Week 6 FAQ

Are our concepts based solely on our individual cognitions?

No.

Or are they also influenced by our environment, upbringing, opportunities etc?

Yes.

It is possible that the use of vignettes produce bias? For instance, those who are familiar with the hypothetical situations might respond very differently from those who are not.

They might, they might not. All sorts of things *might* happen. If you suspected that familiarity was an issue then you'd want to try and measure that and it would become a predictor of the thresholds.

Given that data are collected systematically by public and private entities in much larger quantities than before, what does the future hold for classic sociological surveys? Granted, newer forms of data do not solve issues of conceptualization, but might very well provide less measurement error and allow insight into previously untouched concepts. Examples for this can range from registry data to Google search histories.

The future is looking grim for surveys mainly because response rates are falling through the floor everywhere. Administrative data is used a lot in some countries for example all the Nordic countries have very good population register data. In some countries ie Germany public resistance to use of that sort of data is potentially huge. In the UK official resistance is still an issue, though it is getting somewhat easier. There are of course lots of other "big-data" sources. Not my field, talk to the OII people. My impression is there is lots of potential, but also lots of drawbacks – massive sample selection issues and unclear what the population is.

Regarding the issue of reliability, what would be the best way (if any) to measure reliability for qualitative research - test/retest or inter-item reliability?

Facetious answer: why would you ask me? Why not as someone who concerns themselves with qualitative research? This course isn't about qualitative research. More sensible answer: Unless you spell out what you mean by qualitative research it's very difficult to give an answer to this. As I never tire of saying, qualitative research isn't a thing, it is many things. Some sorts of research that get described as qualitative would fit the test-retest paradigm ie anything where there is say coding of text to do. Standard tool, as you have no doubt been taught is coding by two or more people and calculation of level of agreement between coders.

Could you give us more examples of Criterion-related Validity and Construct Validity?(p.25).

Not really. But if you care to look at the reading I recommend for the lecture you will find some more examples. Also the relevant section in the Hoyle et al. textbook is useful.

Whilst I understand why some researchers need to test theory quantitatively, surely a huge amount of data is lost in reducing what is said to quantitative measures. For example, the difference in respondents answers will be full of dialogical meaning, which can give rich information about what the DV means to them given the different power axis that work within them (i.e. class, ethnicity, gender) etc.

Not entirely sure I know what you mean by "dialogic meaning". Do you mean in Bakhtin's sense? If so you'll have to explain the relevance of the remark to the issue in hand. Also I don't really know what it means to have a power axis working within me. It sounds kind of painful.

I think I get the general point though. It's a version of the world is very complicated argument. The answer is, yes it is & what do you want to do about it? If your goal is to reproduce the world in all its complexity, then good luck and have a nice life. You are going to very busy & you will die without succeeding. All social science is about simplifying and stripping away inessential idiosyncratic detail. This is not just one way of doing things, it is the only way of doing things. Even the "richest" "deepest" qualitative comparison does that. So the only question of real interest concerns what is essential, what can be ditched, and how should we do this in a principled way that doesn't rely on hand-waving or appeals to private knowledge.

There is no completely general answer to this because it depends on your substantive question. The substantive questions that KMST address are very simple. Are average levels of "political efficacy" higher in Mexico or China? If the question has any meaning at all then we have to have some common reference points. You could of course decide that no such common meanings exist but that would seem to be empirically false. We can for instance travel to foreign countries and somewhat understand what is going on.

If we concede that there are some common reference points, then we can point at the ones we pick out & say there, look, that's the kind of thing I mean. Will that exhaust everything we might want to know about political efficacy? No, of course not. But nobody can do everything at the same time.

Concerning the concept of validity, the lecture mentioned that there often aren't ways to test for this. Since validity is so important for serious empirical work how can a researcher be certain they are measuring what they intend to measure and how can they convince others of this?

If your demand is for certainty, then the empirical sciences aren't for you, and choosing another career would be the rational response. Seriously. Assuming you don't mean 'certainty' literally, I did outline three approaches you can take to producing evidence that is relevant to a judgement about the validity of a measure. If you want more, then you are, I'm afraid, looking for the end of the rainbow.

I am a bit puzzled about a statement you made in the online lecture. You stated that measurement involves the specification of a protocol that says how to do it and that this protocol is intersubjectively available. Perhaps I am confused because I have yet to encounter the phrase "intersubjectively available", but I do not know what you mean by this statement. If intersubjectively means "existing between conscious minds; shared by more than one conscious mind" (per the definition), how exactly is a given protocol intersubjectively available?

By existing as an explicit set of codified rules and procedures that can be taught by one person to another or learned from a common source. The point is simply that everyone should be singing from the same hymn sheet and private ways of making measurements that can't be explained to anyone else aren't allowed.

