Jon Krosnick: Well, in order to stay somewhat on schedule, I think we’ll call this session to a close and transition to our first remote presentation. Skip Lupia is co-principal investigator of the American National Election Studies, and has been a co-director of a program of work going on for a couple of years now on open-ended coding of data. So he was asked to teach in Ann Arbor today, so he is going to be – that might be his tie in the upper right – no, there is his head. So there’s his tie. And so we’re going to – Dave is making the technology work here so that Skip can see and we can see him.

Jon Krosnick: So Skip is a professor at the University of Michigan and is at the Institute for Social Research. Skip, thanks for doing this.

Skip Lupia: So thanks – I’m sorry I can't be there today. I have big undergraduate class that I have to teach in the afternoon, so – and they like having me there. What I’d like to talk about today is open-ended coding and some things that Jon and I learned while were principal investigators of the American National Election Studies. What I’m going to tell you about is based on the work and experiences of a lot of different people, including the folks up on the board.

I understand that Matt DeBell is in the room. Matt Berent, who has also been very helpful to us. I understand that there might be an Ascribe representative in the room, so I’ll talk to you about that later, and lots of great staff. An outline what I’m going to cover in the next few minutes is some background and challenges with respect to the question of open-ended coding. I’ll give you an example that was our trial by fire in this domain.

I’ll talk about some general attributes of the approach that we have tried to develop and are taking at the ANES to make open-ended coding something that’s more legitimate and credible than it has been. And time permitting, Jon, I’ll ask you for a cue when I get to that point, whether we want to cover the second example, or you may want the Ascribe folks to say something. But when we get to that point, just let me know if you want me to cut it or continue, and then a brief conclusion.

So the ANES, I’m not sure who’s in the room. So the brief overview is it’s widely considered a gold standard of election studies. Its origins date back to the University of Michigan in the 1940s. And our objective is to try and provide data that can facilitate lots of different hypotheses in political science and the social sciences pertaining to elections, not just the political aspect of it, but also the neat scientific opportunities that arise when you have hundreds of millions of people making a comparable choice on the same day.
So now Jon and I were the principal investigators from 2005-2009, and now it’s capably run by new principal investigators at Michigan and Stanford. Some of the experiences that I’ll tell you about pertain to the 2008 version of the American National Election Studies time series. So the main study that the ANES is known for is a time series that takes a core set of questions, asks them every election period, and then adds new ones as circumstances request.

So as you can see in 2008, we have a – each version of this survey happens in two waves: once in the two months before the election, and once about the six-weeks afterwards. We try to get in the field for the post election wave as soon as possible after Election Day, and really try to clean it up before people start going away for the holidays. In 2008, we had about 164 minutes of interview time, and so lots of questions.

Most of the questions we asked are close-ended, you know, typical and very few response options. But there are a couple of questions we ask where we want the respondent to give a verbatim response. And so some of those examples would be: “What’s the most important problem facing the nation? Tell us things you like and dislike about the candidates, and then tell us who you like and dislike about the party.”

Probably one of the more used and famous open-ended questions on the ANES has to do with other responsibility to recall certain things. So here’s the beginning of a question that’s been asked for close to 30 years, and you can see now we have a set of questions concerning various public figures. We wanted to see how much information about them gets out to the public. And then we’ll name one.

And so 2004, for example, one of the questions was: “What about William Rehnquist? What political job or office does he now hold?” So that’s the type of question that we’re dealing with. The users of this survey have some expectations about what we’re going to do with answers to this question. So ANES users expect us to convert open-ended answers to numbers. These numbers are then typically used in correlations, like regressions and things of that nature, to make inferences about all kinds of things about choice and the election.

