Methods for Assessing the Reliability of Coding

Matthew Lombard

Thank you. I'm glad to be here. I'm sorry I had to miss yesterday's presentation. I had to get up at 4:00am this morning to get here so if I conk out half way through and start snoring you will know why. I come from a communication background. I just want to tell you a little bit about why I got interested in this topic in the first place and then I will talk about these other topics quickly [pointing to slide]: The study that we did of content analyses to see how they reported and dealt with the issue of reliability. Then some challenges that we all face in improving reliability in content analyses especially in communication. And then some quick discussion about relevance of all this in open-ended coding and a proposal to think of reliability in even a broader way than is usually done.

I got interested in this because we started a too ambitious content analysis of the structural features of television in the mid-1990's; that means how long your credits last, how much clutter is there in an image, shot and camera angle, things like that. All of which is completely manifest content that you would think would be very easy to code. We had variables at all levels of measurement, especially ratio levels because we had all these duration variables – how many seconds does something last or how many cuts in a 10-second interval? We had 7 coders code about 450 programs and 4 coders code about 1000 randomly selected timepoints in 2 hour recordings. It was a very complicated project made even more complicated by the fact that we were using old VHS machines and we didn't have all this material in digital form.

We wanted to do reliability right, being well trained researchers – we thought it would be easy, we would just look in the book to find out how to do it. What we found was there were lots of indices, very few of them were flexible. With the variety of variables that we had, there were very few guidelines. There was Krippendorff's 1980 book but this was before the new edition that came out in the last couple of years. Krippendorff's Alpha seemed like the best measure to use because it gives you a nice number of measurements for each variable at each level. It gives you all the information so you can pick the right value. It is very useful. But it is just almost impossible, for the ratio level and the interval level computations, to do by hand; it is just way too complicated. There was not and really there is not, it is very difficult to find, software that will reliably and accurately calculate it for you. That was a huge problem. We also had the problem that some variables don't vary much, for instance "color vs. black," where 99% of the time what people were coding was color. Everyone said color, great. Everyone agreed most of the time so if we use percent agreement we are fine but if you use Krippendorff's Alpha and some other available formulas for inter-coder reliability you get a very low number. Because you didn't have enough changes for the coders to spot "black and white" accurately and code it that way. There are just a couple of "black and white" clips in there. One coder got it and another didn't, so according to the Krippendorff formula you failed to demonstrate reliability. There is an argument about that but on the other hand we were very comfortable in our conclusion that most of the television that was coded was color. We didn't think we needed Krippendorff's reliability to be high for that. As part of this project and sort of arising out of it we did a literature review on this whole topic of inter-coder reliability and how to do it and what people had investigated in the past concerning different indices and how well they were reported in the literature. That was really the research project that came out of our initial study – how well the researchers, specifically in this case communication but we looked for anything about reliability in the literature in other fields as well - how well did they deal with reliability. We wanted to do it right for our project. We really worked hard but it was quite the challenge.

What we found in the literature review was that almost everyone says that inter-coder reliability is important. We all tend to agree, everyone says it is important. But previous studies across different fields, especially in communication but in other fields as well – it is a surprisingly large percentage of studies that don't even mention reliability. Content analysis is a key word but they don't mention reliability. So to make matters more complicated Popper came up with a list of 39 different measures of inter-coder reliability, all for nominal variables, so that means Krippendorff wouldn't be included completely as that kind of index. There are other indices for other levels of measurements and tons of things out there. The only thing that the experts really agree on, at least at the time we did this review of

