. Okay. So we are the Collaborating Center for Question Design and Evaluation Research. And the main thing that we do that we focus on is question evaluation. So what does that mean? What is the focus here? So, very basically, we want to understand how do we -- respondents understand the survey question? Do respondents understand the question the same way? Does the question mean the same in all the different culture and socioeconomic groups that it's asked and in all the different language that it's asked? Because I want to say that one thing that we know for sure is that when respondents go about answering survey questions, they come from it from their own life experience -- so necessarily gender, race class, all of these kinds of social, structural things that serve as filters for how people understand the survey question. Then, even more to the point, we want to know, to what extent are the data -- to what -- hold on -- let's -- hold on. We're having a little. Oh, so we want to know, to what extent are the data elicited from the question really a true representation of what we want to know? And then in what ways is our picture -- the way we are seeing things is distorted? Because we aren't realizing that our questions aren't capturing what we -- what we think they are. And then I think, even more importantly, what important discovery are we not making because we are aware that our picture is distorted, because we have poor data quality? So how we're organized is a sort of -- (inaudible) -- in three parts. First, we are a methodological research program. So we study the question-response process, specifically in relationship to data quality. This is where we get into looking at how people from different

sociocultural groups, how -- interpret questions differently. And then we're also -- we have a question-design research lab where -- in our building, we have a facility where respondents can come in and we can do these question evaluation projects. A lot of this, this is applied work. A lot of the work we do is through reimbursable interagency projects with various CC programs. And I'm hoping that there are a number of you that are out there that we're already working with. And then we also -- we also do a bit of developing question-evaluation applications. And one that I specifically want to draw your attention to is Q-Bank, which is an online library of question-evaluation findings that we have. And I want to say also that Stephanie Wilson's cognitive-interviewing report, that, she's just going to be summarizing today, but it's located on Q-Bank and we'll follow up with that at the end. So, just topics -- we cover a lot of different topics. Here's just a few examples. Also, we use a lot of different question-evaluation methods. And here you can see this is a book we edited that you can check out. But more -- we're mostly known for our work in cognitive interviewing. NCHS was the first cognitive lab at -- in the Federal Statistical System. And our focus really at NCHS has been to evolve the cognitive-interviewing methodology. So it's a bit different from typical cognitive interviewing studies that you may know about. We have evolved it into more of a scientific study and thinking of it, seeing it more as a validation study. I'll get a bit more into that. Right now I'll get a bit more into that. So you can think of analytic goals for cognitive-interviewing studies. This is very basic -- is to conduct interviews with respondents just to look for problems. And this is probably what -- how you understand what cognitive interviewing is. It's basically, if somebody didn't understand this question, oh, they couldn't remember this thing. So you're kind of making an inventory of all the different kinds of things that go wrong. But what we do at NCHS is -- and we definitely do that, but what we're -- what we do is more of a construct validity study. So, specifically, we're identifying the constructs that are captured by the individual questions by identifying the specific things that respondents are thinking about. So what goes on in their mind, what their -- what -- how they're responding to the question, those -- that particular phenomenon, that is what the question -- that is what the variable is going to measure. That is the construct. So then you can take it even further, which we do -- is -- see it as more of a comparability study. So where you can determine where the different -- where different -- where respondents might be thinking of different things and not -- so not consistently capturing across the different groups. We've also expanded our cognitive-interviewing methodology to have a quantitative component. And Paul is going to talk about this in a little bit and how we used it for our COVID studies. But now what we're going to do is we're going to turn to thinking about -- to looking at how we went about studying COVID, or question design and question evaluation during the COVID times. And just to get us set, just everyone -- it would be great if everybody could think back to where they were last March and how everything just so dramatically changed. All of a sudden, we were -- we were not going into work. Some people were unemployed. Life was in an upheaval. So, and then, with all that, we had different, conflicting understandings of the virus. We had new vocabularies coming in and -- but, then, we -- also at CDC, we had this need for quick, fast data collection, which posed problems for us. Because we were now at having to ask new questions that we have never asked before and obviously haven't been able to test. And even more so problematic is that questions that we knew operated one way were now operating a different way, because we had a different -- a new context. And now I'm going to turn it over now to Stephanie Wilson, who's going to talk about our cognitive-interviewing study, and just giving highlights from this work. And, again, I'll -- we'll come back and I'll show you -- we'll show you where to get an actual copy of this work.