So people take the numbers that we produce and they draw an inference. And the critical thing about this process is that the users base their inferences on beliefs about what each of these numbers that we product means. So that’s the key here. The beliefs that people have and are they [inaudible]? Now when we think about the beliefs that people have about these numbers, many people believe that open-ended coding is easy to do, it generates valid measures, and it’s performed well by survey organizations.
In fact, for the 30-years that we’ve been doing the recall questions, we had little to no record of users asking us questions: “Well, how did you make decisions about what answers were correct or incorrect, or things of that nature?” No, we don’t have a record of requests for things like reliability statistics or so forth, and, yet, the questions are widely used. And so this represents a belief in the user community about what I have in front of you.

Now while Jon and I were principal investigators of the National Election Studies, we discovered a different reality in terms of how numbers were derived and what they meant. And a lot of what we learned surprised us and disappointed us, and so we started a further investigation. What we found was the proxies that the National Election Studies were using to produce numbers were not unusual.

So here’s the fundamental question in front of us, and then I’ll tell you about some of our experiences. What is the correct inference for a user to draw from an open-ended response – from a coded open-ended response to a survey question? And the answer to that question is going to depend on what question we’ve asked, what the respondent says, and then some decisions that are made after the interview is conducted about converting those words into numbers.

And so most of the emphasis now will be on those decisions and what we know about them. Now I have a goal in mind for how we might think about what we should do. And the goal is to produce measures that are credible and legitimate. Okay. So by credible, I mean that the numbers have some property that make them believable, that make people trust that when see a number, they actually know what it means. So a lot of social scientists are interested in credibility in their own inferences.

I’m also – I think we’re also interested in legitimacy. That is when people want to know – people have questions about, “What does this number mean?” If you can say, “Well, we derived this number using a set of principles and standards that we can defend and that you might be willing to accept. So that when you go run regressions, or whatever, the number means something.” So legitimacy is also something that we’re aiming for. So those are the big picture goals, and the things that the ANES tried to do to achieve those goals was to increase procedural transparency. As you see, that was badly needed. A lot of people in the user community really didn’t understand how the codes were produced. The second thing is documentation rigor. And one of the reasons that people didn’t understand this is because our study, and lots of studies like us, we didn’t really write things down.
In terms of the instructions that one person gave to another person, in the sequence that led from words being converted to numbers. So by doing those things, we hope to increase credibility. Well, I’ll take a moment now to tell you about our trial by fire. And that has to do with what political scientists call political knowledge. So people have a lot of beliefs about what the public knows about politics.

A quote I have – hopefully you’re seeing it on the screen now, is representative of a wide field of critique that’s says close to one-third of Americans can be categorized as know-nothings, which is not to say that the other two-thirds are well-informed, right. And so the basic idea there is that we have a knowledge base, where we ask people pretty simple questions, and they appear not to be able to give correct answers.

So this is not just a popular thing – Bob Luskin is an academic who studies this as well. He uses data such as what the National Election Study does, and this is not an atypical verdict for scholars like Luskin. So what is the knowledge base from which this type of claim is drawn? And the National Election Studies supply a template for a lot of people, and the actual data for others.

So I’m going to show you a question now, and I’m going to do it in the context of a study done by Jim Gibson, of Washington University, and Greg Caldera. And their work really brought to Jon and my attention a big challenge we had with open-ended coding. So here again is a question from the 2004 National Election Studies about William Rehnquist. And if you look at the dataset, what was reported, you know, and how people used that data.

In the 2004 National Election Studies, 12 percent of the people in the sample were coding as having answered this question correctly. So that’s interesting. And this type of result confirms the type of claims we saw earlier about how ignorant the public is about these general thoughts. So let’s think about where that 12 percent number came from. So prior to 2008, the way that this data was released to the users was we asked the recall questions, that’s what you just saw here; that’s a recall question.

The recall questions were asked in an open-ended format, but only numerical codes were released to users. Prior to 2008, the transcribed verbatim responses, what the respondent actually said were not released. We had beliefs about what would happen if we released those things; there were some concerns about privacy, and so the longstanding policy of the National Election Study is you only release the numerical codes.