To what extent are vignettes appropriate for the testing of hypothesis and theories?

Well, that's what KMST think they are doing.

It seems that vignettes can contain too much context and leave room for interpretation. Furthermore, it seems much more difficult to formulate vignettes in a consistent matter, so that sentence structures, vocabulary and other stylistic features don't affect the response. Wouldn't it be better if one tries to decompose a concept in as many directly observable indicators as possible?

The points you make are not specific to vignettes, they are present to a greater or lesser degree in all attempts to ask people questions. They actually go to some lengths in constructing vignettes to try and exclude all unnecessary contextual detail. The answer to your last question is, all other things being equal, yes, and nothing in KMST's methodology inhibits that. In fact if you read the article carefully you will find that they actually advocate it.

I have a question about Adcock and Collier's Measurement Validity paper. They refer to the "limitations of inferring validity from a high correlation among indicators." They advise seeking to rule out alternative reasons for the high correlation, but I'm unclear how this exercise could resolve the validity problem. If you are able to explain the high correlation, does that then mean your inferences are valid? I took away from last term the idea that multicollinearity was sort of a death blow to your model.

What they mean is very unclear to me and seems rather confused. Observable responses to items that can be thought of as indicators of a latent construct should be correlated because they are "caused" by the individual's position on the latent construct. There are however other measurement models that use so called "causal indicators". In this scenario the indicators define the construct and they need not be correlated. Multicollinearity is a red herring. If you work with micro-data you are almost certain never to come across it.

Isn't to say that we can adjust for personal biases by noting variation across a scale, just a sneaky way of trying to introduce more subjective data points with which to make comparisons?

No

In other words, the problem of subjectivity is inherent in the interpretation of each vignette and the measured scale that is supposed to be used to adjust the data is similarly subjective. There is no objective element in which to "anchor" the analysis and it seems to me that the process is simply adding more subjective measurements into the mix.

No. What do you think the vignettes are? We choose them. We know what they mean. They are objectively the same for all respondents.

Increasing the number of subjective observations can increase the likelihood that you detect variation, but to "adjust" in any direction is to do so blindly. Am I missing something?

Yes.

In King et al.'s piece, they "recommend asking the self-assessment first, followed by the vignettes randomly ordered" (p.194). This seems to me to be a problematic recommendation because separating the two stages from each other allows for two different participant mindsets. While the answer to the self assessment would be uninfluenced by external information, the answer to every vignette would be influenced by the information given by every other vignette. Would it not then be better to have the participant be exposed to all the vignettes first, then to answer the self assessment and the vignettes so as to further ensure response consistency?

Good question, which I confess I didn't know the answer to. King justifies it here: https://gking.harvard.edu/anchoring-vignettes-faqs

Is it important to develop a standard measurement for a variable across a discipline so that results can be compared and discussed? It seems difficult enough to develop a reliable and valid

measurement for a variable within just one research project, but that could lead to many different conclusions about similar concepts due to the varying measurement strategies.

Unless we want to live in our own solipsistic worlds the answer is probably yes & that is how it pretty much works in some disciplines ie psychology, medicine – symptom scales & classifications. Of course, differences are a matter of degree. There are lots of different ways of measuring some things which are all pretty equivalent so it doesn't matter too much which exact one you use.

As the writers in Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research mention about measuring different types of regions 'where different regions are truly unique and variables take on completely different meanings, then any procedure, including this one, will fail to produce comparable measures'. Are there any alternatives to compare different types of regions?

They don't mention "measuring different types of regions". The point they are making is that if you really believe that different places and cultures are fundamentally non comparable, then the future for you is Area Studies or butterfly collecting. Both are coherent occupations.

What is 'Monte Carlo' evidence/ analysis?

Define a data generating model. Generate some data with it. Fit your statistical models. See how well it estimates the known parameters of your data generation model.

The article discusses the distinction between questions with Differential Item Functioning and those without. Would all questions not have, to some extent, an influence of DIF?

Possibly, but it is likely to be very marked when your response scales uses vague quantifiers

The anchoring vignette method in King's (2004) research seems to generate a lot of measurement error when applied in different contexts. Does this therefore negate the internal validity advantage (through minimising DIF) that King claims this method can provide?

Does it? How do you reach that conclusion? I don't see that myself. "Internal validity" and "validity" in a measurement sense are two rather different things that shouldn't be confused. The point is to increase validity in the second sense.

King, Murray, Heath and Salomon (2004) mention benefits of anchor designs and vignettes in reducing bias and increase efficiency. However, does not the vignette approach also introduce a new source of variation to the survey. How do we know whether this variation deals with incomparability or is just another source of bias?

That's what the Monte-Carlo simulation demonstrates.

