Jon Krosnick: So, Nora Cate Schaeffer generously agreed to speak to us from her office, maybe, at the University of Wisconsin.

Nora Cate Schaeffer: Yes.

Jon Krosnick: And we are very grateful for that. And I can think of no one more exciting to talk about how interviewers misbehave, so we will begin this little segment with Nora Cate.

Nora Cate Schaeffer: Thank you. Is that Jon introducing me?

Jon Krosnick: It was.

Nora Cate Schaeffer: Thank you, Jon. I – you’ll see how grateful you are when I’m finished, I guess.

Thank you very much for the opportunity. The topic that I was given was deviations from the script.

Nora Cate Schaeffer: The first question, of course, that I had, but I did not ask because I thought I’d make up my own answer was, “Exactly which script?” and there are at least two different scripts that could be involved. One is the script of the survey questions themselves; and the other is the script of the rules of standardization. So, I’m going to talk about both of those things in various ways. And what I’m not going to do is to talk about my own work to any great extent and I’m not going to try to do an overview of the literature; but, I’m going to just try to talk about some studies that I think might be thought provoking as a way of thinking about what some of the issues are. And then, I’m going to make some – provide some illustrations of what some of the issues are in various ways; and then, I’ll talk a little bit about some thoughts about what might be useful to do research on, but to think – some of the challenges facing research in this area in the future.

So, I’m not sure. The slide gives my topic of interviewer deviations from the script and I don’t know if I need to think about or worry about this so much with this audience, but I thought I would say a couple of words about what I think about interviewers and then talk about the social and kind of the context for these – think – for in which behavior in the survey interview occurs; then,
say some things about deviations of different kinds and their effect on data using some of the illustrations I just referred to. If there’s time, then I’m going to try to stay around 30 minutes so that there’s room for the next speaker, something about – just present some transcripts showing some cites of tension between conversational practices in standardized interviewing; and then, finish up by saying some things about studying interviewers and interviewee.

So, why I think about interviewers at all? Well, interviewers will probably continue to be important in recruiting the first stage of critical panels and, at least for us, CAPI surveys tend to become increasingly complex so that, in addition to asking questions, interviewers can have to do complex histories and timelines, anthropometrics and tests of different kinds, blood spots, other bodily fluids; and then, persuading respondents to give permission to obtain and link records, such as social security records and telephone interviews also continue to be important for panel studies. So, I think even though self-administered modes have increased in importance over the last decade or so, interviewers will continue to play an important role for the foreseeable future.

So, how – what is the context in which we could put our thinking about what happens in the interview? So, this is a very – I’m sorry.

[Laughter]

It’s a very messy picture because I made it and I’m not very good at this, but this is a model that Jen Dykema and I have been working on in various combinations over the last few years. And it’s an attempt to try to think about what are things that go into influencing the behavior of the interviewer and the behavior of the respondent as we observe it in the survey interview. And one thing that we don’t often pay a lot of attention to because we take it for granted is the role of technology in influencing both the behavior of the interviewer directly as she has to manage the instrument, but also indirectly by limiting and shaping the characteristics of the survey questions. So, we have technology as an important, you know, sort of distant cause of the behavior in the interview. The question characteristics, as I mentioned, are pretty – I think we’re used to thinking about question characteristics as an influence on the behavior of the interviewer. The behavior of the interviewer is also shaped by interviewing practices and the training that they receive in standardization and how to behave in standardized ways. And, less obviously, until we look at recordings, which now is
easier and easier to do, the behavior of the interviewer is also shaped by what we could think of as interactional practices at various times, such as, I mean, a very obvious one is the practice that interviewers have of saying, “Okay” when they hear an answer or when they return to the agenda from some digression. So, if the interviewer asks a question and the respondent gives an answer, the interview says, “Okay” as a way of both closing out that sequence and returning to the agenda. That’s very much an interactional practice that’s been observed in other kinds of contexts.

And then, finally, the behavior of the respondent is probably, in terms of deviations from the script of standardization, possibly one of the most important influences on the behavior of the interviewer. So, all of these different pieces – the technology, the question characteristics, the rules of standardization, conversational practices, all of these things affect the behavior of the interviewer, in some cases directly or indirectly, but also the behavior of the respondent and through that, the behavior of the interviewer.

