Jon Krosnick: Okay, I like Jan’s concept of the “send ahead.” Is that what that was called? So, now we have something even ahead of the send ahead, which is supplementing server data collection by preparing appropriately through cognitive interviewing.

You know that pretesting, in retrospect – I can say this, ‘cause I’m young enough – pretesting was embarrassingly oversimplified for a really long time in survey research, and we came to realize, partly thanks to psychology, that there is more to it that we can learn and improve in a questionnaire through more sophisticated data collection methods.

And Gordon’s been in this since the very beginning, so I’m particularly delighted that he wrote his book, and that he’s willing to come give us his insights into all this. Gordon Willis.

Gordon Willis: All right, I want to thank Jon for inviting me. I see he brought a newspaper to read during my talk, which gives me a lot of confidence here.

But what I’d like to talk about is mainly cognitive interviewing, which we hope is a science. So, we’re talking about state of the science as opposed to state of the art.

One of the issues we really have to deal with here is how much is idiosyncratic, is art versus a strict science? And I think there’s a lot of research that could be done here.

To start, to put things in context, I see what we do within a total survey error paradigm. So, everybody has their own favorite source of error, which to some extent different presenters have talked about. So, we have coverage error, sampling error, non-response error, measurement error due to the interviewer, measurement error due to the respondents, post-survey error with analysis and all that.

Well, this is where I live. You know, it’s sometimes called response error. We’re talking about error that’s controllable, presumably through either question or questionnaire design. And the notion here is that this is worthy of attention and research, the fact that small changes in questionnaire design/wording/format can make a big difference.

I hope I’m preaching to the converted on that. But I have one example which I really like, not only because it’s mine, but I’ve had three different reasons for showing this example. One, I think
it serves as a way of showing what cognitive interviewing is by way of example. I also think it shows the importance of making small changes in questionnaires, and they’re not always trivial.

I know Nora Cate Schaeffer is listening in from Wisconsin, and she probably would clap or agree with me, even though we can’t hear her doing so.

And finally, I think this example shows the kind of research that I think should actually be done more, ‘cause this is from several years ago. The context here is way, way back, when I worked at the National Center for Health Statistics Cognitive Lab, and we had a production kind of system going.

We’d test survey questions like this one, presented to us to measure strenuous physical activity. So, we’re presented this, “It probably works, but see if there’s any problem with it.” So, on a typical day, how much time do you spend doing strenuous physical activity, such as lifting, pushing, and pulling?

Go into our Cognitive Lab, do a few interviews, and it becomes very clear very quickly that there could be a problem here, because we’ll have people say something like, “Three hours,” “Four hours.”

And then you probe them and say, “Okay, tell me about your typical day,” and you find out they do nothing more strenuous each day than, like, reloading the photocopy machine. Right? So, what’s going on? Well, really it looks like it’s a demand effect classically. Who wants to say, “None,” right? You look like a couch potato.

So, the alternative we tried instead was to use a filtered approach. You just add a yes/no question, basic resolution to a common problem. Say, “On a typical day, do you spend any time doing this stuff?” If it’s, “No,” well, we’re done. And it looked like respondents were happy to say, “No, I work in an office.” It seemed like a good idea.

If they say, “Yes,” then, “Oh.” Then you branch to the second question and add the follow-up. Well, at that point, there was a senior level number of the establishment who said, “Yeah, well, you guys are adding another question. We don’t want to do that really. And, you know, you’re probably right; it probably makes a little difference. But, you know, you’ve heard this before, surveys are a blunt instrument. Does it really make a big difference?”
I thought it did, and I got a chance to test that out empirically. And I think this is what we should do more of is to predict things, put our money where our mouth is for pretesting, and rather than saying, “We know it’s true because it came out of the Cognitive Lab,” let’s at least see if our effects occur in actual ecologically valid kind of environment.

So, you know, you think I was right that it makes a big difference? Well, of course. Otherwise, I wouldn’t be showing it to you – would I? So, yeah, it can matter a lot.

So, here’s what happens when we use both versions. And what I mean by this, here’s no filter was the original, which we said was going to produce relatively few people that report no physical activity. And the filtered version with the additional questions should raise the proportion of people who report no activity.

And this is significant. You’re talking about 72 percent versus 32 percent in a field pretest. Did it again, just in case I didn’t believe my first results, and get even stronger results with a second study. And the effects are enormous, you know, from almost nothing to half the sample.

So, you know, on that basis, we think that those that are worried about, like, three percent sampling errors – give me a break. You know, we got huge orders of magnitude of response error. So, this is what we’re dealing with, with cognitive testing.