But we did have a way for people who are interested in seeing the verbatim response to do through a restricted data access program to
basically sign their life and their children away in exchange for being able to look at some of our data that we don’t release. So Gibson and Caldera did that. Because at the same time that we ran our study, they had run a study about public knowledge of the Supreme Court, and they had asked a question just like this one.

And, yet, the answer that they got was very different than ours. They found upwards of 40-50 percent of the public who could answer this question correctly, and they couldn’t understand why they got such a different answer than the National Election Studies. Well, then they told us. So when they looked at the analysis, what they determined was that in 2004, a respondent was graded as answering the question about William Rehnquist correctly only if their answer said both chief justice and Supreme Court.

So a respondent who said only one of those phrases was graded as incorrect. And it turns out that another 30 percent of respondents actually identified William Rehnquist as a Supreme Court justice, right, but were marked as incorrect because they didn’t say both chief justice and Supreme Court. Now what’s interesting is in Gibson and Caldera’s study, they looked at different ways of answering the question. And just to give you some sense of what the 12 percent number means, I’m going back here.

The one that the ANES released; at the same time that they ran their own question, they also ran a multiple-choice question. You can see it here asking whether – who was chief justice at the time; was it William Rehnquist, Louis Powell, or Byron White? Seventy-one percent of people got that answer right and could pick that out. So this type of result raised some serious questions for us about what was going on.

We went back and looked at the 2000 ANES, where the same coding scheme was used. And we found that of the nearly 1,600 respondents, close to a quarter of either set that William Rehnquist was a judge, or that he was on the Supreme Court. And, yet, if they didn’t say both things, they were coded as having answered incorrectly. Here is a sample of other incorrect answers from the 2000 ANES about William Rehnquist.

So you can see we have a problem here, but this wasn’t the end of our problems because Jon and I then started trying to figure out how is this happening? What was the process that led to this type of answer? And in the process of lifting up rocks, we found another error. So one of the other identification recall questions asks about Tony Blair, remember him? There he is. And the question is: “What job or political office does he now hold?”
Now what’s interesting about this question is that people who code the data are given some instructions. We’ve looked for instructions and we couldn’t find very many, but we were able to find a paragraph about what to do in this case, and the paragraph was changed in 2004. The paragraph read here: The reference must be specifically to Great Britain. The United Kingdom is not acceptable because Blair is not the end of Ireland.

Okay, so anyone in 2004 who said that Tony Blair was the prime minister of the United Kingdom was marked incorrect. Perhaps from the colorful map here you can see that there’s a problem with that. So now this is pretty serious, and our question is how did this happen? How did this happen not just once, but over a period of multiple years? So the typically ANES coding practice, we have an interviewer who asks the question and they transcribe the response.

One thing that we can talk about in the Q&A is the quality of these transcriptions as went through various things that we’ll tell you about, we actually found the transcriptions to be of varying quality. We have actually attempted to record the interviews, and we think there might be some virtues there in recording all of these interviews so that we don’t have to deal with the variation in transcriptions.

But, anyways, after the transcriptions occur, several weeks after – you can see we have a very short field period, so all the energies of focusing on collecting the data, and then at some point afterwards, typically a couple of weeks or months afterwards, a staff implements a coding scheme. Now as we looked into this coding scheme it was reminding me of people looking for weapons of mass destruction in Iraq. We kept expecting to find the documentation about exactly how this was done and what instructions were followed, what validation – how things were validated.

But what we found was really little to no record of instructions to the staff. Basically the way that the coding was done was staff members would hire undergraduates or other people and tell them verbatim, you know, they’d give them a list of codes and then basically tell them how do it, and then make a judgment about whether they had done it well. And there was really no documentation of reliability analyses; there were many cases in which the coding was done by a single coder.

So this was problematic, and we had an initial response. The first thing we did was send out a letter to the user community basically alerting them to the situation and talking about what we would do next. So what we would do next is based on the principle that you see on the screen now. A basic expectation of scientific research is that you be prepared to document, archive, and share everything so that what you do is available
for scrutiny by others. I think this is how you develop legitimacy and credibility.