So, I wanted to, sort of, provide an example of how these things might connect to each other in leading to deviations from the script. And, for this, I drew some work by Mark Fuchs that I think also involved, in various combinations, Mick Couper and Sue Ellen Hanson. This was from a while ago. And some studies of these were a set of studies, they think, that looked at how to implement grids. And grids are a kind of, you know, format that we sometimes have used to display data, you know, display, sort of, the information we wanted the interviewer to collect in paper instruments. And when we moved to CAPI, we kind of turned them into series of very irritating questions partly because there was no way, and it’s still very difficult I think for some software, to get a grid on the screen in a way – in the same nice way that we used to be able to do on paper. This study that Fuchs is summarizing here was looking at, sort of, what was the implication for the interaction of having a grid that was organized by topic and, item by item, walk through the grid by topic, a grid that went through an item by item sequence of questions organized by topic, a grid organized by person and an item by item walk through the data collection organized by person? And this was clearly a fairly small study, but one of the things they found was that, in some cases, respondents volunteered a lot of information all at once. So, that’s an instance of a kind of conversational practice occasioned by the fact that the respondent can see what you’re up to, kind of gets the idea of what you want. And so, they tell you, “Everybody
who lives in this household is white”, for example. And when that happens, if the interviewer is in a situation which they have to still walk through all the items, then they have to figure out how to manage that, right? Because they still have to follow the rules of standardization, but they know the answers now to all the information, or at least what the respondent thinks is the answer, to all the information that the interviewer was going to ask for and so forth.

So, this is a kind of situation in which there can be deviations from the script, both in terms of the interviewer not asking all the questions that they were supposed to ask; but, also, in terms of the interviewer’s possibly not following the rules of standardization because they’re trying to manage the information that the respondent has applied unexpectedly.

Now, when respondents do this, this table is – this graph might be a little hard to interpret because the order of these bars is different from the order on the preceding graph. So, I’ll just tell you what it says, which is that if the instrument is designed in a – with – as a grid organized by topic, the information is provided much more quickly and, presumably, that’s because the respondent does things like I just suggested, saying, “All of us are white who live in this house” or “Everybody here is a citizen of the United States” or whatever the question is and just dispenses with a whole lot at once.

So, with that as just, kind of, an illustration of how technology interviewing practices, respondent behavior, question characteristics and so forth might all be tied up together, and an example that we’ll return to in a little while if I don’t use up all my time in digressing, we’ll not turn to some different kinds of deviations.

So, in thinking about the sorts of deviations that you all might be interested in hearing about, of course, the obvious one is not reading the question as worded. Another is follow up behaviors and, here, there’s not an actual script, but there’s principles of standardization that interviewers could follow or deviate from.

One kind of follow up behavior that is of interest to people is providing definitions and others’ feedback. And then, there are other kinds of follow up behaviors.

Then, there’s interviewing for information that comes in a variable form, such as interview – such as event history calendars or
timelines or the sort of grid that I was just talking about where the respondent just provides variable amounts of information at different locations and then, the interviewer needs to sort out whether to take it all then, ask the questions that appear on the screen and so forth.

And then, there’s other kinds of intrusions of conversational practices that present challenges for the rules of standardization and, if we have time, we’ll look at a few of those.

So, Jen Dykema and I, in a paper not so long ago, tried to summarize some of the things we think we know about the relationship between behaviors of survey interviewers and the results and quality of measurement. And we restricted ourselves to studies that were, sort of, record check studies or had some formal analysis in the case of variable errors, some formal analysis of those variable errors. So, there’s the very wonderful study by Groves and Magilavy from 1986 that – it’s – I don’t think they looked at the question reading the way we think of it in an interaction coding sense. I – it was a – it’s a little hard for me to understand exactly what they mean by the question reading and clarity and pace of question. But, whatever they meant by it, they didn’t find any consistent relationship with interviewer variance when they examined a set of 25 items.

And then, there’s the Hessinger and Bushery paper that found no effect of exact reading of question on the test/retest index of inconsistency in their examination of, I think it was, 34 questions.

Then, if we look at validity, sort of, in the sense of – and, here, I’m thinking about just in the record check study sense, there’s the Dykema, Lubkowski and Bliss study, which has – which was the – how fields study data comparing answers to health clinic or hospital records and they found that a substantive change in the reading of the question had no effect for 9 of 10 items and increased accuracy in one case. This is a summary of their – at the question analysis and their various results. They have fairly complicated tables and there’s several variants of this result. But, this one is more comparable to the one below in which, looking at 11 items, there was no effect with question reading for 9, decreased accuracy for 1 and increased for another.

