The Challenges of Transparency in Collecting, Coding,
and Analyzing Open-ended Survey Data
Arthur Lupia, American National Election Studies Principal Investigator

I want to say a little bit about what we want to accomplish with respect to the survey research field in general and to the American National Election Studies, in particular. We think that what is at stake in the argument that we are making is the credibility of both endeavors.

For example, this year you were exposed to a lot of poll numbers. You probably had a sense that some of those poll numbers were more reliable than others. One question you could ask is, "Which poll should I believe?" Or you could ask whether any of the polls were reliable. We contend that one way to distinguish between polls is to consider their credibility, where their credibility depends on the procedures that they use to produce the data.

Some pollsters' procedures have properties you might find appealing if you are trying to use their numbers to characterize the views or attitudes of a broad population. Others pollsters use procedures that, if they were brought to light, you would likely say, how on earth did you come up with that procedure? Conversations about credibility and procedure are the ones we at the ANES and funders at the National Science Foundation want to have. Such concerns motivate this conference. They supply its basic theme.

In election polls you see a lot of things like this [pointing to screen] – a survey will say 67% of Americans have an opinion. A lot of reports also state the margin of error. Typically, the margin of error is offered in small print at the bottom of the page or screen. The margin of error is usually supposed to give comfort to the audience. If the margin of error is reported as plus or minus four percent, you are supposed to feel more comfortable with the report. You are supposed to feel that the true percentage may not be exactly 67% but it is probably close.

The question I want to ask you right now is – why do you believe this margin of error claim? That claim is based on a set of assumptions about the relationship between how the data produced and the population that it is meant to characterize. These assumptions that are not simply handed down from high or true by default. They are assumptions whose validity are a product of procedural decisions made by survey producers and living, breathing human beings who, in various ways, implement those procedures to produce the surveys.

Some of the key assumptions are purely statistical. A margin of error claim is based in part of the law of large numbers, but the validity of the error claim depends on more than this law. The validity depends on assumptions made about how the data was collected, including who was interviewed and how they were interviewed. The validity of an error claim, then, depends on more than statistics. Hence, the initial assumptions about important aspects of the data collection are incorrect, then the margin of error claim may be invalid. Its truth value may be less than 1.

Many assumptions relevant to the validity of poll-based claims are based not on statistics but on survey design or production decisions. One of the pushes that we want to make in our call for greater clarity and rigor in the domain of open-ended coding [pointing to screen] is to bring important questions about the validity of survey data to light. These questions are important to raise not just not for people like ourselves [gesturing to the ANES PIs] who produce surveys, but for the tens of thousands of people who use survey and poll data draw inferences.

When many academics, survey analysts, reporters, or other observers, base their reputation on a claim that 67% of the American public have a particular opinion, or when they base their reputation on +/- 4% claim, they are often making a claim that is based on a set of assumptions whose truths value they may not be fully aware of. That is the general problem. Every survey-based claim about elections is based on assumptions like these -- every single claim. Audiences can get into a situation

where they passively accept the numbers as true, but the truth values of all of the claims depend on the veracity of a set of assumptions.

This point brings me to the basic methodological critique that motivates this conference. The critique pertains to limited introspection amongst the scientific community about these underlying assumptions. This kind of claim is not unique to us. There is a whole field of survey research devoted to this matter considered generally. But the critique is our basic starting point.

At this conference we are looking at the effects of limited introspection with respect to open-ended coding. That is where we are coming from. Our basic idea about a solution is that the way to deal with this problem is to increase transparency. Why would increasing transparency work? It is not just about me publicizing my procedures; it is about publicizing my procedures so that you can evaluate the credibility of the claims that I make and hold me accountable for decisions that I have made. Or it is about my publicizing the procedures that ANES adopted so that when Jim [a reference to conference attendee James L. Gibson] uses our data you can call him on his underlying assumptions when he makes a claim. We think the credibility of survey research can be moved forward if those questions are in play and if people think about them. We think that the credibility and value of survey research will be improved if people have a better understanding of the relationship between the numbers they are using and the phenomena they are trying to describe.

Again, in the survey research field you can kind of split – you have a division of labor – those who produce surveys and those who analyze surveys. On the analysis side, people have had this kind of conversation about transparency for several decades now. A lot of people notice the statistical methodology of analyzing survey data, they think hard about the proper use of statistics, and they develop new and different kinds of statistics to work with survey data. If you take a step back and say what the methodological push about – it is about rigorous inference to be sure, but the more important point is that it is about transparency. The move to improve survey analytic methods have demonstrated the credibility benefits of being clear about the statistical assumptions we are making when using survey data to draw inferences.

In these fields, the push is going from "passive acceptance of data or analysis as true" towards something like a lab book approach, where you are expected to be able to document and defend theoretically the analytic procedures that you use. The push is towards a tradition of scholarship that begins with transparent and clearly stated first principles – what are my assumptions and theoretical framework – all the way to – here is the way I took the number three at this point in this data matrix to mean something about this individual's response to a specific survey question. It is a push to transparency. Now, sometimes statistics can get a little crazy and it may seem that the methods push is getting away from transparency. But at the end of the day the push has been towards transparency on the analytic side in the field of survey research.

We think that all of us can do better. An argument that we would make that on the production side of surveys, this type of debate [about procedural transparency] is not as salient or as present. But for the credibility of the survey research field and for the credibility of the scholars making the claim, understanding the relationship between data collection procedures and inference it is every bit as important understanding the relationship between statistical procedures and inference.