In this respect, we’re channeling Richard Feynman – my background is from CalTech so I met him long ago. But Feynman at the 74th Commencement Address talked about another honesty, and the idea of giving people all the information, to help people judge not just the judgment that you make, but how you got there. And this type of philosophy was really what we started to aim the coding process towards.

So what we did immediately is to the extent that we could for 2008, we made redacted transcripts available. And the reason we were able to do this for 2008 is because we had written the consent form in a way that would allow us to do this. We would have liked to, and may still like to release earlier transcripts, but there’s a legal concern about whether our consent forms from those years actually – whether the respondents would consent to this. Since it’s not there, we haven't released them.

We approached NSF, who was incredibly supportive and realized the types of problems that not just we had, but realized that a lot of people had problems with documentation. And we ran a conference in Ann Arbor bringing together a number of people that are in the room right now, some machine coding people, linguists and things like that, to try to work through and develop better best practices for coding. And then we started to assemble expert committees for each of the ANES open-ended coding for it to have them help us identify mutually exclusive and collectively exhausted coding schemes.

So when we started to look at some of the coding schemes that the ANES was using, we found big piles of codes where we were actually questioning where we would put certain responses because you’d have 10 or 15 codes that to us looked pretty much identical, and then you’d have wide spaces of uncovered territory. So we tried to get expert committees in each of this issues to help us figure out a scheme where we could have, as close to possible, as to one-to-one relationship between the words that people say and the numbers that we produce.

We also wanted to be replicable. One of the consequences of what we found is we delayed a number of years, the release of the open-ended codes. And we’d questions of people saying: “You know, I really need the codes to make some sort of – to compare 2000 to 1996 and so forth. When are the codes coming out?” And one of the things that we had to say in a constructive and friendly way as possible is: “We don’t have the document – there is no documentation from previous years. We don’t have the instructions from previous years.
We could try to pretend to replicate what they’ve done, but really I don’t think we can do it. So we’re trying to put forward a set of materials that would allow people to replicate whatever it is that we produce.” And you’ll see how we’ve done that. Now our first committee went back to the recall questions. And so it’s interesting that a big part of that debate had to do with which responses should we count as correct; which should we count as incorrect, and how should we deal with partial knowledge?

Now what’s interesting about the question, “What’s the job or political office,” you know, when you ask the question about Dick Cheney in 2004, a lot of people say “vice president”, but some people say “anti-Christ”, and some people say “chief puppeteer”. And the question is what do you do with that? And so a lot of people wanted to count things like that as some sort of knowledge.

[Laughter]

So this debate actually went on and on within this group, and it was the focus of here’s what we should do with the coding scheme. And for reasons I’m happy to answer in a Q&A, it’s actually not that simple to find a way to grade partial knowledge. There are a lot of – some slippery slopes in there. So I’d say that we’ve reached a breakthrough when we actually asked the committee to do something, which you wouldn’t think you would have to ask a group of users to do, which is go back and read the question.

Because there was a temptation to use the question to try to – you know, if they say, “Nancy Pelosi is from California,” we count it. But the question doesn’t ask you to free associate about these individuals. It asks you, “What is the job or political office that the person now holds?” And so as we started to think about how to code that question, we want to code it with respect to this question.

So if you say that Dick Cheney shot his friend in a hunting accident, you have decided not to answer this question. You may be giving us knowledge about Dick Cheney, but now we’re getting, maybe, a bias sample of what you know. Because to get that type of information, we have to a) count on you not to answer the question that we actually asked, and then b) work on some other things. So this whole concept of partial knowledge was changed by, well, it partial knowledge with respect to this question?

So the coding framework that we now use basically first focuses on political office. Did the respondent say something about the political office? Did they identify any part of the title of this person’s political office correctly? Some people have multiple offices; for example, the vice
president of the United States is also president of the senate. The person who is the speaker of the house is also a congressperson.