STEPHANIE WILSON: Okay. So thanks for that introduction, Kristen, to kind of set people up for what we do. So as she suggested -- okay. I'm going to talk a little bit about the cognitive-interviewing study. Hold on. Okay. I'm going to jump ahead of myself. Okay. So here's kind of an overview of where we work. So, as she suggested, the study's qualitative. They were one-hour interviews that are in-depth, one-on-one interviews. And, even though we often do these in person, because of the pandemic, in particular, they were all conducted virtually over Zoom, in particular. And the way that we conducted the interview itself was that first we administer the survey question to the respondent. And then we do follow-up probing to understand how they answered the way they did and why they answered the way they did. And all of the interviews took place between September and November of 2020. And we got a total of 50 interviews. It was the purpose of sample design. And in addition to demographic diversity, we also wanted to make sure we got a variety of people, some of which had never been tested for COVID-19; and then others who had been tested. And then, among that group, we wanted some who tested negative and some who tested positive. In terms of the demographic background, it was a pretty diverse sample in terms of gender and race. So I'm not going to go over this, because if you wanted to, you can get the report in Q-Bank -- but also in terms of age and education, a variety of people, different educational attainments and age. I'm not going to list all of the questions here because it would take too much time, but -- just to give you an idea of the topics that the questions covered -- we had COVID-related work and health-insurance interruption, different questions related to access to healthcare and including things like telehealth or telemedicine, any symptoms that people had, and questions related to COVID-19 testing, like the type of tests that they

had, the results that they were given and whether they quarantined before or after testing. And then we had a couple of psychosocial effects of the pandemic like being sad or lonely or feeling disconnected from people. So, like I said, I'm not going to go question by question to talk about the findings. But, instead, what I'm going to do is highlight the key themes of the findings that we had and there's essentially four. The first is the pandemic frames the question intent. The second is people had to fit new phenomena into existing frameworks. The third is the evolving timeline of the pandemic. And the fourth key finding is that there's new vocabulary or new meanings for existing vocabulary. So let's take them one by one, starting with the first: that the pandemic framed the question intent, right? So for -- in other words, every question was kind of filtered through people's experience of the pandemic. And for many questions, that's fine. But, for some questions, they were not meant to be related to COVID-19. So here's an example of what I mean. There was one question that asked: which of the following symptoms have you had at any point in time since March 1st, 2020? And then you can see that there's a whole list of symptoms. It was a mark-all-that-apply question. And the thing is with this question -- it was intended to capture any and all symptoms that they had. In other words, it wasn't just supposed to be a measure of COVID-19 symptoms. However, many people reported only symptoms that they judged to be caused by COVID-19. And so, in essence, for many people, the question functioned as a prevalence question. Some people thought that it was kind of trying to ask, did you have COVID-19? So here's an example of a quote that illustrates this, one person who said I would have thought it was asking literally have I felt those things, but maybe I also would have been self-conscious, because, like, I know that my symptoms are not all COVID-19 related. Some are pregnancy related. I would have thought the question was trying to determine whether or not I might have had COVID-19 and just not known it. So this is -- this is a situation in which people are not going to report symptoms if they didn't attribute those symptoms to COVID-19. Okay. The second key finding -- finding new -- fitting new experiences into existing frameworks. So, in this situation, I'm talking about new organizational structures for healthcare delivery that popped up during the pandemic and new types of healthcare workers that popped up. People had to take this into account. So here's a question where this happened. It was, has a doctor or other health professional ever told you that you had or likely had coronavirus or COVID-19? And so what people had to think about here was, well, who counts? And so what happened is that some healthcare workers got excluded from this, because there was this new thing going on, like contact tracers or staff at drive-up testing sites. For example, one person said, it was just the testing site I went to. I'm not sure if they're doctors or not. The third key finding -- evolving timeline, right? As we all know, if you think about your own experiences, these -- our knowledge and experiences have undergone changes over the course of this pandemic. And judgments have to be made on which time frame to consider when you hear a question. So here's one example of what I mean, one question that read, did you seek medical care for coronavirus or COVID-19? And another one was, did you suspect that you ever had the coronavirus or COVID-19? And so answers to this question are contingent on timeline. So here's two ways this could go. One is that a respondent may have sought medical care, but at the time they didn't think it was for COVID-19 until they tested positive later. Or, conversely, they may have originally suspected they had COVID-19 until later they tested negative. And so the judgment here in the question-response process is, do they answer based on their understandings prior to or after testing? And so here's how this could go, for example, right? So for the question, did you seek medical care -- there were false negative answers to this, right? So one person who said, I thought it was just like a cold or a flu or stomach virus or something. I didn't know about all the symptoms of coronavirus. Or another person who said, when I went to seek medical care, I wasn't feeling well. But I assumed it was a cold and sinus. I didn't go thinking I had COVID-19. And similarly false-negative examples for the did-you-ever-suspect-you-had-it question, right? One person said now I don't because I tested negative for it. I mean, obviously I did have COVID-19 but I tested negative -- I mean, I thought I did have COVID-19 but I tested negative for it. So obviously I didn't. So those two questions, then, are not are not capturing the people we want to capture. And then the last one here -- new vocabulary or new meanings for existing vocabulary. So examples of new vocabulary -- things like telemedicine might not have been exactly new but it was new for a lot of respondents. This wasn't a common word. Social distancing was an emergent word that was new vocabulary. And then you had new meanings for existing vocabulary, things like a pandemic or COVID-19 testing or employment status or quarantine. I'll just -- I'll just focus here and give you an example on that last one, quarantine. We had a question that said, have you isolated or quarantined yourself because of the coronavirus, yes or no? So this quarantine word -- here you have a familiar word but it takes on new meanings, potentially, right? So there's two ways to go on this. We did have some respondents who kind of had a more of a strict medical definition for this, if you will. In other words, they thought about a person who was isolated from the rest of society and in fact, even from the rest of their family, in a bid to stop the spread of COVID-19. But then you had other people who interpreted this question and this word as asking about the new norms of pandemic-appropriate behavior, things like mask wearing, hand washing, staying home when possible, and social distancing. So here's an example of what I mean. And what would happen is you would get false positives on this, people answering