So, look, this is clearly a fairly limited range of infor – I mean, you know, this is – these are not – there aren’t a lot of studies out there that one can compile. And, but, when we look at them together, one thing we might say as well, what the interviewer – the way the
interviewer reads and delivers the question doesn’t really have much effect. However, it’s very important to keep in mind that we’re looking at standardized interviewers and the level of standardization varies over these studies, but it’s not exactly a free for all out there and most of these are telephone studies where the interviewers are pretty carefully monitored and supervised. So, it might be more accurate to say that when standardized interviewers are carefully monitored and supervised, the kinds of deviations from question wording that they engage in don’t seem to have a big effect on the quality of the data.

The other thing we can say is that there are these striking examples that it’s very easy to become absorbed by thinking about, and which it seems that deviations from standardized – from the exact question wording made some difference, either improving accuracy or decreasing accuracy; but we don’t know because the analyses weren’t able to go to that next step of, “Well, golly, what actually happened? Was this just one or two interviewers who were responsible for this effect or was there some specific behavior that improved or worsened a question? And, if so, what were those interviewers doing to make the question or better or worse?” Those would be really good things to know, but we don’t know them, and they’re kind of expensive questions to answer ‘cause you have to – you can find the cases, but then you have to go rummaging through the recordings and these are not new studies. So, rummaging through the recordings is slow.

There are other kinds – the other kinds of deviations from the script we might think about, other than question reading, include, as I already suggested, follow up behaviors. And one of the things we know about follow up behaviors is that probing – when probing happens, data quality is reduced. Now, and that’s – I don’t mean to say that in a causal way because, as I’ll say in a few moments, it’s probably not the case. But, probing is associated with increased interviewer effects and follow up behaviors by the interviewer are associated by increased accuracy – I’m sorry – decreased accuracy. And that’s whether the interviewer is following the rules of standardization or not when she does the follow up and that’s also the case in event history calendar interviewing according to Belli. Oh, I’m sorry, this was not an event history calendar. This was one of the health field study interviews. So, when probing and when there’s any probing or feedback, when they have significant effects, they seem to be associated with reduced accuracy, as well as increased interviewer variability.
Now, when these follow up behaviors happen and the – what’s usually going on is that there’s a problem of some kind, either the question has a problem or, and what often seems to be the case, is that there’s a lack of fit between the question and the respondent’s situation or the respondent can’t remember or there’s some behavior of the respondent that is – in which the respondent is expressing some difficulty with answering the question so that it’s – I don’t think most of us think of it as the interviewer causing these problems, but, rather, the be – the – it’s not that the interviewer is doing a bad job of probing. It’s just that when probing is needed, we’re in a difficult situation.

The contribution of the design of the question to this situation is illustrated in this very nice picture from a paper by Shell and Kreuter and, probably, you could do some – could have done something similar with an O’Muircheartaigh and Campanelli ’98 paper, if they had this nice plot. But, what this plot shows is they, in this paper, Shell and Kreuter made an index of things about items that might be associated with increased interviewer variability. And what they find is that the fraction of the total cluster variance that is due to the interviewer gets bigger as – and more consistently big when the survey question has some of these harmful properties, such as being an open question or things like that.

Okay. There are other kinds of follow up behaviors that we might be interested in and one of them is providing definitions. There’s a very famous experiment by Conrad and Schober that compared both standardized and flexible interviewers. Both sets of interviewers were trained to read the question as worded, but the respondents in the flexible condition were given this extensive additional instructions about how important it was to ask questions and the flexible interviewers were trained to follow up, as needed, to help the respondent understand. The techniques that were used appeared to be effective at improving the understanding of the respondent when the respondent scenario wasn’t a good match to the question, which is one of the situations I referred to a while ago. But, the interviews took more time.

In reflecting on providing definitions and I have some examples that I can see already we won’t have time to look at, but, in which interviewers might be providing definitions. But, it seems as though providing definitions can improve respondents’ understanding of complex concepts when the respondent situation needs it. But, we don’t have studies that compare that method of providing definitions with other methods of providing definitions,
many of which are discussed by Schober and Conrad. And some of the issues, very important issues, are discussed in their paper. So, it would be very helpful, I think, in thinking about production interviewing to think about expanding the comparison, not just to providing definitions this way versus no definitions, but thinking about other ways of providing definitions and, with any luck, doing – or with enough money – doing it in a way that both variable errors and bias could be assessed simultaneously, which is one of the things that doesn’t happen very often.