As a way of background – I’ll go through it quickly for anybody who doesn’t know a lot about cognitive testing – we’re developing, evaluating testing, either a questionnaire or other survey-related materials: consent forms, introductory letters that we mail to people. They’re all amenable to cognitive testing. It’s a flexible method. Right?

We recruit people differently than we do for regular surveys, so it’s different. We pay them, ‘cause we want their undivided attention, and we conduct one-on-one interviews, typically in a cognitive lab, but that doesn’t mean we have something necessarily very sophisticated. But again, we go to where we can find people.

The basic approach here involves verbal probing approaches. We either probe, like I was just talking about, and which I’ll give some examples of, or use classic type of think-aloud, like hopefully sit there and listen and let them just think aloud while they go through
the questions and listen. That doesn’t always work, so we do a lot of probing.

The most common – not only the most common model for cognitive testing is to find problems and then fix them, to make repairs. That seems obvious, but these are different skills that are involved in those two parts, and, you know, each one, perhaps, could be amenable to different types of research.

Just as a general note, cognitive testing in any form seems to be best done as an iterative type process, where we do small rounds and make changes to questionnaires in between rounds. It doesn’t seem to be necessary to do a lot of testing before we change, although that does bring up some interesting issues related to sample size, which I will get to.

Okay, the theory for NSF staff, I’ve seen proposals always should contain some element of theory. There is a theory – not a whole lot – behind cognitive interviewing. What has held up really well, I think, over the years is this basic yet phenomenal, in a sense, model by Roger Tourangeau, now going on 30 years. Can you believe that? We’re getting old, dude, since that came out.

*Roger Tourangeau:* Speak for yourself.

*Gordon Willis:* Yeah, well, Roger Tourangeau’s grandfather came out with this model way back, and I think he still believes in it. But the notion is people have to understand the question – like what a podiatrist is – be able to retrieve the information; do they have a chance of even knowing the answer; decide to tell the truth, even given they fully understand the question, they damn well remember the answer.

But say it’s a national health interview survey, and we’re asking, “How many sex partners have you had in the past 12 months?” You have to envision this happening with a spouse listening intently on the couch. You know, there could be some bias in the response. It’s a decision process issue.

And then finally, we have to ask for responses in the way that they’re going to give them to us. That’s the basic model. I think we’ve widened this view over the last 10-20 years to incorporate more sociodemographic, sociocultural type of issues. So, it’s not only just plain cognitive psychology, but it’s a good model.

We operationalize cognitive testing through doing these probes like, “What does the term dental sealant mean to you,” something
that’s readily misunderstood. Paraphrasing. “Can you repeat the question in your own words,” and see what you get back. Asking how confident people are in their responses. Yeah, how do they recall something that had a very long reference period?

We devise specific probes, and my favorite, perhaps, when I can’t think of anything else to probe is, “Can you tell me more about that?” Very useful for getting more information.

So, with that as a very quick background, there are a number of unresolved scientific and methodological issues that we have to grapple with. I picked out the ones that I think are the most important unresolved issues.

The first is basic, “Is cognitive testing reliable and valid?” And that breaks down into just, “Overall, is this an effective thing to be spending our money and our time and our effort on, and hiring staff?”

And the second thing is what I call “parametric evaluation.” I don’t mean parametric in the statistical sense. I think I just meant procedural evaluation. I should have changed it, but it’s easier to apologize than to fix slides sometimes. So, we’ll go with that.

And then secondly, looking ahead, the question is, “Is cognitive testing useful for – quote – the survey of the future, whatever that’s going to encompass?” And I think that involves both this issue of moving towards mixed modes and also the kind of novel administration procedures that Michael Link talks about.

Secondly, cross-cultural applications are getting to be really important. So, this is where I am with this, saying, “It’s a fact that cognitive testing is fairly widespread. It gets a lot more respect than it used to; it’s accepted now.” Look at RFPs. I’m very gratified to see they always have a call in them, like, “We’ll include cognitive testing of your survey materials of course.”

And so, there are several government laboratories that are devoted to this, and at private contractors as well. So, there are labs at NCHS, at the Bureau of Labor Statistics, Census Bureau. And it’s a fairly widespread, big deal. But we don’t really know, for one, whether these independent labs, testing the same questionnaire, would really come to the same conclusions. Okay?

How idiosyncratic is this all? There have been a few studies on this, with kind of mixed results. One that I was involved with
compared three laboratories and found, yeah, the same questions were problematic, but for different reasons across the different laboratories. Is that a good thing? Bad thing? It wasn’t really clear.