So for people like this, we’ll take any mention of a title that they actually hold as correct. We also have a code for people who identify part of the title correctly, but only a fragment of it. And didn’t it say anything incorrectly? And sometimes we wanted to distinguish that from someone who said part of the title, said “speaker of”, and then said ______, if they were talking about Nancy Pelosi. You’d want to distinguish that type of thing.

So the first thing we focused on is did you say anything correct or partially correct about the political office? Since the question asks about political officer’s job, a correct answer to the question that was asked would also constitute descriptions of what this person does, of what their job is, maybe making legislation or organizing a political party, and so forth. So for each of the questions we asked, we’ve gone through, you know, textbooks and things of that nature to identify a long list of jobs that a particular person has.

So these are now – unlike before where we basically just gave a correct or incorrect code for each question, now we have a political office code: Do you give a completely correct answer or a partially correct answer? We have do you give a complete description of a job or an incomplete description? And then finally there’s “other”. And “other” is anything that the person says that is not pertaining to the job or political office of that person.

Now with respect to the other responses, what we’ve done with them is we haven't made any judgments about them. What we have a code saying “other”. The truth-value of the various claims that are made is you might think it’s easy, but there’s, you know, if somebody says, “Nancy Pelosi is a liberal devil,” you know, what is the truth-value of that statement? And so we don’t have a code for a partial correct, we have a code for: Did you name the political office, did you name the job, or have you said anything else?

And now the verbatims or at least the transcripts for the verbatims are in the public dataset. And so if someone wants to go into the other responses and create their own variable for what they think is partially correct, they can do that and the field can argue about it. But, again, we decided not to go this way because I think the idea that I was trying to code anything you say about this person really goes against the question that was asked. It is not a good question to elicit free association
So if what we want is general recall, we’d need a different question. So the main attributes of this coding scheme is that it’s theoretically defensible. I mean you could think of different ways of coding this, but I think in this case, we can go back to a set of first principles and say, “If you want legitimacy and transparency, here’s how you get it.” It turns out that when we wrote the instructions for this and explained it to coders at a scribe through an iterative process, we had very high inner coder reliability.

So we have some confidence that if we showed these instructions and this data to any human, let’s say, in the temporary United States, we get the same seven numbers. This coding scheme has the property of being mutually exclusive or collectively exhausted, so you can't both give a partial – you can't both be right and wrong about political office unless we put you in the category where you said one right thing and one wrong thing. But if you’re in that category, that’s where you are.

And, again, scholars can use the public data that we have available, the transcripts. If they want to think of a variation of what we’ve done, they can do that. So with that example in mind, let me just talk about a few general attributes of the approach, and then I’ll ask Jon for the cue as to whether I should hand it off or go somewhere else, or show you a second example.

So what we did is for each of our open-ended questions, we first tried to identify a theoretical framework. And so the basic idea there is that words don’t define their own categories. These words, when analysts want to use them, are with respect to some sort of theory about how the world is organized and what concepts are important and so forth.

So what we did is we really leaned on and worked with the expert communities to figure out is there a consensual theoretical framework from which we can develop a set of categories that will be analytically useful to a large and general population. And then we would take that theoretical framework and develop the code frame, develop the relationship between numbers and sets of words. And these are things where with a theoretical framework; it was really an iterative process with the experts.

On the code frame, we would take what the experts gave us and then do the math ourselves, and then bring it back to the experts and say, “Is this an implementation of what you spoke to us about?” Then we would go through a process called chunking. And the idea of chunking is if I give a longer response to a question, what we do is try and break it down to discrete utterances and discrete thoughts to allow for the fact that people may say different things during an answer.
The chunking, we developed that – you know, we hired a vendor, Ascribe is in the room so they can tell you about what they do, but we would say, you know, we don’t want a general impression of an answer, we want as precise, as possible, a numeric representation of each thing that a person said in their answer. So we have a chunking framework, we developed that in cooperation with the vendor in terms of how to implement it, and then finally we get to coding.