Feedback is another kind of possible deviation from their script and there aren’t a lot of studies that do this. One study, Groves and Magilavy, looked at one of, I think it was Charlie Cannell’s studies in the set of studies that they examined, which had an experiment that compared completely scripted feedback to very limited range of feedback. But, even within this very narrow range, they saw a tendency, although not significant, for the rho sub n to be smaller for the group with the scripted feedback.

One of the other few studies that’s out there that I know of, experiments that look at feedback is Dexter’s 1987 study, in which they tried to use personal feedback to motivate respondents and I, personally, I think this issue of respondent motivation is very important and I think, likely, to be increasingly important if we’re going to be, you know, sort of – when we get to peoples’ houses and it costs us so much to get there, we want to make them work really hard and do all kinds of things for hours on end, some of ‘em rather distasteful. I think paying attention to respondent motivation is a very important thing for us to do and one of the things that Dykema did was experiment with this and found, even with a small number of interviewers and using the right degrees of freedom as nearly as I can tell to do the test for the interviewer level variable as opposed to the respondent level variable, found, you know, a couple of significant effects. So, something that’s worth thinking about some more.

One kind of interview that, sort of, deviates completely from any script, not really, but just sort of in a, you know, just in a rhetorical kind of way, is event history calendar interviewing because it’s very flexible and there’s no, in some versions, there’s at least an elementary script. In some versions, there isn’t much of a script. There’s just interviewer training. But, it certainly illustrates the very close relationship among technology, instrument and interviewing practices, but it also illustrates a lot of the challenges that come up in trying to design studies to evaluate different kinds of interviewing. So, if you want to compare event history calendar
interviews with standardized interviews, just finding the right place to make the comparison is very hard. So, coding the – you have different behaviors that occur in the two different styles of interviews, not surprisingly, because you trained the interviewers to behave differently. One set of interviewers has a script. The other set of interviewers doesn’t. So, it’s very – how to make assessments of one group versus the other group isn’t straightforward and when you observe differences, it’s very hard to know what you should attribute the difference to. And it could be that some of it is to behaviors that we’re not even tracking, such as motivating feedback or something like that.

The other thing that we see in thinking about this kind of study is how important it is to have the right kind of statistical design and analysis to think about the – how to assess the results because the design of such studies is so complicated.

So, I wanted to return quickly to the idea of a grid and some of the things that comes up when we try to think about another kind of, like – so, you know, the flexible interviewing in which the respondent was given definitions is designed to deviate a little bit from standardization to try to improve comprehension. Event history calendars deviate from standards – traditional standardized interviewing to try to improve recall. But, the example that we started out with, the grid, sort of – respondents spontaneously do things that deviate from what we – the way – the standardized interview assumes they’re going to behave. And we might think about how we would design a method of data collection to get complex information of the kind obtained by grids.

So, one of the studies that we worked on recent – that we developed recently at the UW Survey Center for one of our clients, and they were interested in complex family structure, and we were looking at ways – and this was a very large team of people. I took out the slide acknowledging all of them for the sake of time, but this is not my product. This is a very collaborative product. But, the collaboration referred to in the title is between the interviewer and the respondent and it’s a label taken from the old Suchman and Jordan paper when they talk about having an instrument where the respondent and interviewer could both see the form that was being filled out. And since our client wanted such repetitive information, we wanted some kind of aid that would convey the structure of the task to the respondent, made collecting the data more efficient, kind of work on the motivation issue, as well as recall, and then display the information to help the respondent really see the structure of the task and also give the respondent the opportunity to
correct any information that was recorded incorrectly. We also wanted the interviewer to be able to record answers the way the respondent provided the answers. And so, sort of, to follow the kind of conversational practices that we thought we saw in looking at what Fuchs, Couper and Hanson had to say about their grid experience.

And when we think about visual aids and such, as they’re used in interviews, it’s something that we really don’t know very much about. I’m not going to spend much time here so that I can show the next picture. Let me go – let me just say one thing from this slide. If we’re thinking about using technology to give us better visual aids, one of the things we have to think about is what are the interviewing practices that go with the technology? And it seems – I think it seemed to us, as we were working on this, that every technology, whether it’s paper or CAPI or some other kind of technology, needs a set of interviewing practices that somehow helps the interviewer manage conversational practices, understanding, reducing interviewer variability, the motivation of the respondent and, you know, still keeps reliability and validity in mind. So, we need – when we introduce something new, we need to think about, “What are the rules for interviewing that goes with it?”