yes when probably the answer should have been no. And here's some examples of what I mean -- one person who said, I was isolated or quarantined because of the fact that we asked everyone to not come into the office and work remotely as we figure this out. Or someone else who said, me and my husband, as things completely shut down, we stayed in our house. We didn't go out anywhere unless we, unfortunately, had to get to get some groceries or some important stuff. For the most part, we wore face coverings and stayed inside. And lastly, only leaving the house when necessary, consistently washing my hands, and making sure I'm clean and wearing a face mask when I do exit the house. So the bottom line is, then, that the data quality that might come out on these survey estimates is that these things are capturing things that we might not realize it's capturing and it might be missing things. And what's more important is that the error -- the measurement error here is not random. Rather, it seems more systematic. It's patterned in ways that are theoretically meaningful. That means it kind of -- it renders the possibility that statistics that come from this can have bias. But, at this point, I'm going to turn this over to Paul, who can talk a little bit more about a quantitative approach to evaluation.

PAUL SCANLON: Thank you, everybody. Thanks, Stephanie. So I'm going to take a few minutes and talk about another question-evaluation approach we take at NCHS, which is our Research and Development Survey, RANDS. So I'm first going to give a brief overview of the RANDS program and the specific RANDS during COVID-19 series that we conducted. And then I'll jump into some of the question-evaluation findings. Okay. So NCHS has been conducting surveys in the RANDS program since about 2015. And, to date, we've run seven separate rounds of data collection. What really sets RANDS apart from other NCHS programs is the fact that we use commercial-survey panels. They collect our data. And, to this point, we've used two separate panels. The first two rounds of RANDS were conducted using the Gallup panel. And then, since then, we've used NORC's AmeriSpeak panel. So RANDS was originally designed to be the center's methodological-survey platform -- so somewhere NCHS researchers could perform, research, and investigate the properties of commercial-survey panels. So, at a high level, we break this methodological work down into two areas. So, first, we use RANDS to conduct studies exploring measurement error and evaluating questions. And that's obviously what we're talking about today. But there's another group of NCHS researchers that uses RANDS to explore how commercial-survey panel data can be used to produce or supplement the health statistics that NCHS releases to the public. So a term I've used a few times so far has been commercial-survey panels and I don't want to get too far into the weeds but let me just take a second to explain what I mean by that. So these are standing panels of potential survey respondents that commercial firms recruit, maintain, and then they contact for service. Most of these panels are web-based, though some, such as AmeriSpeak, include phone respondents as a way to mitigate some coverage bias and make sure we can get more people into those surveys. I should note here that even the high-quality statistically sampled panels, such as AmeriSpeak, do have major data-quality limitations as compared to NCHS's traditional household surveys. And these include, you know, things like sample size, nonresponse and coverage bias. But the one major upside that these survey panels present is speed. So we're able to develop and field a survey in a fraction of the time it takes to field one of our traditional household surveys. And, at the beginning of COVID, when we really wanted to collect some information not only on how questions were formed but actually collect some information about COVID outcomes, we were able to use these panel surveys. So that leads me into the RANDS during COVID-19 series. This was a special set of RANDS surveys that we conducted during and due to the pandemic. So, at the beginning of the pandemic, many of NCHS's other surveys were impacted by things like the stay-at-home orders. And so we decided to run a series of RANDS focused on -- directly on coronavirus-related topics. And that would not only allow us to do the important question evaluation and development work that we're talking about today but also, as I said, release some experimental estimates, which you can actually find on the NCHS website. As with other recent rounds of RANDS, these were conducted using NORC's AmeriSpeak panel. But we also actually in this case included data from a supplemental opt-in panel just to give ourselves some extra sample. The opt-in data wasn't included in the publicly released experimental estimates, as you can find on the NCHS website. Though I am including it in the evaluation findings I'll be showing you a few minutes. There are three rounds of rounds during COVID-19, two of which were conducted last year. And the third is in the field as we speak, actually, and we hope and -- hope to have the data from the third round in the next week or so. There's a universe of possibilities for using commercial-survey panel data for question evaluation. But today I'm going to focus on two approaches that RANDS and the quantitative or mixed methodologies that stem from this work allow us to do. And I'll give an example of each. So I'll start by showing how we used RANDS during COVID-19 to conduct a head-to-head comparison of questions asking about COVID-19 testing and then show how we leverage the large sample size of RANDS to do some subgroup analysis of these findings. One of the most useful things we can do with a methodological survey is compare questions head to head, quote, unquote, which is -- so basically using experimental design and getting half the sample one question, and the other half a different version, and then comparing the outcomes. A good example of how we did this in the COVID case was actually