Just because you don’t know the path that converted an NSF dollar or an NIH dollar to the data point that you are using in your analysis doesn’t mean that that path is irrelevant to your analysis. It doesn’t mean that the steps on that path taken are irrelevant to the truth value of the claim you are attempting to make. So you might ask, "Can analysts draw credible inferences from survey data?" We are survey researchers. We love surveys. We have devoted a lot of our careers to it. Even though the question seems to threaten what we do, we [gesturing to the ANES PIs] think that question is in play. It is in play, particularly when you are producing the data yourselves, you learn that properties of the data are products of strings of human decisions.

Thousands of people are involved in the production of one of these data sets. We are all making decisions. A lot of these decisions are affecting the truth value of what each of the elements of the matrix in the data means. Right now many of the elements of the production path are not public, which is why in this conference – one of the things we are trying to emphasize – is procedural transparency. Documenting what is going on.

[A new slide appears.] What is the best way to document open-ended coding? That may seem like the lesser of the two problems – we are also going to talk about the content of the coding itself – but in the debate about survey research we think these are co-equal problems. We have noted that in our own studies we can't explain the codes. We don't know where they came from on the prior data sets. We don't know because there is no prior record of how the decision were made. In part that is because there was no norm in the field of making that documentation.

The directions we want to push now are on documentation but also with the relationship to the theoretical framework. Consider the answer that a respondent gives in response to an open-ended question about a political event or person There is, in principle, an infinite number of ways to code that response. There are multiple theories of language or cognition that you can apply. From the survey perspective, operationally, as Jon was saying, the field wants us to clean up these open-ended responses and produce a small number of discrete numerical] codes. From a labor standpoint within a survey organization, we can do a good job of providing a small number of codes that we can defend and document very rigorously or we can take a less responsible approach and not care as much about documentation and try to produce a lot of codes. I'm not saying this is a one-to-one trade off, but the tradeoff between the number of codes we produce and their quality is something that we are forced to consider. So we have to ask ourselves, "What are the relevant theoretical frameworks to which the data will likely be applied?" "From what types of first principles should we draw if we intend to create new codes? Questions such as these are of a kind that we believe a lot of folks in this room have paid a lot of attention to. We are interested in learning more about these choices from you. As Jon mentioned, at least from the ANES perspective, we are ready to rebuild from the ground up. How can we most effectively code our open-ended responses? We want to implement improved practices in the 2008 study we are doing right now that is still in the field – when we release the data next year. We want the 2008 studies to have a new and improved coding basis and, then, in the future – we want to go back and improve the codes for the older [ANES] surveys so people can make better use of them.

Of course, people might ask, "Why not just recode everything now? There are some challenges. One is labor. It will take human beings to redo the coding. This is very labor intensive. Our studies have been done with 2,000, 3,000 or 4,000 respondents. Respondents are giving answers to, in some cases, 20 open-ended questions with answers all over the place. It can be very labor intensive. How far down that path do we want to go? What are the most valuable types of coding endeavors – what are the most valuable ones we can engage in?

Another thing is when you go down the road to greater transparency – as we have with the ANES in many ways over our term as PIs – we have opened up a lot of attributes of the production process to the entire public. We want people to know about, and to question, what we are doing. In questioning us, people will have a better understanding of what we are producing.

Another advantage of opening ourselves up to new ideas about coding while also being more transparent about the project's past, is that when you go transparent and have a continuing operation like ours, you are going to find mistakes. My own view is that I would rather be mistaken than silent. At least if I am mistaken then someone can engage me about this.

For projects like the ANES, that have a longer life, there is a risk in such transparency. You can worry about feeding the idea that – ANES screwed this up and then they screwed this up – why should we fund them? That is an issue that people take into account in trying to figure out how far down this [transparency] road to go. But we think that legitimacy comes from transparency, so to us the risk is worth it.

Then there is the privacy concern. A lot of people ask us why we don't release the open-ended codes or why we don't make it easier to get. Again this is a live question. The privacy question is a big one for us. We are funded by the National Science Foundation which itself is funded by Congress. The first day that our study makes a headline in the New York Times with a story about the violation of someone's privacy, that is a really bad day for us. These are the types of things that are really bad for us. These are the barriers we face in releasing more of the open-ended material to the public. But we think that with respect to the opportunities inherent in improving our coding, making it more transparent, these costs as a general matter are relatively small, the risks are relatively small. We are ready to rethink this thing.

This is the basic conclusion that we have – this conclusion has nothing to do with open-ended coding itself, it is representative of our entire approach – but it is because of this approach that we come to you. Basically we think that what is at stake here, through a range of decisions, is the credibility of survey based research. You can run surveys in a lot of different ways now. Some of them are less valid than others. Some are less reliable than others. Some of them are simply fraudulent. But the problem is, if we aren't having a broad discussion about what distinguishes reliable from unreliable surveys, then there is a danger that important constituencies will think that all polls and surveys are garbage.

You may think that some surveys are garbage and that is fine but what I would ask you to think about that in that sense – if we don't run surveys – what is the alternative? Right now, for a lot of different fields, the alternatives – whatever you think of surveys – the alternatives are a lot worse. We think transparency and rigor in the development of surveys is the way to go. Those are the basic foundations of what we want to accomplish.

What we can do now is to take questions. We would be happy to do that. If you don't want to, then we coan just go to the next speaker because we are on a tight schedule. With respect to speakers – we are going to keep you tight on time. We haven't built in extra time because we wanted to make sure to get everyone in. On the 45 minute sessions our plan is for the speakers to speak for about 30 minutes and 15 minute of Q&A. So apologies in advance, as we will really shut you down after 30 minutes. On the 30 minute talks this morning we will be a little more liberal on how much time we take in setting up the problem. Are there questions before we proceed?