So, this is the visual aid that we were working with for the Swiss mom study and what we devised was a kind of dynamic household roster. So, this – what you’re seeing here is just a screen, but this is something that’s filled out dynamically. The respondent holds the iPad and then, as they give the names of the people in the household, the names appear. As they identify the relationship and the birth date, all that information appears to the respondent so that the respondent has the big picture of what we’re trying to do and can also see what information the interviewer has entered.

We had something similar for the, kind of, it’s not really an event history calendar, it’s more a timeline about co-residents, when the children were born and when various fathers and boyfriends co-resided with the mother and the children. So, this, again, in the interview, is a dynamic display of all the children we’ve learned about, all the fathers we’ve learned about, and then the time period from the time the oldest child was born.

We also had a lot of repetitive questions about who contributed to the household, in which we were able to list all the people that we were going to ask about, all the different kinds of contributions
they were going to ask about and, again, display that to the respondent.

The – but, doing all this required coming up with rules for how to train the interviewers and this was one of the things that I was working on and one of the things that I think wasn’t done as well as it should have been because we didn’t know what –

[Laughter]

– what the interviewers and respondents would actually – how they would actually end up using this. So, you know, we’ve got this pile of 200 recordings and, at some point, we’ll be able to listen to them and find out. But, for example, one of the things we needed to develop was rules for how to do a verification. And we wanted rules that were more detailed than any rules we were able to find guidance on. And, as we worked on it, we realized we needed to be able to train interviewers to notice inconsistencies and distinctions of pretty complicated kinds, you know? Sort of, you know, so Andrew and Sarah are brother and sister, but they could have different fathers and we want the interviewers to be aware of that and be able to give back to the respondent the appropriate questions in order to get confirmation of what the correct information is. We also –

Nora Cate Schaeffer:

So, here’s an issue – a presentation that shows – a slide that shows the verification – the information we were trying to train interviewers to use in verification for repeated information. And you can see that we have all the people in the household, all the children, and then, the different meals that they could get at their daycare. But, we wanted the interviewer to be able to take information volunteered by the respondent in a fairly flexible way. Not as well done as we might have done, but, you know, a good first draft.

So, some of the challenges is how do respondents provide information and how do they use the display and how can interviewers manage the variable provision of information by the respondents and what are the set of interviewing practices that support the use of the display? I’m going to skip over the material, as I projected I would, about the conversational practices and the cites of tension and I’ll just – you’ll have to watch it flip by. And we’ll go to some slides about the future and I hope I’ll be – I hope I’ll be on time.
So, some of the things that we’ll probably need to look at in future research on interviewing is interviewing on complex topics, continuing work on things like the event history calendar, household structure and things like that; complex tasks, including physical measurements, cognitive assessments and so forth; how to design questions that are sensitive to conversational practices and stimulating recall; and now, I just want to say, this isn’t all the subset topics concern – this is just a subset of topics concerned with interviewing that I could have mentioned. This is just those concerned with deviating from the script.

What do we need in study designs? We need manipulation checks when we have experiments; identification of key behaviors that can be coded reliably; and we need outcomes built into the design that can be compared across conditions so that we can come to some conclusions. We need experiments that are large enough to have a sufficient number of interviewers and appropriate assignment of respondents to interviewers when we’re measuring and assessing an interviewer level behavior. And then, we need the right analytic models, as well as designs that let us look at, ideally, validity and reliability together, but, you know, at least something, some kind of criterion. And these studies need to use some kind of realistic research context that can serve as a model for large scale production. And then, include some kind of development and assessment of interviewer training and monitoring. Monitoring is very important. And I think those are – continues to be a role for small scale lab experiments, probably a role for single method studies of a single method of interviewing, although, I have, kind of, complicated thoughts and feelings about that; and then, of course, large scale field experiments, but they tend to be quite expensive.

And that’s it. I took the exclamation point off my thank you so that I – so that Colm would feel more comfortable.

*Jon Krosnick:* Is it you? He had your exclamation point and you have his not exclamation point?

*Nora Cate Schaeffer:* Right.

*[Laughter]*