COVID-19 testing, which is something Stephanie just mentioned. So, in the first round of RANDS during COVID-19, we asked a question about whether people had been tested for COVID. And then we followed up with an open-ended probe to try to tease out what kind of tests they were thinking about. So the first round of RANDS was actually in the field before the cognitive interview, as Stephanie had just mentioned, before we were able to actually do those cognitive interviews. So we really had no idea how people were interpreting some of these questions. And so we used open-ended probes to try to collect some qualitative information about how people understood questions. And when I say an open ended probe, this is basically a text box where you would say, in the previous question, what were you thinking about? And we would just give them text to -- a text field to open and to write in. And so, in that open-text data, for the COVID-19 testing question, we saw that some people who said they had taken a test were counting things like temperature checks and symptom checklists. And they were counting those as COVID tests and that's obviously measurement error. We thought that this measurement error or the potential for measurement error may have been related to the format of the question. So, in the second round, we decided to test an alternative approach. So here you can see the two versions of the question we tested in the second round. The first version, which is on the less -- on the left, is just a single question about whether or not a person had been tested. And this is also the version we used in the first round. The alternative version was a two-questions set with more specific questions about the specific types of tests and you can see that on the right. So half the sample in the second round received the version on the left and the other half received the version on the right. So, after running this experiment, we can see, you know, that there was no actual -- no statistically significant difference in the prevalence of testing across these two versions. But what's really cool about what we do in RANDS is that prevalence isn't the only outcome measure that we can look at. So here you see a web probe that we administered to all respondents who said they had tested for COVID using either version. And so basically anyone who said they had tested for COVID received this probe. And what we're trying to do here is suss out what specific activities respondents were, quote, unquote, counting when they were answering the COVID-19 testing question or questions. And so you'll note the two of these answer categories -- answer category five and six -- are the out-of-scope categories that we saw in the open-text data from the first round. And so people who indicate these answer categories on the probe, we would say that's a potential for measurement error. So, given that we administered this probe to all respondents who said they received a COVID test, we were able to compare the prevalence of these out-of-scope patterns across the two versions of the question. It turns out that we didn't see a significant difference in the amount of out-of-scope response between the two versions. So what we'll say from that is on a kind of a first-level order here where you don't see one question or the two-question approach producing any more measurement error than the other. However, given the large sample size of RANDS, we can take this analysis a step further and look at whether or not we see subgroup differences. So, in the evaluation context, what we care about in terms of subgroup differences -- and this is something that Stephanie kind of mentioned at the end of her section -- is largely whether or not respondent or groups of respondents understand the same question in different ways. And if so, measurement error can be concentrated in specific ones of these groups. And that can have a massive impact on the reliability of any eventual estimates that come out of these survey items. So the advantage of doing subgroup analysis using RANDS data is that we typically have a large-enough sample size to be able to detect significant differences. And depending on the data that we're using, we can make inferences to the population. So, typically, with both cognitive interviewing and with more quantitative question-evaluation techniques, there are specific subgroups that our experience and, you know, the scientific literature indicate that we should focus on, things like gender, race, age, education. But it's important to note that we could look at groups based on anything that we have access to in the data file -- so health differences, geography, and anything like that. So, just to briefly show you some results, first, here is educational attainment. This is the percent of respondents who use the out-of-scope interpretation by educational attainment. And we can see that respondents with a high school or -- a high-school degree or less use the out-of-scope patterns significantly more than respondents with at least a high-school degree or -- I'm sorry -- with more than a high-school degree. And now here we can break this down further by looking at that COVID-tested version. So we can see that -- the same general pattern within each question format near the one- or the two- question approach. We can see the respondents with lower educational attainment use the out-of-scope patter significantly more than respondents with higher levels of educational attainment. And while I'm just showing you educational attainment right now, today, you can see that we would -- we could do the same sort of analysis, you know, based on other demographic categories, like age or race or health conditions or people that have usual access to health care or not. So doing subgroup analyses like these allows us -- allow us to get a much better resolution look at how survey questions reform and how that may impact the validity of the estimates derived from these questions. And, with that, I'll turn it back over to Kristen.