A more recent study by DeMaio and Landreth at the Census Bureau – they did also a three-team comparison. They were unable to tell which is best in the horse race sense, or whether there were certain procedures that were better than others, except, now again, from this procedural point of view, they did at least decide that it’s really important to listen to recordings of the interviews in order to have reliable results. So, we get some information, in any case, from doing these studies.

And then, something that I will go into a little more detail in because I was involved in it, I did a four-lab comparison. I want to show an example of what this is like, and what we think we found. ‘Cause again, I think this is useful research to do, just to reassure ourselves that we’re doing something that’s meaningful.

So, for this example, a few years, had a self-administered questionnaire at the Cancer Institute on perceptions of either breast or prostate cancer risk, depending whether you’re male or female, where the functioning of the questions was unknown in advance. I mean newly scripted questions. So, I thought it was a good opportunity to have four different labs do parallel testing. Use whatever your techniques are, test this thing, tell me what you find, and then I can compare the answers.

A fairly good size cognitive interviewing study, almost 150 interviews in 4 different cultures or languages. So then, again, are the written results going to be very similar, or very different? So, there’s the way we wound up with 148 interviews between National Cancer Institute, Westat, NCHS, Public Health Institute, in those 4 languages, and went at it that way.

So, I don’t know if people can read this in the back. This is the questionnaire. I’ll read it to you. It says, “Please circle the single number, on a scale from 1 to 5, that best describes how concerned you feel right now about the following things.”

And then we have – we actually have this, yeah. “Feelings of concern now,” to remind them, and then things like, “breast cancer occurring in me, my family’s history of cancer,” eh, down to, like, 24 things. And then you do, “Not at all,” to, “Extremely.” Common kind of self-administered form. Right?
So, this is the notion how this is supposed to work cognitively. The respondent looks at the first question, thinks about, okay, feeling of concern. Connects those cognitively, and then picks a response. And, okay, we test it, and we find out if that’s what, in fact, happens.

And we try to anticipate what would go wrong, and people came up with notions like, well, for example, “Breast cancer occurring in me.” Well, that’s kind of clunky; it’s not the best wording. That’s what I thought would come up.

This is what came up instead. This is a model of what respondents actually do. Don’t read the instructions at all. We know that, that people don’t read instructions. They also don’t read this thing about feelings of concern now at all. So, they’re not thinking about concern. They see, “Breast cancer occurring in me,” they see, “Not at all,” to, “Extremely,” and they just go for it and circle something, whether it’s – well, how likely is this, or how much has this occurred, or something else. But the something else is other than how concerned I am about it.

So, for example, when you get down here, “What having breast cancer would do to my body,” you get people that would say, “Well, it would do very much to my body.” That’s not what we’re trying to ask. And the one you can’t see down here is, “My chances of dying of breast cancer,” and the person says, “Somewhat.” It has nothing to do with how concerned they are about that.

So, interesting, but I think what’s more interesting here was this finding was true ubiquitously. That’s a word. Across all the different labs, the different languages, the notes I got back basically said, “This approach does not measure perceptions of degree of concern because concern was ignore.” Okay.

And again, the important point is nobody saw this coming. I mean maybe we should of, would of, could of, but that’s why we do testing when we don’t see these things. And it was nice to see – reassuring, really, to see this come up repeatedly across the four labs. So, end of story – right?

Well, yeah, that constitutes evidence of reliability, but there’s also been opposite results. Kristen Miller, NCHS, did multi-country studies and find out that real problems, you get big differences in
results and don’t know why. And it’s got a lot to do, I think, with different training between staffs of different labs. A lot of stuff.

But the point is, this issue still needs to be addressed. Under what conditions are the cognitive interviewing results reliable? What do we have to do to make that happen?

Okay, so second unresolved issue is this thing about procedural evaluation. This means down in the trenches or nuts and bolts here. Stuff that only a cognitive interviewer could love like, “Which of our particular procedures are useful?” And there used to be a fair amount – to the extent it’s been done in this field – a fair amount of research in this issue. So, we have think-aloud, which is what it sounds like, versus retrospective probing, meaning you administer the questionnaire; you leave all the probing till the end. It’s debriefing.

My census colleagues, going back a ways – this is mainly older research. I’m saying that for a reason. It compared and contrasted these two approaches on a paper questionnaire, and they found, lo and behold, that the findings were very similar, except – almost parenthetically, but I think’s really important – was that subjects with low educational level tended to blow the skip patterns on the questionnaire under think-aloud.