That as we take each chunk, each utterance that can represent an idea, and turn that into a number, right, and that is executed by the vendor, where we do – that’s not executed by us. We develop with the vendor a set of rigorous instructions and things of that nature so that the coding can be implemented by a human being that is really at arm’s length from us, which is a way to validate whether our code frame and whether our instructions really lead to a set of numbers that don’t depend on having come from us, that depend on it coming from a logic that people can understand.

So that’s the sequence. Now what we want to try and maximizes legitimacy and increase transparency, we want to document everything. And that was really our goal. So what we thought was to document our written correspondence with working groups. Our goals are we have – most of the – through the beauty of the Internet, a lot of our communication was electronic and we have a written record. We’re hoping to make so much of that available so people can see how the theoretical frameworks were developed.

We have written correspondences with the vendor about implementation. We’d like people to be able to see how the code frame was implemented so they can figure out whether our practices led to – whether there is something about our practices that skew the numbers. Written documentation of all decisions, right, of every decision we make when we ran into trouble. Written documentation of all conversations we’d have, like with the vendor when the – maybe we’d run instructions and then coders would be confused by them.

We’d like people who are using the data, if they want to, or people, who are developing their own coding schemes, to be able see what did other people do. And if I’m getting different results than the ANES, is it because of a difference in practice? And then we also want multiple independent assessments of the decisions. So what we have now, you can notice in italics I have “ideal”. Actually producing this is incredibly time-consuming.
And so I think the idea is this is very time-consuming and expensive, but what we tried to do is approach it as quickly as possible. Our constraints – we actually – we decided to unshackle ourselves of the time constraint. We did not rush to get the codes out, and so our main constraint in doing this was really financial because as Ascribe will tell you, we actually asked them for all of this. But it’d be very hard for them to do what they needed to do with the money that we had for them and to supply this as well.

But this is – I think we want to inspire to complete procedural transparency, this is what we want to try to accomplish. But, for example, to give a sense of just how thorough we’re being, if a coder had a question, what we asked Ascribe to do is not have that conversation verbally; we asked them to have it in writing so that we could see it and record it. So that in the future – like in 2012, when the ANES is trying to do their coding scheme, they can see what types of challenges we had and perhaps decide to make the same decision from the same logic we had, or perhaps decide that in a different logic.

But to the extent that conversations like the one following this question are hidden, makes replication harder, makes comparability harder, makes it harder to answer questions about what does that number actually mean? So what you see on the board now represents, I think, our current practices. We have much better documentation at every stage of this process. We have evaluations of various kinds at every stage of this process. We have increased procedural transparency as a result.

And because we’ve had great communities and great vendors to work with, we’ve actually accomplished through these interactions, very high inter-coder reliability, statistics that we will be releasing in lots of different ways.

Skip Lupia:

Okay. So the second example I want to talk about is with our most important problem question. And here the challenge is going to be – I’m not going to go through all the various problems, but the question here is going to be: How do you develop a code frame for a question where there’s very little guidance? So it’s more kind of how you develop sort of the theory.

So here’s the question: “What do you think is the most political problem facing the United States today?” Variance of this question had been asked on the National Election Studies for thirty-ish years. I don't remember its origin. But it’s a question that a lot of people are interested in. It’s a question that is copied on National Election Studies and other studies around the world when they try to study elections.
Now when we started looking into the coding scheme for the most important problem question, we identified some challenges. One was that the coding categories seem to vary a lot from year-to-year. There would be codes dropped, codes added. And, again, because of the lack of documentation, we really couldn’t figure out what the criteria were for adding things.

So, for example, here’s a code frame education that was used in 2000. So I don’t remember what the number was – 837 would be the code for education. And so the idea here would be if someone had mentioned financial assistance or quality of education, or so forth, where education they would be this code frame in 2000. Now 2004, same number, same code is used, but also was added a different modifier also to include the high cost of college.