MS. MILLER: Okay. So, just to -- I want to sum up before we ask some -- go through some questions. I hope the one thing that you take away is -- from this if you're going to take only one thing away -- the idea that a good question

obviously has to be relevant to the research agenda. But it also has to be relevant to each potential respondent's experience and knowledge. And the only way that we can really know that is through studying, through studying how people understand the questions. Otherwise, what we're going to be getting is a question that's only relevant to people who write survey questions and that also -- the shifting context of the pandemic really affects the question performance. And, then again, bottom line is that question-evaluation studies really ensure that intended constructs are captured for all respondent groups. And it's just really not okay -- even though we're in a situation where we need to crank out -- crank out the data, we really need to be thinking about how we're going to do question evaluation during this type of response. And so we have -- a couple lessons or notions that we think going forward for rapid response is that -- is three things. First, to integrate data collection with question evaluation. And so, to do that, for two things, first off, you want to be able to interpret the question or to interpret the data correctly. But then, also, you're going to be asking these questions again. So this is going to improve -- you can improve the questions on the next round. It's important to leverage mixed methods, as we did with the cognitive interviewing and our RANDS survey. And, then again, using post-survey question evaluation. A lot of times we think, oh, question evaluation is just pretesting but it's not -- it's not a waste of time to do question evaluation, even after you have the data collected, because this, again, is going to help you to be able to understand how to accurately interpret the data. Okay. So that's -- oh, let me just say, this is our -- the Q-Bank, which, again, is our online library of cognitive-interviewing reports. And you can find our cognitive -- the COVID 19 report is there. And this is the information to that. Questions?

MS. WILMOT: So thank you very much, indeed. That was really informative. We're now going to be accepting questions through the Q-and-A function in Zoom. And we have a few coming in right now. So the presenters will put their cameras on and we can begin. Some comments saying how interesting and useful the presentation was. I agree. And a question asking if there are standardized questions that are being developed by or have been developed by NCHS related to COVID. You know, if they're creating a questionnaire themselves, can they -- how can they find out what questions we might be recommending or not?

MS. MILLER: So it -- traditionally, it hasn't been in our job scope to come up with standardized questions. What we have done is -- and maybe we want to rethink this for that -- going down the future. But what we do do is -- again, is we test these questions, we -- in these reports, you will see specifically what the -- each individual question captures. So what we would like to see people do is to get on, read these reports, see, is this the question? Is this question capturing what I need it to capture? And then you making the decision, yes, I'm going to go with this question. So, again, not anything standardized but information provided to us so that you can, you know, choose the question that is best going to fit your research question.