the literature, was you can't use percent agreement as an accurate measurement of inter-coder reliability. That happens all the time. That is the most used measure, not just in communication, but generally! It is flawed because it doesn't account for the possibility that people will agree just randomly, and there are lots of good demonstrations of that. It is a very inflated measure. Some of the articles we looked at actually took results from previous studies and said – if you didn't use percent agreement you'd lose all of these variables and it wouldn't be reliable anymore. They also said don't use Cronbach's alpha, r, or any kind of associational index and do use indices that consider chance agreement. And this literature review reinforced our perception that there isn't easily accessible software. The only thing that we could find, and we really scoured the environment, was beta versions, including Krippendorff Alpha's beta version. All these special software programs required a separate set up – you had to change your data all around and feed it into the program to get it to work. It was the kind of thing that would discourage people who weren't motivated to do reliability right. With the literature being a part of this study we actually read and coded these content analyses in Communication Abstracts which is the index for the communication field. 5 years – 1994 to 1998, 200 articles. We had 2 coders. We used several different coder reliability indices so we could show the differences that they would produce. 80% of the studies were strictly quantitative, 8% were qualitative and 12% were both. They coded them like this [pointing to slides]. The top category was newspapers. Notice only 3% were "respondent reports." 16% were "other." These are all different kinds of stimuli or materials, similar to previous studies we had looked at.

Just under 70% of the articles discussed inter-coder reliability at all. Discussed is a fancy word. Mentioned would count them as discussed here. Of the 69% only 41% reported reliability for specific variables. That is a very big problem. 9% didn't report the number of reliability coders. The mean number of sentences anywhere in the text, including the footnotes, was 4.5. Only 6% included reliability variables in a table. Only 45% cited an article or book related to reliability.

Here are the methods that were most used [pointing to slide]. Notice that I grouped percent agreement here. That includes 16% where they said" inter-coder reliability"; we assumed they meant percent agreement. Often when they use an ambiguous term we assume they used that simplest and most common type of agreement. 36% used other techniques, so it is all over the place.

This is almost always without any explanation of why. It is just – "here is what we did." You can read this – more disappointing, lack of information provided in a lot of these articles. 91% did not report the amount of coder training it took to reach those levels of reliability. We really thought that was a problem.

The conclusion – not surprisingly: inter-coder reliability is inconsistently assessed and reported in communication research reports. Just a couple of examples [pointing to slides]: Here is the entire discussion of reliability – remember this is one of the 69% where they did deal with reliability – [reading] "All 168 articles were trained by four trained researchers. The Holsti coefficient of inter-coder reliability was .93." They are reporting across a whole bunch of variables. Overall the number was .93 but what was it for the individual variables, and there is lots more to say about reliability! Here is one where they say more, it is longer but arguably not better: [reading] "Correlations were computed between the second author's ratings and the two undergraduates' rating for all the variables." Then there are the mean correlations stuck in there. 'If you want more information, let us know.' That is not particularly helpful.

We took up the challenge a little here and thought about how we could improve the reliability in our field. We argue it's especially important in communication because we feel people that study mass media use all kinds of methods but if there is one that makes sense especially for what they study it's content analysis. Ironically our study was of the *form* of a medium. But it is the same argument. Our thought was that we should publicize and really argue for the importance of inter-coder reliability especially in communication. We should also not make perfect the enemy of the good. If we can get people to just do some sort of inter-coder reliability, preferably not just percent agreement, we thought that would be quite an accomplishment. Krippendorff himself gave us some trouble after the article came out. He was finding things in the third decimal place. He is totally right. You should have complete

accuracy, but I'm not nearly concerned so much with that but just getting people to do inter-coder reliability in the first place and explain their choices.

We proposed a set of guidelines and described some available tools in this HCR article (Human Communication Research), a journal in our field. Also in this there is a link to an online supplement that we try to keep up with. It's a shortened version of the article and a list of software and a lot of details on how to use the different software, how to set the data up to get results. Of course, this talk is another extension of that article.