There are lots of examples like this, but the point is these code frame were shifting from year-to-year. Now if this was one off survey, where people just looked at a particular year and then left, these changes wouldn’t be a problem. But one of the most common uses of the ANES is to compare from year-to-year. And so this type of variation, particularly if users don’t understand that it’s happening, can cause big problems cause of the fact that education is chosen by 8 percent in 2004, and 4 percent in 2000.

And this case could have something to do with the decisions made by not the people who wrote the question or the people answering them, but people six weeks later. So when we looked at the 2004 code framework, there are 154 categories, those are how many places you could put a – how many different buckets there were to enter a representation of what people said. We saw no clear theoretical framework. There was no hierarchical logical organization of what was happening. Or there was, but it was really diffused.

A lot of the categories, as I mentioned before, were not mutually exclusive and collectively exhausted. So the notion of any sort of one-to-one or one-to relationship between what was said and the number, it was impossible for us to see how that could have happened in the past. We couldn’t find written instructions. We did not find validations statistics. I spent the better part of a year on the Institute for Social Research in Michigan – and a lot of great people work here.

I spent the better part of a year trying to find this, like Jon and I calling like every person in the building, who could have had a record like this. They don’t exist. And what we also found is when we looked at how people used these categories; actually 154 aren't used. What we found is most users trying to take the initial categories, convert them to seven mega categories, and then using that. And it’s also worthwhile saying in 2004;
only 14 of the 154 categories had more than five non-zero values. So in terms of this coding scheme, its effectiveness or efficiency, I think there are some problems there.

So here we were really [inaudible due to audio cutting out]. We had an expert committee. We talked to them and we weren't coming to a consensus about what we should do. But one thing that did come up is lots of people ask the most important problem question. So what the committee and Jon and I asked Matt Berent to do was conduct extensive interviews of other polling firms that made great use of a most important problem questions. And you can see the groups we talked to on the board. And they were wonderful.

Now a difference between them and us, of course, is most of them are using these questions in one off surveys. And so the comparability over time, I think, is not the same issue for them. So what we found when talking to them is that they have very limited code frames and did not have anywhere near 154 categories. I don’t remember what the medium number is, but it was very small; it may be 15. That’s my hunch. Jon, is that close?

*Jon Krosnick:* It depends on how you count.

*Skip Lupia:* Okay. All right, 15-ish. When we talked to a lot of the groups, where wasn’t a set rule about how codes were added or subtracted? And, in fact, a common thing we heard is that the organization would collect the data first, and then develop the code for frame later, which, again, for that one off study is reasonable; for us, really not plausible.

What we didn’t see in these houses, no one reported to us systematical analyses of coding reliability either way. It may have happened, it just didn’t come out in those conversations. So this was the challenge for us then, like what do we do, given that we have a code frame in the ANES that people aren't using most of it and that we really can replicate.

So we wanted to start with the idea that if you’ve got a coding scheme of a question like this, it has to be defined with respect to some theory of language and some theory of meaning. And so as the result of a lot of conversations, we once come up with, again, a mutually exclusive and can collectively exhaust a frame that is stable. And that can be used over and over again every four years that is replicable, that if we wrote this the right way, other people could people take it, use it, and get comparable answers to us, and it reflects common theoretical concerns in the discipline.

And so we came up with the following basis: Federal budget categories. In the new most important problem-coding scheme, the main eight
categories reflect the main eight categories in the federal budget. And I’ll explain why that is in a second. Then there was a secondary source, and that was called the rule of two. And what we did there was looked at the other main credible houses that ran the most important problem question, and just try to figure out what do they ask reliably, like what codes did they use reliably to communicate to their users?

So this is a hierarchical thing where our basic categories are federal budget, and then we used the rule of two. With respect to the federal budget categories, what’s interesting about them is, and I’ll show you an image of a federal budget page in a second, but these categories have been stable since 1940. That is if you look at the federal budget and the main categories were all things are organized, the federal categories are actually stable; they don’t change.

