Open-ended Questions in the GSS

Tom W. Smith

As some of you may know I'm the big proponent of the total survey error perspective. One of the nicest things about this conference is that when we usually use that perspective we think of sample, non-response, question wording, and mode. And then way back in the corner it shows the flow of survey error from coding. We are finally getting the spotlight on it here and I think that is wonderful. As we have already heard, and I am going to add to the choir, yes, there is measurement error in coding.. The problem starts with collection of the data. Good open-ended coding starts with good open-ended response capturing. That starts with good general training of interviewers about the task at hand. And very specific item level, question level coding.

So for example, you have already heard about industry and occupational coding, I'm going to be talking about that some. On the General Social Survey interviewers have 2 pages of instructions about how they should ask the 4 questions that allow the coding of occupational and industry. Those 2 pages of instruction are 10 times as much instruction as any other single question gets on our survey. And even though it is a magnitude of order greater it is at least that much more needed because of the complexity of it. So that is the first thing – general training, very good instructions – you can still end up screwing up and doing poor open-ended coding by messing up the back end but if you don't get the detailed accurate information at the front end, it doesn't make any difference what you do on the back end.

It is nice now that , even in the field in personal interviewing using laptop computers, we now have the increased capacity to get verbatims. Getting the verbatims is very important because of the example in which you said with the actual audio is being decided to get racial patterns of speech. They are also very important in doing something called discourse analysis where you need a very fine grain of understanding the actual language. They are also very important because the interviewers – one of the things for example when you ask interviewers and you immediately see two different styles – you ask an open-ended question and one interviewer wrote down "he" said that patriotism is important. That clearly is not what the person just said to you. The person may have just said "I" think patriotism important. He clearly didn't say that "he" said. Others will record closer to the natural language but unless you are recording it, none of them are going to get the natural language.

I also want to talk about – open-ended questions are not all of one type. We are mostly talking about a subset of open-ended questions. I wanted to indicate the umbrella is much broader. The first is a group of questions which I call technically only open-ended questions. They are questions like – IN what year were you born? How many children have you had? The coding scheme is pretty simple for both of those. We are not really worry about how you code those but they are open-ended questions. Then we can ramp it up a little bit.From the technically open-ended questions we can go to the minimally open-ended question – what language do you speak at home? Clearly that is more complex than what year you were born or number of kids you have ever had. You have to figure out what they gave you as an answer if it really is a language.

One of the issues when we asked this question was if we counted American Sign Language an appropriate response. We decided yes, that is one of our language codes. Another one is ethnicity; "From what country or part of the world did your ancestors come?" Again, it gives a lot of things. The Census for example, when they ask their ethnicity question they will not accept the response Jewish. It is a religion. They are not allowed to collect religious information so they do not code that even though they said Jewish or Hindu. Those are all religions and unacceptable codes so you have to deal with this even in these minimally open-ended questions.

Then you get to the full, complex open-ended question. Pretty much like NLSY, we only have one open-ended question that is part of our core and is asked basically throughout the entire history and that is the occupational and industry question. But we do have a much larger battery of open-ended questions that are asked on an occasional basis. I will talk about those in a minute. Let me first talk a

little more in-depth about the occupational and industry coding question. It is very challenging. It is naturally challenging because – think of all the complexity. Think of all the different jobs and occupations and employers that there are. It is a naturally complex topic.

On the other hand, the Census produces a thick book for coding this. It is not quite the coding scheme that you saw earlier, it is a look up book, now of course computerized. You are given terms – if you say plumber – you look it up – you say Joe the Plumber, who wasn't a plumber of course but that is another story. You get a code for plumber. If you are lucky the key terms are given, matched and you get a clear coding scheme. But of course often times it is not so simple. There are and no matter how often you update your manuals, there will always be new occupations that are just emerging for which there is no code. There might be something that you can put it under but clearly there is a new occupation in which there will in the future be a new code for but there isn't now. You have to deal with that. You have to deal with people with multiple jobs.

There needs to be a rule to handle this. The rule might be to code their main job or it might be code multiple jobs. Usually you only code one and you code their main job. But there is no explicit question in the questionnaire, at least in our version, saying, what is your main job, something that explicitly pulls out that information. What happens is that the interviewer records and/or writes down the information about their jobs. Then a coder tries to figure out what their main job is. Two coders may come to different conclusions about which was the main job and therefore which should get coded.

What I think is even a bigger problem is that many people have jobs that cover many different occupations. Now, if I ask someone what they did at work and they say they file reports, answer the phone, and make appointments for their boss – we might think – a secretary. Those different tasks could eventually fall under one well defined occupation. This is one that was an actual case we had – what does a person do? I sell clothing and I design clothing. There is an occupation code for selling clothing and one for designing clothing. They are both clear codes. This person worked in a small apparel company and did both of those major functions and as far as we could tell, equally. It wasn't at all clear as to how we could code that.

Coding occupation can be improved on the front end if you get interviewers to get the fullest and most complete descriptions but by the examples I just gave that often will not solve the problem. It also will not solve the problem when you have a poor respondent. Or a respondent that is not very articulate or clear who can only talk in jargon about their occupation. "I'm a third grade…" Tt means nothing to anyone else outside the organization. It is a perfectly accurate description for someone within the organization. But neither the interviewer nor the coder is going to have a clue what that means. Respondents, either because of being too specific or too institution specific or because of just being inarticulate, we get responses back – I'm sure they are included in the 40-50 uncoded ones – what do you do at work? "We make things." A good interviewer would probe but sometimes the probe will take you – what kinds of things do you make – "lots of different things." Or what do you make – "stuff". We get people, even after probing; give us nothing more than that.