Just quickly, here is an abridged version of the guidelines just to give you an idea of the things in our recommendations. Since we did this, Klaus came out with some modified versions of this. There is the new Nunendorff book, and there are some other sources. This is certainly not private ownership and other views are generally comparable. So [pointing to slide], obviously do reliability and second of all report a bunch of stuff including: the size of and the method used to create the reliability sample; the relationship of the sample to the full sample (was it the full set or a subset or a separate sample); how many reliability coders were there; did they include the researcher (some people have written that it is a bad idea to include the researcher – you want to show that people outside the experts can code, that any smart person can be trained to do this); the amount of coding conducted by each reliability and non-reliability coder; obviously or critically important: which index or indices were selected to calculate the reliability and justification of this/these – this is admittedly somewhat hard to argue because there is disagreement about the strengths and weaknesses about each one; what was the level of reliability for each variable, for each index, not just an overview; how much training (in hours) did it take to get the coders to be able to code reliably at the levels reported; how are disagreements – invariably there are disagreements – how were the disagreements in the coding resolved in the report of the actual data; and where can the interested reader get more information.

We (Cheryl Bracken and Jennifer Snyder-Duch – now alums, they were Ph.D. students at the time, they are professors now - we still have a fighting interest in this. We're still trying to advocate, push, cajole, encourage and whatever else to get more agreement among the experts about these guidelines and especially which indices to use when. One relevant example here is Cohen's Kappa. Some people say that should be the measure of record; that is the one that everyone should use. Krippendorff has argued fairly persuasively that there are some serious questions about whether it is appropriate. Everyone else is left stuck, because as someone mentioned earlier, SPSS has this one that you can use. If we could get consensus – I am not a mathematician – when is it appropriate and when isn't it appropriate? What index should different people use? I'm talking about it from the perspective of someone who wants to do a content analysis and they are not particularly interested in all the details of this thing, they just want to know what to do. Really not that there is much that one can do but we're trying to help along the process of developing software – especially not separate stand alone software where you have to set everything up separately and transform your data just to calculate reliability – but things like SPSS; there are all kinds of macros for SPSS and SAS and other programs. Just recently a Ph.D. student came up with an online way of calculating several of the measures but you have to submit your data to the website and it calculates it for you.

In terms of open-ended coding, this is only a slight tangent but I have heard the phrase "qualitative content analysis." I know based on the way I was brought up, content analysis is supposedly quantitative, from the social science tradition, but obviously there are all kinds of cases where - I like the word quantitize – where we quantitize the quality of the data but you do a quantitative analysis. There are all these sort of hybrids and in between types of analysis. A very simplistic and impractical bias would be – I have heard this from some qualitative researchers: if data is qualitative and you have open-ended responses and not just a sentence that needs to be turned into a quantitative, coded variable – if you have qualitative data you have rich data and why do you want to turn it into numbers? And on the other hand, if you've got something that can be turned into numbers if it is simple, why not find out enough information so you can form a good coding scheme and then use that coding scheme as the question? If only the world was that simple. That is a bias that I have encountered and think about. If it is really a qualitative variable don't lose the richness by coding it to a number and turning it into a quantitative variable. If it is quantitative, make it quantitative using closed-ended questions.

Obviously a lot of the time data are open-ended and verbatim and essay. Clearly if humans are going to code that data, reliability is critical according to the reasons we said (I didn't give you all the quotes that people did in the literature – 'without inter-coder reliability content analysis is useless' – all these dramatic statements that you find in the literature).

Reliability with open-ended data is an especially interesting challenge. Someone called me about it a year ago from a very large world wide company where they do verbatim analysis for marketing and other applications. She was about to give a presentation about inter-coder reliability. She contacted me for that reason. She said what they did – the first person, one of their employees, goes through the data and basically comes up with a coding scheme and then everyone else just uses that same scheme. She was asking me about it – how do you make sure they are using the scheme consistently? If you have the first person form the scheme then there is a level of problem even there. You are creating the scheme with the input of just one person and this is not a trained expert, this is just someone that works there. Clearly if you are basing the data on a scheme that already exists, there are issues regarding reliability that directly relate to the validity of the scheme in the first place. Then getting multiple people to code consistently according to the scheme is a challenge.