Some categories change because the government evolves, but the main things that the government – at least labels how it labels what it does, has to remain constant. Every federal governmental program and activity is listed within this framework, or at least everyone that needs funding, which is just most of the ones that we pay attention to. And so when you look at this budget and you look at the subcategories and super categories, what you find that is that all major federal government functions are categorized.

And the reason that is relevant for the National Election Studies is, of course, when you ask voters around an election, what is the most important problem facing the nation, they’re usually thinking about things the government should do or shouldn’t do, and that’s the type of thing that a lot of them are naming. So this is a code frame – I don’t know what the resolution is going to be like, but this is actually a cover page from a recent federal budget document.

And you can see these numbers that end in zeros – hopefully you can see them. There are these labels in bold font; they end in zeros, those are the federal budget super categories. For our coding scheme, ours are identical. So that’s the basic framework that we work from. And now when we get down to the subcategories, these are – now we’re getting to a level of details where most of our respondents don’t join us. And so that’s where we try to look to how people were using this data, both our users and the major polling organizations, to help us fill in subcategories.

So our rule of two was we took any category that was used by two or more of the ANES and then the other groups up there. And if they had used a category repeated times regularly, that would be one of our subcategories. That would be something that we would take and place under the federal
budget. I should just say that not everyone, when they talk about a most
important problem, mentions a government activity.

And so we had another nine category, you know, starting with 900, which
is “other”, and that would mention those types of things that end up in the
rule of two. Now one thing I can say about the advantages of this code
frame; again, we have aimed for mutual exclusivity and collective
exhaustion. This is derivable from a transparent logic – of course there are
others. But one of the nice things ex-post is with this code frame, we have
developed very high inner coded reliability, and I think a set of concepts
that users are more likely to find useful.

There are other ways to do this, right. So what we have is one that is, I
think, is defensible and replicable. So the last thing I want to say is the
challenge for us is that documentation and validation of what we do, of
how we turn words into numbers, it’s time consuming and expensive. I
think Jon would agree, I mean some of it was frustrating, but actually a lot
of it was fascinating. Personally, I learned a lot in the process from lifting
up rocks to producing new numbers.

But I think the bigger picture, you know, if we go back to the conference
and how many people had the same kinds of challenges that we did, is that
we benefit in terms of legitimacy and credibility, and society benefits in
terms of their ability to trust what we say. If we had rigorous public
accounts of how we produce those numbers.

Because at the end of the day, the numbers we produce are a function of
what we write, a function of our samples, a function of what respondents
say, and then a function of a lot of decisions that we make that at least in
our case we hadn’t written down. So that was leading to people making
erroneous inferences about what we were producing and putting their own
credibility in that. But we’ve really tried to turn that around and we’ve
had a lot of support in doing that particularly from the foundation. So I’d
like to acknowledge that in where we ended up.

Thank you Skip. So I want to add a couple little footnotes on this before
we transaction to the next brief presentation. So the first thing is just for
your entertainment value, what Skip didn’t show you that I thought he
might is a spreadsheet that Matt Berent prepared that had, across the top,
the various survey firms we love: Pew, Gallup, CBS, New York Times,
and so on. In the rows were problems like unemployment and the
environment and so on.

And then little x’s in cells to show in the coding scheme used, which
organizations had each of those problems. And what Skip told you is that
we use the rule of two because if we found any category where at least
two organizations did it, we used it. The reason we used the rule of two instead of the rule of three or four or five or six, is that if we had used the rule of three or more, we would have none.

The disintegration of the coding of such a simple question, “What’s the most important problem facing the country,” is really mind-boggling. And, clearly, there is no logic to the way it’s being done. And so the federal budget might seem radical and you couldn’t really scrutinize it because Skip showed it to you in such a tiny little print there. But that may be at least a rationalizable approach.