MS. WILMOT: And sort of along similar lines, someone was asking about, you know, in a rapid-deployment situation, maybe there's not time to conduct a full evaluation. You know, if we're looking at pop-up vaccination activities or something like that, what would you recommend in terms of trying to ensure at least that the questions capture, if you don't have time for a full -- (inaudible) --

MS. MILLER: So I fully appreciate the problem. And, at the same time, I think that whenever we're writing questions, we need to have a concerted effort to come up to have to have -- to have a plan of how we're going to go about question evaluation. So, again, it's like -- it's keeping track of the questions that go into the field, having mixed-method or having, like, these follow-up probe questions that we had on RANDS to be able to see, okay, this is going to be -- there's going to be error in this question. It's going to be more rare for less-educated people. Let's keep that in mind as we interpret the data that's coming in. But, then, again, keeping track of what we're asking so that we can improve our questions. I mean, I would like to think that, you know, we're -- what -- how many a year? We're more than a year into this but, we are so far into this pandemic, I'd like to think that we have a track -- we -- our questions that we're asking are much more improved from the questions that we began with, you know, when it first started. So, again, it's just really having a question-evaluation plan going forward.

MR. SCANLON: But that plan -- so, in our instance, with the RANDS survey,, because we were using the RANDS to actually produce public estimates, you know? And this -- we were in the field before we were able to do the cognitive testing that Stephanie talked about. So in that -- I think that's kind of a similar situation. You know, we had to run questions and evaluate them at the same time. And, you know, we did that iteratively, because we had multiple rounds. And that was helpful. So we were able to start in the first round of some kind of open-ended questions and then, as you saw, do more targeted questions later on. But just the idea, as Kristen said, of having a plan to say, okay, we know we want to evaluate this in some way. You know, is it experimental design? Is it using probes? Whatever.

MS. MILLER: And I just -- even to follow up on Paul, I think it's really important for when -- the fact that we had this understandings from our question-evaluation studies with the RANDS, that we were connected then to other agencies that were collecting COVID data and we were able to tell them what we were finding so that they could take our -- to

take our findings to improve their questions. And so I think that's what's -- that's also what's really important, is to be able to network. And even if we can't really write reports, and make everything all official, at least we can -- we can network and improve off of each other.

MS. WILMOT: So still on the lines of the discussion we're currently having, we have a query about how long does this whole process take from, you know, start to finish if -- to do a full -- a full evaluation, as you describe?

MS. MILLER: So it depends on the level of work and the complexity of the question. We can -- if you're doing a small study of 20 respondents, you know, that could take a couple months. If it's a real important question and you really need to make sure that everybody -- like, a demographic question, you really need to make sure that, you know, everyone across the country really understands this the way you need to understand it in the different languages. This can be an extremely involved process and could take a year or two, even, to make sure you get it right. I don't know if you guys have any follow-up on that.

MR. SCANLON: It depends on the -- it depends on what we're doing, too. This goes back to how we're choosing to evaluate. If we do a cognitive-interviewing project, that takes a certain amount of time, right? Because we have to recruit people and then, you know, we do that kind of analysis. If we're doing web probing, we need to have a survey to put them on. And so it's -- everything takes time but you just kind of have to figure out the level of evaluation that you're comfortable with, in that specific instance.

MS. WILMOT: And, kind of along those lines. Paul, the -- thinking about the RANDS survey, I think I've understood this question correctly. You know, how quickly can you get stuff out there? Or do you have a -- is there a mechanism for other people to be able to use the RANDS survey?

MR. SCANLON: Oh, yeah. So there's no mechanism for other people to use the RANDS through us. You know, there are mechanisms to engage with NCHS. In terms of speed, when we got the -- these COVID surveys, I think we pulled them off in about a month and a half, two months. You know, we were sprinting for that. It probably would have been a little faster but there's obviously approval processes that we have to go through. But I think in terms of field time, we are -- we're in the field for usually about a month. And so analysis, then, would be on top of that.

MS. WILMOT: Okay. And then, fixing a bit more on the cognitive testing, we have a question about the vaccine-hesitancy questions that you've tested in the past and whether or not they needed to be retested. But I think they were adapted in the situation. Were they -- were they retested, Paul? Do you know?