Well, yeah, I think that makes sense. If you’re trying to think aloud and also follow the questionnaire, you’re not very good at some of this stuff, you tend to mess up. So, that may not be really illustrative of what happens if you’re not forced to think aloud. So, it does matter sometimes which procedure you use here. But, they wound up recommending, “Eh, do think-aloud for some, and retrospective probing for others, and you’ll find something useful.”

I just – a couple of other studies, again older, where the author is again Census Bureau. Most of this is government research, actually, I noticed, in the examples I have. That’s where the research has been done on concurrent and retrospective probing versus pure think-aloud, and they said, “Eh, in general the problems were similar in both of these for self-administered questionnaires. A lot of the problems are navigational.

Susan Schechter at NCHS, similar thing and again said, “It’s good to mix techniques.” But that research kind of stopped for a while, and we just accepted what we do seems to be okay, but we haven’t taken it to the next level, really, which I think is bad.
What has been of more recent significant interests is a contentious issue that’s really cooked up is that of appropriate sample size. Cognitive interviewing is a qualitative technique, so we don’t do hundreds of interviews usually. And I put “can’t do power analysis.” You probably could; I don’t know how appropriate it is to do so.

But anyway, the sample sizes are typically small. We’re doing no more than 30 interviews often, and sometimes people do less than 10 even. I mean what do we get from that? Is that useful? Do we reach what they call, in the qualitative field, “saturation?” Which means doing more interviews doesn’t give you much more information.

And how many is enough? Well, think about it further; that’s a complex question because it could break down to meaning, “How many is enough to identify a problem?” That is, how many interviews do you have to do before that problem first occurs? Or do we mean something like validating in the sense that we set a criterion and say, “Well, I want to see a problem happen in three of ten, or eight of ten interviews, or whatever, before I conclude that it’s a problem.”

This has not been well studied. This is something that we have to grapple with. There’s been a little bit of work on it here by somebody sitting in this room. Johnny Blair and Fred Conrad did a nice study, I think, assessing what happens when you increase the number of interviews you do. What happens to the number of problems you find? I mean do you get a nice asymptote? Do you reach saturation?

And I’ll just – Johnny can explain it more, I guess, but here’s a critical graph from their paper, in POQ, which does show a negatively accelerated kind of relationship, except down in this lower level that we’re usually working at. There’s still a lot of stuff going on as you exceed that level, meaning finding more problems. So, they conclude that additional interviews continue to produce observations of new problems, although the rate decreased. But the argument could be made that when you’re down here, you really should be up here and doing more interviews.

Now, there is also pushback against this. I know other people who say that Johnny and Fred picked the wrong dependent variable. I mean – anybody that’s motivated to not do more interviews can
find all kinds of counter arguments. Right? But this is something that we need to look at a little more closely.

So, turning to the survey of the future, we get into mixed modes and novel administration methods issue as well as is cognitive testing useful for these purposes? I think it is. I think it’s a natural; we’re well positioned. Because cognitive interviewing has a strong history of attention to administration mode, in particular because interviewer base versus self-administration are very different cognitively as to whether they deal with auditory-based information versus reading/visually-based information.

So, we’re already there. I mean what we’re dealing with that. We’ve increasingly focused on Web usability. Jennifer Romano-Bergstrom here knows about this. She’s done a lot of usability cognitive testing. They’re starting to converge – right? – in a sense. I mean you have to agree with me – right? Yeah, because – and that seems like a good thing because it enables us to use our current methods as we go to new technologies that increasingly involve computerization.

I also think there’s interesting new research that’s being done. For example, Jennifer Edgar, Bureau of Labor Statistics, on more electronic types of cognitive interviews, either Skype or Internet-based, where you just have your respondents – subjects, whatever you want to call them – fill out questionnaires and answer probe questions kind of by writing and sending you back the stuff.

You think, “Wow, that’s radical. I’m not going to do verbal probing with them? I mean that’ll make my head explode.” But you can do a lot of interviews for the same amount of money as one interviewer-administered cognitive interview. So, if you lose some information, maybe it’s worth it. I’m open-minded enough to consider the possibility. I’m not that old – right?

So, the other issue here that’s really important, I think, is cross-cultural applications. What we all know, this is a hugely important issue in survey methods – right? Because we no longer take the Hispanics, which is a very small group, and then fold them into the overall estimate and say, “That doesn’t really matter whether they’re in or not.” Now, they’re a bigger part of the population, as are Asians – Russians even, if you’re doing a California health interview survey. It matters. Especially when we’re doing comparative estimates. If we want to compare estimates between groups, then we have to avoid the apples to oranges problem. I mean are we really getting the same information?
And then when we go five groups, then, yeah, it becomes not just apples and oranges, but apples to oranges to pears to guanábana to rambutan. Guanábana I’ve had; that’s the Hispanic fruit. And I just found rambutan is a very, very common Asian fruit. And you’ve got to know this stuff – right? – to do the cross-cultural work, or else you’re just not asking the right questions.