The next issue is that if you have open-ended data and it is being coded automatically – it is fascinating some of the discussion here this afternoon about all the different degrees of, on the one side people take these open-ended responses and code themselves or give it over to the computer and stand back and let it do its thing. But in any case reliability is still an issue wherever you are on that continuum. The computer is reliable. I think you can certainly count on the algorithms that are in the computer working the way they are supposed to. I'm so frustrated with computers, but that is one of the few benefits of computers – it does what it is supposed to do (as long as it doesn't crash). But obviously there is going to be the issue of whether the scheme that it is applying, the logic that you are using to code these open-ended things, is *that* consistent? Here I am arguing for this broader definition of reliability. Not just a person and a person, or a person and a few coders, or a person and a computer, as in the earlier discussion that I thought was really fascinating where we train the computer to be a person basically, like my junk mail filter: I put mail in my junk folder so my junk mail artificial intelligence system will learn how to do that.

Regardless of what entity is doing this, you still have this issue of – is it consistent with other people and other projects where they are doing this work too and using – in election studies it is also over time, where you ask year to year or each election cycle you ask these questions – does the same scheme get used in different papers that come out of a larger data set or across researchers where they ask similar questions or the same question – for common open-ended items, items that are going to be asked by the same people over time or by people in a field. Even if you have a computer doing it we want to increase the consistency of coding schemes themselves so there's reliability of the coding schemes across studies, across researcher so we can compare results. The work I'm doing these days has to do with a concept called Telepresence, the sense of 'being there' in a mediated environment, virtual reality or high definition television. The frustration that I often have is that people - I'm sure you will be able to relate to this in whatever background you have - people use the term to mean different things and they use all kinds of variations of it to mean different things and they measure it a billion different ways. I may be exaggerating that but not by much. And these lead to all kinds of ways of measuring, so when you see a study that says they found a relationship between presence and enjoyment or presence and whatever the dependent variable is – you can't trust that that is the case because what kind of presence are they talking about and how did they measure it? It could be everything from, "well I asked them how much presence did they feel" – they probably didn't understand what I was asking – to a 50 item matrix of questions filled with indices and so there are all kinds of variations across schemes because they are essentially different category schemes. It means that I just can't trust their conclusions. You can't make a nice little map of what you know in that field.

How do we advocate a better consistency across – where it is possible, because it isn't always going to be possible – across studies and across researchers? At events like this – this is a wonderful effort and obviously there is a practical focus to a lot of what we are talking about. But this larger issue comes through: How do you keep a collaborative effort – this is purposely multidisciplinary but even in another

way of doing this another step is having people *within* fields get together and talk about their particular questions that they ask. I have an interest in some of the politics because I have an interest in political communication but that is not my main area. Other people here, that is their main area and there are people in all kinds of areas where this could be done – and the technologies that we all use for other things really can be useful in this context. Wikis, listservs, shared databases – even just being able to put those guidelines on the web has gotten a lot more attention to that article that we did than the hard copy article itself. As I know has been discussed because Joe sent me the slides from one of the introductory talks yesterday, it is this issue of using theoretical rationales for developing these coding schemes. I think that is a very good idea, a strong thing to do that should be advocated. And also the slides mention and I strongly endorse requiring publication, either with the database materials or papers that come from the database that is being analyzed, of enough of the details about the schemes or how they were done so they can be replicated. If you can't have consistency in use across different databases, people don't know how the database coding was done. Really all of this comes down, in some sense, to community building in the sense of shared obligations. We all have a motivation to do our study our own way and get it done as quickly as possible. If we can see that working together on these kinds of things will improve the quality of work that we produce collectively, I think that is a strong argument. All of this is quite easier said than done, but worth doing. If I have stepped on anyone's assumptions about what is appropriate or not appropriate to do with inter-coder reliability I certainly don't mean to offend. I just learned a lesson in approaching the whole topic over a decade ago and I have a bug up my butt about this. It is important!