Now let me talk about some of the other examples that we have. We do ask a lot of open-ended questions. But, unlike occupation and industry, they are typically only asked on a single survey. I will give you a few examples. For example we asked a series of questions about mental health experiences, seeking treatment, asking people to explain the troubles they have had and so on. A second example is, we asked a question on the ERA. We asked first of all if they favored or opposed the equal rights amendment and then we asked as a follow up why they favored or opposed the ERA. Clearly a classic, broad scoped open-ended question. We coded it into a little less than 100 different codes. Another example is that we asked people if they have ever changed their religion; what religion they were raised in; did you ever change your religion. As a follow up we asked why did you become an X – we mentioned whatever religion they changed to. if they started out as Catholic and they are now Baptist we asked why they became a Baptist.

One of the most challenging ones, one that was specifically sponsored by the National Science Foundation, was essentially a question saying – what does the word science mean to you? In other

words, define science. Not only is that a conceptually difficult thing but for the average person you can imagine that is an extremely difficult challenging thing for them to do. Spiritual transformation, we asked people if they had undergone a spiritual transformation and asked them what led to the transformation. And then a second question, what happened at the point of the transformation? What went on when you were transformed? And the third open-ended question was essentially what were the consequences of the transformation, how has your life changed? We had open-ended codes for all that. You can imagine it was very complex.

Another example, there is a well established series of questions that makes a subset in an IQ scale in which you are given different objects and you are asked how they are alike. You might have a chair and a table – how are they alike? These are field coded into a scheme – 2 points – I will read some of the coding instructions. "You get 2 points for any general classification which is primarily pertinent to both members of the pair. For example, an eye and an ear are both sense organs. A poem and a statue are both works of art." You get 2 points. Imagine how you would extend those examples to all the different kinds of things people are likely to say. You get one point for any specific property or function which is common to both and constitutes a relative similarity.

For example, an egg and a seed both have shells. A poem and a statue both stir feelings. You see the distinction with the poem and statue, stirring feelings. Yes, it is something that they both do but it is not the essential nature of what they are. You get one point instead of two points there. You get zero points for specific properties of each member of the pair, generalizations which are incorrect or not pertinent, differences between the members of the pair or clearly wrong answers. I like that last one, if all the other stuff isn't – possibly wrong answers or uncertain whether it is wrong or not. But it all gets together. Here we are differentiating different degrees of correctness – like you suggested with the Supreme Court example.

Perhaps equal to the challenge of the science question is the following question – "The next questions are about freedom in America today. Freedom means many different things to people. When you think about freedom what comes to mind? Can you tell me in a couple of sentences what does freedom mean to you?" Challenging question to answer and at least as challenging to code.

Now I want to talk a little bit about all of it. That will be my last example. I realize I am the final thing between you and lunch. I don't want to hold you up too much here. How do we go about coding these questions? I shifted out at this point because as you are seeing from this example how it was coded. Of course, most of the coding, and the occupation/industry coding as was already referenced to, is done by having trained coders who are familiar with the thing, under supervisors who are trained coding supervisors. So you have, just like those people down in the basement of the Census Bureau, professional coders. That is done with occupation and industry at GSS. Sometimes, but not as a standard thing, dual coding is done. Dual coding is not a standard procedure there.

A second way that coding is done which we did for example with the question about the spiritual transformation – what led to it, what happened during it, what were the consequences of it – graduate students under the PI's supervision within the project itself coded that. So a number of our questions, particularly when the PI have been the designers of the particular module, as I was for the spiritual transformation questions, we take responsibility for and directly supervise the coding. The advantages are that you have closer, more intimate control over it. But you don't have professional coders who have some skills that you would like to draw on. You have graduate students for whom this may have been the only coding they have ever done.

The third thing that we do, and this brings us back to the freedom question, sometimes we have the coding done by external people who are the PIs of modules. Some of the mental health questions that I alluded to earlier were coded by staff with substantive expertise in studying mental health at Indiana University. In the case of the freedom question, these questions were coded at Harvard University by people working for Orlando Patterson. I use this example because if we had coded these questions, we would have coded them in a completely different way because for his purpose this question about what freedom means – was coded in 8 categories and that is it; single digit

variables – just 8 categories. If I had sat down there I would have at least 3 digits, a much more complex code and nuances of freedom and specific examples and whatnot. But from the principle investigator's point of view – these were the 8 categories that he needed.

We have, at least in recent years, the last 10 years, we have mostly typed verbatim text, the occupation and industry questions and the spiritual transformation questions – we have them available under special permission. People can access this. But here is the interesting situation. Under our sensitive data procedures we get about 50 requests a year. Of those 50 requests 48 are for geographic detail. For the basic occupation and industry question, I could be wrong, I cannot remember a single request for the occupation and industry – even though we have been collecting it since 1972 and have it in a readily accessible form for at least the last 10 years – there has never been a request.

That is our experience in terms of occupational coding and in terms of open-ended coding in general. I will end by saying that I think it is wonderful that this conference has been organized to look at the neglected but very important areas of a survey.