So, the issue is, can we use our evaluation pretesting techniques that we already have fairly well established to obtain cross-cultural comparability, or do we have to go back to the drawing board?

I would argue that there are strong issues of cultural misapplication. We’re just asking the wrong questions. You can’t ask poor Hispanic women about physical activity, using a questionnaire – and this has been done – a questionnaire that was validated on Harvard grads. Okay? And it kind of has this vibe like, “Well, the last time you were playing polo, and you got injured, who did you go see?” I mean it just – it doesn’t work.

And then there’s the linguistic things. We have mistranslations, which seem to be a real problem. Even basic translation, I’ve found, can go very badly if we ignore basic pretesting practice. And here I’ve thrown in something which – I always try to find an excuse to include this in any talk, because it’s such a great example.

Anyone seen this before? This may have gone viral, or whatever. In particular, anybody speak Welsh? No? Good. Okay, the British want to be good about this, translate things into Welsh in Wales. So, this sign, “No entry for heavy goods vehicles. Residential site only,” they went through their usual procedure, sending it to their bilingual Welsh-speaking colleague. You know, via Internet, and received something back from him and put it on the sign and went with it.

And then, before too long at all, Welsh people started complaining. ‘Cause what this actually says is, “I am not in the office. Send any work to be translated.” And, I mean, it looks about the right length. I mean nid could be no. I mean so that’ll pass first muster. And there’s a Welsh politician who said, “You know, it’s really nice of them to try, but they really ought to get somebody who speaks Welsh before they go to press.” And this is kind of an encapsulation of what happens more subtly often where I work. But so we have to have methodological research into these cross-cultural applications in particular.
Getting back to cognitive interviewing, the question that we always wonder, doing these cross-cultural, multilingual, cognitive interviewing investigations, do cognitive interviews themselves function similarly across groups?

This is something, again, my colleagues at the Census Bureau are really interested in, because if people respond differently to cognitive probes, I mean you can come to the wrong conclusions. Like something looks like it’s a real problem in Spanish, and if it’s just an artifact of the probing measurement-based process, then we’re really not doing our job right.

So, the issue really is what kinds of modifications do we need to, in particular, analysis procedures. This is another place that’s ripe – as ripe as a guanábana or whatever for further work. Here again, Kristen Miller is kind of leading the way here.

For behavior coding, this – I can force this in, I think, as a cognitive type procedure because, you know, we are dealing with cognitive issues, especially across culturally, because again, the issue is for behavior coding. And what that is, is, unlike cognitive interviewing, we don’t probe to get under the surface, we simply eavesdrop.

We listen to a number of interviews, and we listen to what’s going on for overt indicators there’s problems in the interaction. Do our respondents interrupt? Do they ask for clarification? Do they provide uncodable responses? We’re looking for some indication from them that there’s something amiss. And again, the issue is, you know, do cultural differences here preclude real comparability between groups?

And in particular, do different groups vary in their willingness to basically say, “That’s a stupid question,” or, “I don’t know what you’re talking about,” or that type of thing. Because there’s always that stereotype, you know, like the Asian, the Japanese is just going to go and say, “Yes, yes.” Then you have Russians, who hate everything you do, and they’ll let you know. You know? But is that true? We can’t just say it. I mean we have to establish it.

I think the critical study here, at least to date, the only study I know that’s really looked at this, by Tim Johnson and colleagues, and I think this is brilliant, too – actually, partly NSF funded. I don’t know who at NIH sponsored it; it wasn’t me, but the idea is, “Let’s
embed some really bad questions and do some behavior coding and see what happens, see whether respondents say anything to us, or they just let it go.”

So, here’s just part of the study. But here’s the questions they made up. “In the past ten years, how frequently have you visited a cerarium?” The second one is, “Do you support or oppose a law to ban the import of photams into the U.S.?” And it turned out that overall there was evidence from behavior coding of comprehension problems in – what is that? – 61 percent of the interviews on the cerarium question, and a full 83 percent on photams – I don’t know how you – photam – I don’t know what the appropriate pronunciation is, Tim.

But also what I was wondering here, is when someone did say, “What the blank is that,” did you revert to the old Census standard, “Whatever it means to you?” Yeah, okay, very helpful. But the issue here – all right, so we’ve established that, yeah, you can get respondents to react to that. The question is, how does that work cross culturally? I mean do different groups deal with this the same way?