MR. SCANLON: We didn't do a new -- so we did a massive cognitive-interviewing project on designing and evaluating those questions. And you can also find that report on Q-Bank. We then -- for the newest round of RANDS -- and I know we have some partners throughout CDC who are also using those questions and they have adapted them for COVID. They haven't been tested in terms of cognitive interviewing but, on the current RANDS during COVID-19 that's in the field right now, we are doing extensive web probing on those to see how our modifications to ask specifically about COVID-19 vaccines are working.

MS. WILMOT: So we also have some comments on the way in which the cognitive testing or the cognitive evaluation is carried out, some comments querying whether those in higher-education groups understood the questions or the question intent better than those in lower socioeconomic groups. And I think there was another question that it looks like someone deleted. But it was whether or not the respondents were influenced by their fear of COVID when they were answering the questions or if Stephanie had any feel for that when she was conducting the interviews.

MS. WILSON: Right. Actually, that's a very astute observation. Because, again, the experience that people had -- you had people who -- right -- thought it was a hoax over here to people who were super afraid of getting it, right? So that absolutely was filtered through. One example I didn't get a chance to talk about was this need -- there were questions about, did you need medical care for something but not get it because of the pandemic? And so that kind of fear, the idea of need was filtered through people's experience with the pandemic and whether -- how afraid they were of catching it. So certain things were missed. Because suddenly now I don't need to get a checkup. I don't need a well-woman visit. I don't need a cancer screening because of the pandemic, right? Where in non-pandemic times, yes, I feel like I need those things. So that is an element of fear that absolutely did factor into interpretations.

MS. WILMOT: Sorry. I'm smiling. One of the -- one of the comments -- are your staff experiencing burnout? What do you -- what are you doing to combat these quick turnarounds? Thank you for acknowledging that. Okay. I have a -- there's a question about the Dynata sample pool, if -- whether you could say a bit more about how you incorporate the nonprobability Dynata sample into your findings.

MR. SCANLON: Right. So, for the findings I presented today, we -- I literally just lumped the probability and nonprobability together. And what I showed you today -- and I apologize for not saying this -- this was unweighted data. NORC is -- has explored and provided us with some combination weights where you could actually link or weight the probability and nonprobability together. We at NCHS haven't done enough studies on our own to understand how that

work and kind of the resulting quality to present that data. So that's how we're approaching it. You know, the Dynata sample -- I didn't really talk about this because I was -- you know, didn't want to linger too long on web -- the properties of web panels but Dynata is an opt-in sample and what that means is that people bring themselves to the survey. So you've probably seen this on -- when you click on a news story and it, you know, tries to -- you know, to get access to that news story you have to answer a few questions or it asks you to take a survey. You know, that's a form of opt-in. And so the problem with combining the AmeriSpeak or other high-quality panel surveys with the opt-in is that, you know, you really don't know who's in a -- in that opt-in group and you don't know what they represent in terms of the population. So you certainly do need to be careful when combining them, which is why I chose to just use unweighted data. So, you know, the results I presented today are just, within that sample, you know, here are the percentage of respondents who use that out-of-scope. So --

MS. WILMOT: Okay. And another good question sort of nowadays, now we're all kind of moving towards web surveys -- what are your thoughts on kind of routinely including web probing in mainstage web survey for ongoing evaluation?

MR. SCANLON: A huge fan. I'm a huge fan of that idea. You know, I think this kind of goes back to Kristen's point earlier -- is that you should always have a -- some sort of a question-evaluation plan. And, you know, as people are using web surveys more and more as kind of these quick-response vehicles, to get data out right away, inserting these probes gives us an opportunity to understand the data quality, if people aren't doing full cognitive-interviewing projects. And I think we should say that doing a full cognitive-interviewing project, like what Stephanie did is a gold standard in terms of understanding what the patterns of interpretation are. I mean, obviously, doing web probing can give you something else. It provides us with more information. But the way we like to approach it as -- at NCHS is that you do them both together. So you kind of have both the qualitative and the quantitative approach and you have a much better resolution of how these questions work. But I think -- including probes and trying to understand how people understand questions in everyday surveys I think is a great idea.