How can a researcher decide whether a goat is a good substitute for DVDs for Ethiopians in comparison to New Yorkers? (see Murray, Heath and Salomon 2004:204).

By thinking and drawing on as much local knowledge as possible.

I found the vignette approach in King et al. (2004) interesting and possibly useful, but they did not discuss much about how to achieve vignette equivalence (Pp. 194). They only talked about testing the equivalence (Pp. 199) and not much about the process of achieving equivalence in the first place (e.g. how to "properly" translate a vignette). Hence, my question: is there any systematic way to achieve linguistic equivalence?

KMST's use of the term "vignette equivalence" is actually very specific. What it means is that the vignettes should all tap a single dimension of the construct they purport to mention and that the respondents should understand them as doing so. This would be violated if in the example KMST use some of the vignettes really related to a construct other than political efficacy or, perhaps more likely some of the respondents perceived them to be so. In the latter case not all respondents would be reacting to the same kind of thing.
Linguistic equivalence is really a slightly different issue. It could be important if, as sometimes must be the case, though a semantic equivalent can be easily found, the words used have accreted slightly different connotations in the different languages.

I have serious doubts about researchers who are keen on conducting cross-cultural research projects but do not adequately understand the languages involved. One can maybe hire an interpreter, but I think there are limits on how much you can rely on them.

So do I, which is why I tend to steer clear of comparative research unless I have at least a passing familiarity with the languages involved. The professionals, of course, use back translation as a check on semantic equivalence. You start by formulating the question in the lingua franca – usually English, translate it to the target language and then get someone else to translate it back to English. Then you see

what has been lost and what has been added in. But yes, bilingual competence is needed.

The article about DIF was interesting that vignettes ground people with certain examples and the responses could be adjusted based on assessments. However, at the same time, I thought people could understand the variables in the vignettes in a different manner across context, which can cause another internal validity and lack of comparability.

Which variables? How difficult is it to understand a lack of clean drinking water? How difficult is it to understand feeling that nobody listens & nobody will do anything for you? These seem like pretty universal things to me.

In the case of the test-retest method regarding how to measure the change in short period of time. And there you cited the example of testing mathematical abilities among children over the short period of time by using test-retest method (where observers would use slightly different instruments to detect the changes and to avoid maturation among children). And in many research studies, researchers do consider the fact that, to detect the long term change, a same place has to be surveyed after long intervals of time. So in that case how would one control for - death of respondents, or natural calamities in the area, then rehabilitation/ displacement of people from that area??? which would eventually not facilitate a study like test - retest to detect the underlying changes in a society or region over a long time intervals.

There are so many levels of confusion in this question that it is difficult for me to know where to start. What you appear to be talking about has little to do with evaluating the measurement properties of an instrument – which is what I was talking about. You actually seem to have something in mind like a panel survey. How to deal in that context with selective attrition is an important topic. However it is not particularly relevant to what I am talking about.

What is the Monte Carlo method as mentioned in the King et al. article? How does one perform a Monte Carlo simulation, and how does it compare to other types of simulation methods?

I do a very simple MC simulation in my second lecture. The code is here: http://users.ox.ac.uk/~sfos0015/experiment_random_distrib.do

What other kinds of simulations did you have in mind?

How realistic is the assumption of vignette equivalence (p194) even within the same culture? are our personal experiences/ beliefs similar enough for practical use of this assumption?

Well, a facetious answer is that broadly speaking the members of the same society understand each other well enough for stable expectations to be formed which work most, if not all, of the time. So there must be some substantial overlap in the way we see the world. Then there is the design of the vignettes. They are meant to be the verbal equivalent of pointing at something and saying, look, that's what I mean by X. So an example of KMST's is pornography. It's difficult to define and people might not quite understand it in the same way. But if I get an explicit photograph and (with your permission) show it to you I can say: look, this is what I mean. Vignette equivalence means something very specific in the article – see answer I give elsewhere.

My question for the week is on the perils of interpretation when conducting studies that involve different cultures. The reading by King et al. touches on this in terms of levels of interpretation, however the solution of measuring "response category incomparability" seems limited in the sense that it may not account for misinterpretation of response on the side of the researcher himself, if the context of asking questions to respondents of a different culture to the researcher. It would be interesting to see how this question could be analysed in more detail.

After you have corrected the vague quantifier response scale for DIF what would be the nature of the researcher misinterpretation? I don't really understand what you are asking.

Does validity really presume reliability (slide 19)? In the majority of cases, yes. But what if the measurement itself were to impact the measured characteristic? Then reliability would be a fairly meaningless concept, because it would be limited to that specific time of measurement. Say I was measuring how much a person likes a certain kind of music and exposed them to several songs, asking them to evaluate them on a Likert scale. However, if I was then to repeat the procedure half an hour, they might report higher liking of the music due to exposure effect. How do we deal with this kind of separation of validity from reliable? Because validity is given in both measurements here, but there is no reliability.