So, now if you look at the African-Americans, Koreans, Mexican-Americans, whites – this is reassuring. Hey, they’re all real high on the cerarium question, even higher here, saying, “What the blank does that mean,” or, “What are you talking about.” Some kind of response like that indicates a comprehension issue for the photam question.

And to me, it’s not that significant that the numbers varied some because our usual criterion in behavior coding is that something more than 15 or 20 percent of the time produces a problem, and that’s our red flag. So, the point is that we’re going to hit that in all these different groups. It’s not going to be the case that it just kind of runs by the Mexican-Americans, who just sit there and say, “No, I don’t like photams,” or something.

So, I think there’s something to be said for cross-cultural comparability of behavior coding as a method, based on this type of research. And it’s good to do. So, to reiterate, proselytize, to be very overt about what I think should be studied, funded, perhaps by NSF, would be, again, cognitive testing reliable overall. We’ve got to figure that out, ‘cause that’s vital to continue to support this endeavor at all, I think.
This procedural parametric evaluation is real important if we ever have a hope of developing – quote – best practices that are going to be common across the government, across the world, across the universe, whatever.

As far as usefulness, survey of the future, we’ve got to get into, of course, the new administration method, so we’re staying current with developments in the field. And again, cross-cultural applications is really where it’s at as far as maintaining, again, our ecological validity.

I’ll add, too, that a lot of this stuff is not that expensive to do. Some of these studies are, you know, relatively on the cheap side, compared – especially compared to 99 percent of what we do at the cancer institute. But this methodological research, related to cognitive pretesting, or pretesting results, pretesting procedures in general, is fairly low cost, and I think you get a decent bang for the buck out of it.

So – okay, I did okay, 32 minutes. Leaves times for questions or a discussion. And I saw – I was pointing to Johnny?

Johnny Blair: Oh, okay. I’ve got a couple of comments. Okay, a couple of comments. One that occurred to me, in looking at the presentations earlier on validation, and it brought to mind that when we’re doing cognitive interviewing, or other pretesting for that matter, that sometimes it would be nice to know, did the four questions, where there is, in fact, a correct answer, did the respondent come close to or get the correct answer?

And while doing things with validation and sort of production pretesting or cognitive interviewing, there are all sorts of obstacles to that. But if you’re going to do research on the pretesting method on cognitive interviewing, I think that there are some opportunities to add into the mix some sort of validation, even if it’s only for a subset of questions by the frame you select from and – well, I won’t go over the different ways that one can come up with validation information, but it seems to me that that’s often a piece that’s missed.

And cognitive interviewing, as anyone who’s done it knows, we find all sorts of things. I mean the more you talk to respondents, the more respondents you talk to, the more you find. And you usually end up kind of judgmentally deciding, “Well, I’ll pay attention to this and not to that, or whatever.” And I think at least there’s one outcome measure – you know, did they get the right
answer or not – is something that I think has been – has been overlooked.

The second comment is in designing these kinds of studies and writing proposals to do research on the method itself, as Gordon suggested. There’s all these different things in motion. You know, you can do probing different ways; you can do think-aloud or not think-aloud; you can give interviewers different amounts of latitude; and you have different types of questions and all sorts of things that are moving.

And I can speak as one who’s been turned down for funding many times by NSF, that is very easy for a proposal to kind of be picked apart based on sort of, you know, what things you chose to look at.

And I have no idea how to go about this, but to the extent that some theoretical support can be developed, or some ideas about theory as to what should you be looking at, and where should you – on the basis of not just intuition, or sort of how you feel that morning – where should you expect, or what kind of results should you expect when using these different things?

And I said two things; I actually want to throw in a third because the other thing occurred to me earlier in – I forget which of the talks. But the first time that I was involved in a study that used cognitive interviewing, which was so long ago I’m not going to actually cite it, but it was about proxy reporting, and it was about how do – the ways that people come up with answers, process information to come up with answers about proxy reports – you know, reports about someone else – differ from how they derive answers to requests for reports about themselves.

The key thing here is that we use cognitive interviewing that is a tool. The study was not about cognitive interviewing, but we learned a lot part ‘cause it was so long ago nobody knew very much. But we learned a lot about cognitive interviewing, and in retrospect, I think if we had set out, as part of this study design, to learn some things about how cognitive interviewing works, we could have learned a lot more.