MS. WILMOT: And another question that's come in about the RANDS panel -- sorry, Paul. You said in your presentation it fulfilled two purposes: one, to evaluate how the questions were actually answered; and, two, to improve questions for future collection. I'm not -- I'm not familiar with the term but the -- (inaudible) -- apparently. Okay. So the question is -- you know, does that actually have the ability or by doing a few -- do you actually have the ability to measure any change? Have I understood the question right? How do you deal with measuring change? Yeah.

MR. SCANLON: Yeah, I'm looking at the question now. You know, the question about change over time is an interesting with these panel surveys, because one of -- one of the unknowns at this point -- and we're doing a lot of work. So we presented today on the question evaluations out of RANDS but there is a whole other aspect of RANDS research that's really digging into what these panels look like, what is the data quality that comes out of them, how can we calibrate that data to make it look more like our traditional surveys, like the National Health Interview Survey. You know, and one of the underlying questions is, how much should these panels change over time. So basically, if we -- if we ran -- like, we did a set of questions in -- I guess we did it in June and then August. You know, can we look at that change over time and trust it? And I think our initial findings are that we were comfortable enough presenting these as experimental estimates. I think there's a lot to really understand about how panel compositions change and whether or not that's trackable over time. We also -- there is a public-health need to what we were doing and we wanted to make sure we could get the data out. And so I think that's why we chose to use it that way.

MS. WILMOT: Okay. And, Stephanie, an interesting question about your experiences of conducting these interviews using remote interviewing and how that might compare in your experience with face-to-face interviews. Do we think we'll continue to use virtual interviews during the kind of post-pandemic?

MS. WILSON: Right, so the interesting thing was -- even though I've been doing this for a long time -- I had never done a virtual interview prior to the pandemic. So I went into it kind of skeptical. But I have to tell you I am a convert. I really feel as though the Zoom interviews really gave the same kind of quality cognitive interview data that face-to-face, you know, an in-person face-to-face interview gave. So I think we should continue to use this. I think that in certain situations -- you know, there's downside maybe in terms of socioeconomic status but an upside to this would be geographical diversity that we can now explore that you can't do -- it takes so much more money to, let's say, go regionally throughout the United States, for example. And we did have actually some geographic diversity in our sample here. So not enough that made it count because we were trying to do it quickly. But, yeah, I think it should continue to be used.

MS. WILMOT: Okay. Thank you. And while you were -- because of the subject matter, can you comment on whether any of the participants -- you may not be able to -- became emotionally upset during the cognitive interview? And how would your interviewers have handled that -- if they did -- about the COVID pandemic, I guess?

MS. WILSON: So I -- you know, not really. Of course, there were some heartfelt conversations but there was nothing

recorded where it became troublesome. And I think that our interviewers, they were all highly skilled interviewers that we use. And I think it was just -- if you just show a human component and connect with them and commiserate sort of, that was just -- it was fine. So, no, there was nothing that was a distraction to the interview. That's the essence of the question. Yeah.

MS. WILMOT: Yes. Okay. Thank you. And a very quick question to Kristen -- are we planning on providing any questionnaire-design training or cognitive training?

MS. MILLER: Well, I think that we definitely should be thinking about, doing that. A very specific purpose for us to do this webinar is to make ourselves -- make our program known and make it known the kind of work that we do. And, certainly, if anybody has an inclination they would like to work with us, please reach out and reach out to me and we can set things -- set something up and talk. At the same time, we have given -- we have given question design and question-evaluation workshops in the past. And I imagine that in the near future we probably will as well.

MS. WILMOT: And the last question, again, to Kristen, that we have time for today -- a very informative presentation but the question is, do we test these questions in different languages, different cultures? Can you comment on --

MS. MILLER: Well, that certainly would be the gold standard, right? And we have -- in our program, we have done a lot of question evaluation in international contexts and so have really worked out the methodologies on how to do that. So that's definitely there. A lot of times, we just do things in English and then that's just not ideal, is it? And so, going forward, we are trying to do more and more in Spanish, so --

MS. WILMOT: Okay. Well, thank you very much. I'm afraid that's all we have the time available for this afternoon. Thank you again to our presenters. If your question has not been addressed, I do apologize but you can direct it to PAOquery@cdc.gov. That's P-A-O-Q-U-E-R-Y@cdc.gov. And we will respond to your email. Again, a recording of this session will be available on the NCHS website. I wish everybody a lovely afternoon.