As I say in the lecture, for reliability to have any meaning there has to be a ceteris paribus clause. If the reality changes, as it does in your example, then test-retest fails as a technique for evaluating reliability because the ceteris paribus clause is violated. However that does not necessarily imply there is no reliability. You are confusing a specific measure of reliability with the concept of reliability itself. If reality is completely reactive to the measurement process then that pretty much rules out our ability to measure reliability empirically. Would we then conclude that

the measure was unreliable? That would be going beyond what would be warranted. All we could say is that we don't know anything about its reliability and if it were completely unreliable then it would, by definition, be invalid.

I'm still unsure of the concept of DIF as discussed in the readings. What is about and why is it important?

Perhaps read the article again? Otherwise I don't know what to recommend.

From the reading, it looks like using vignette and self-assessment is a plausible way to detect the DIF and adjust the measurement error in surveys. But for research that involve in-depth interview, how can the problem of DIF be solved? Or could that be the case that interview-based research does not suffer from DIF since it can better understand the different cultural context of the participants and thus making sure what we want to measure is the same thing that the participants give to us?

I don't think people who do in depth interviewing think about things in this way, though I don't know. Many appear to feel queasy about even mentioning the word "measurement". So basically I don't know. Why not ask someone who does that sort of thing?

How widely used are methods such as, or similar to, what is described in the article? Are researchers engaging in surveys not just with self-evaluation but similar vignette ranking? Also, as the researchers are still assuming 'vignette equivalence' (p. 194) and thus require participants to understand the vignettes the same way, does this approach not suffer from similar issues that regular surveys do in terms of participants' different interpretations of questions/concepts?

Gary King's web-site has a lot of examples.

Is criterion-related validity the same with criterion validity? In the lecture video you have given an example about construct validity but not for criterion-related validity. Can you give an example now?

I don't know what your first question means and therefore can't give you a sensible answer. The statement in your second question is simply false. Go back and watch that section again. If you didn't spot the example then I'm not sure I can help you further.

What is a latent trait model?

It is a model that, in its simplest form assumes a continuous unobserved latent variable which predicts observable binary responses (the responses don't have to be binary, but let's keep it simple). An example would be a model for maths ability. You give a bunch of people a maths test consisting of 10 questions of varying degrees of difficulty. The answers can be right or wrong. The idea is that the probability that a person gives a right answer is a function of their position on the underlying continuum – which we might call maths ability – and the difficulty of the question. The trick works because we assume that conditional on the person's position on the unobserved continuum the observed responses are independent . If we are prepared to believe that we can then work backwards from the observed responses and infer what their position on the unobserved continuum must be.

My question for this seminar is: what are the most efficient ways of reducing social desirability bias?

Good question. Answer: by using a suitable mode of administration. Face to face & telephone interviews are most likely to be prone. So self-completion supplements are sometimes used on the assumption that people won't be so embarrassed to admit to undesirable views or behaviours. If you are just interested in prevalence then a so called randomized response technique can be used. Imagine you want to know whether the respondent smokes dope. You tell them to flip a coin but not reveal the outcome. You then tell them that you want to know whether they smoke dope and that if the coin came up heads they should answer yes, and if it came up tails they should answer truthfully.

Adcock and Collier (2001) suggest some excellent strategies for measurement validation; however, they implicitly assume the existence of multiple datasets with different indicators that's been explicitly designed to pertain to the same systematic concept (e.g. a particular definition of democracy). The main problem that they're trying to solve is scholars' disagreements on which indicators are the best-designed. However, what can we do when these explicitly designed indicators do not exist?

If the problem is lack of data then there is only one solution. Go out and collect more.

In Social Stratification, for example, you mentioned that there's been no data that's been specifically collected to measure social mobility.

Nope. Didn't say that. What I said was that in the UK since the 1970s there have been few (I can think of 2) surveys specifically designed to collect data on social mobility. Others have information you can use, but studying social mobility was not the main motivation or even a very important motivation behind them. Precision ~= pedantry.

Instead, scholars for social mobility rely on adopting the most suitable indicators from datasets that's been collected for other purposes.

Correct.

How can these scholars then assess the validity of their chosen indicators?

See, for instance, Evans & Mills (1998) *ESR*, 14, 1, 87-106.

My question this week is in regards to the King et. al. article: Has their method been adapted as a common practice since the publication in 2004?

See King's website for lots of examples.

And is it as well known/used in sociology as in political science? It seems (at least on paper) to be a good way of increasing the validity of survey answers.

It's known, but little used partly because it is expensive to implement.