And I’m wondering if one avenue that – and I don’t know if this would affect, you know, the likelihood of funding or not, but is to look for cross-disciplinary kind of collaborations, where you’re working with someone who maybe has somewhat sort of a different interest, a psychologist or education person or whatever, and cognitive interviewing is part of that, and building in some sort
of embedded experience that’s related to the cognitive – you know, the cognitive interviewing part. You know, just that I have no idea whether that would affect the likelihood of funding, or how that works; it may.

But there also may be, I think, some usefulness in working that way, in limiting the domain in a particular study, in terms of the subject matter that you’re looking at. It may be other kinds of constraints.

I know my tendency, I don’t know if this happens with others, you try to do everything, you know, ‘cause you want to have enough different kinds of things and do enough different kinds of things that you can say, “Yeah, this’ll work forever, in any circumstance. But there may be, I think, some advantages to also doing some of the other kind of thing, where you’re really looking at a more limited domain, whether it’s in terms of population, subject, or whatever.

Gordon Willis:
Okay, yeah, Johnny, you said a mouthful. You picked a couple of different themes that are important. I like your notion of criterion validation, I would label it, where you have an actual standard that we can compare to. It’s one of the things I was thinking about during the talks on voting. You know, if you know whether someone’s voted or not, that gives you a criterion, although I understand it isn’t perfect.

So, we – where I work, we’re interested in using medical records perhaps as validation, and that might be a good way to test different versions of our health-related, you know, cancer risk factor questions. Although there is a cautionary note here, too, which again has to do with the quality of records. That’s a conundrum, and I’ll just bring an anecdote out, just to keep it alive in the survey methods arena here.

Charlie Cannell, years ago, told us – he used to tell a story about where he was trying to do one of the earlier record check studies. He thought of this, too, way back in the, like, ’60s or ’70s. Well, he made an arrangement with some physicians in a hospital to get access to their records. And so, then they could bring in patients and ask them various forms of questions, and then access the records.

And they went into all the nuts and bolts of how they were going to put this stuff together, and finally, you know, came to an agreement. At the end of the meeting, one of the physicians said,
“This is great; we’re finally going to find out how good our records are.”

And he said, “Ooh, something isn’t right here.” So, that’s the problem we’ve got, like what is it we’re doing to validate what. So, I run into that a lot, and I’d like to do more validation, but it really is – it’s hard finding good validation measures. So, if there’s a possibility of funding something like that, I think that could be really vital.

And I could respond to the other things you said, Johnny they’re good points, but I think there are other questions which we’ll entertain.

*Jon Krosnick:* If I can just follow up just quickly on what you said about validation, and that is I think to look for opportunities where you may have multiple measures of validity as far as the correct answer, maybe debriefing the respondent, looking at other records that the respondent has, in the case of some of these health things, as well as frame variables. That is, if –

Obviously we know that records are imperfect. If you validate on a couple of sources, I think you’ll feel a lot better than if you validate on one. And if, by a miracle, you can find three, that’s even better. But yeah, there are limitations built into that, but, you know, I think they’re – the tradeoff is would you rather try to deal with the limitations or not know sort of if people are able to answer the questions?

*Gordon Willis:* Well, yeah, I should throw in one other thing here, too, which is through this meeting, we’ve mainly been talking about behavioral measures, about things that are observable rather than latent constructs.

There’s this whole other side, too, where if we want to validate based on using latent constructs, then we can use scales, and we can bring in psychometrics into this mix, too, and try to evaluate cognitive interviewing by saying, “Well, can we predict which of two versions of a scale is going to produce a more reliable, valid instrument based on – whatever, a Cronbach’s alpha, item response theory analysis, differential item functioning – all this stuff we can pull in from the psychometricians, and I think that could be a really useful research direction, too.

*Mark Liberman:* Just one quick thing. I’m Mark Liberman from Yorktown, Pennsylvania. The issues in this area remind me very much of
much of the work in human computer interaction, and there’s enormous literature on that subject, and it might be worthwhile to attempt some cross-fertilization.

It’s especially reminiscent, I think, of the issues that arise in trying to design automated conversational systems, where small differences in the way questions are phrased, small differences in the menu which is followed, and so on, lead to large differences in success.

*Gordon Willis:* Yeah, especially as we do tend to migrate more towards computerized, self-administered types of administration procedures and away from interviewer-based telephone and things like that, then we are going to have to –

*Mark Liberman:* Well, then the two things sort of converge, in a certain sense, except that the criterion of success in one case is, “Did the user get what they want,” and the criterion of success in the other case is, “Did you get what you want?”

*Gordon Willis:* Okay. Jennifer?

*Jennifer Romano Bergstrom:* Jennifer Romano Bergstrom from Fors Marsh Group. I have a few comments, and I promise not to talk too long. I just wanted to follow up on one thing that you had mentioned about collaborating with psychologists and other people who do cognitive interviewing. I think that’s a great idea. We oftentimes see people in cognitive testing and usability testing who coach, who lead participants, and then you obviously get very different results. So, I think that is a really good idea and something that, if resources are available, that should be sought out.

Human/computer interaction in cognitive interviewing, that’s something that with usability testing, whether it’s human/computer interaction, or even paper surveys, I think that’s something that I’m personally trying to do more of, doing integrated usability and cognitive testing.

There are issues, obviously, with that. Usually in usability testing, we don’t interrupt them and ask them, you know, what they think about things as they’re going along. In cognitive interviewing, we do that a lot throughout the entire survey oftentimes. You showed some citations where it’s retrospective, but we know there are issues with retrospective as well because memory is fallible. So, if
you wait till the end of the interview to ask people what they thought about a certain item, there are issues there.

And now for my question. In terms of the empirical testing and some of the studies that you’ve talked about, and that I’ve read a lot about, they seem to focus more on the issues that different researchers discover. This is something that I’ve been looking at lately in terms of figuring out the proper sample size, and there’s very limited research that’s been done on that.

What do you think about, instead of focusing on the number of issues that are discovered, since that is really subjective – for example, one person might say there’s one big issue, and it’s a navigation issue. And there’s a little bit more usability.

Somebody else might divide that up into three different issues and say, “You have an issue with your top navigation; you have an issue with your left navigation; you have an issue with this other navigation.

So, it’s really subjective, and when you tend to count up the number of issues different researchers have found, I think there’s a lot of issues there. So, what about, instead, quantifying the actual performance? So, more usability metrics, whether or not people can actually complete a task. So, focusing on these type of metrics instead of the – the more subjective issue.

Gordon Willis: Yeah, the notion of coming up with metrics based on performance and success, makes a lot of sense to me, especially in the usability area. Where we’re dealing with a Web survey, then what I like is there’s certain things that we can measure which are pretty much – I don’t know if “absolute” is the right concept or not, but we can give them a problem-solving task and say, “You know, how would you do this, that, or the other thing,” and see if they can do it.

And that’s a good test of the human/computer interaction, the interactability of the system – right? I don’t know how we do that as much, necessarily, with a paper questionnaire, or an interviewer-administered questionnaire, you know, what metrics we could use there. If you thought of some for a paper questionnaire, that’s great, but I’ve never been able to come up with much there.

Jennifer: I think for a paper survey, it’s very similar to a Web-based survey, when you’re testing usability. It’s still whether or not they can use it, whether or not they can work through the survey. If they’re reading the – if they’re following skip patterns, if they’re using
response options appropriately, I think it might be a little bit different when it’s interviewer administered.

**Gordon Willis:** Yeah, I actually agree with you with respect to the paper-based administered questionnaires. Yeah, for years we’ve been paying a lot of attention to whether the person can get through the questionnaire, whether they blow the skip patterns or not.

These are common metrics. It’s one reason that I prefer, for a self-administered questionnaire, to do the probing after the fact rather than mucking them up with concurrent probing through the administration of a questionnaire.

And people who have done Web surveys, too, purely from the usability point of view, have told me that they look at it exactly the same way. Often they’d rather let somebody go through, watch them, see what they can do, and leave their probing till the end.

But I know that’s an issue in Web surveys the same way it is within cognitive interviewing. I feel like there are these two sciences here that have been pretty much independent, not talking to one another, and it’s time to bring that together, and that seems like an obvious avenue for research to me.

**Gary Langer:** Can I just give one idle, frightening thought? Just – it’s Gary Langer, sorry – I’m just – it just occurred to me, in interviewer-administered surveys, has there been research into not how the respondent understands the question, but how the interviewer understands the question, and what impact that may have on the data collected?

**Gordon Willis:** There are a couple of ways we look at the interviewer side of it. One is with behavior coding, it’s more mechanistic, where you listen to whether they’re asking the question correctly or butchering it or think they have a better way to get at it.

But you’re asking more in-depth about what does the interviewer really understand about the question. And that, I think, is one of the areas we typically do not get at in cognitive interviewing. We don’t look at the interview side enough.

So, therefore, one of the things that we have done is to convene – I guess you’d call it a focus group of interviewers, to get together and talk through the questions with them, either before or after they’ve conducted a field survey. So, yeah, I think that cognitive
interviews of interviewers make as much sense as they do of respondents. And I think I better get off the stage.