SciencesPo

SÉMINAIRE

Centre de recherches Internationales ? Centre d'études européennes et de politique comparée

Les sciences sociales en question :
grandes controverses épistémologiques et méthodologiques

Compte rendu de la 72e séance

Cross-national survey data: traps to avoid

May 27th 2024

Nonna Mayer

Today our two guests will discuss comparative survey data and how to handle it. The development of large cross-national projects (ESS, EVS, ISSP) has considerably boosted comparative research. But comparing the collected data from one country to another is not obvious. Katharina Meitinger is assistant professor at the Department of Methodology and Statistics of Utrecht University and before she worked at GESIS Leibniz Institute for the Social sciences). Drawing from her own research and innovative mixed method approaches, she warns against the major biases to be aware of at all levels (concepts, question items, mode of collect) and the way to overcome them[1]. Emanuele Ferragina, sociologist at CRIS/LIEPP (Sciences Po) is a specialist of international political economy and comparative social policy. He discusses these methodological issues from a public policies analysis perspective[2].

[1] See "Detecting and explaining missing comparability in cross-national studies: The case of citizen evaluation of patriotism", *Survey Research Methods*, 17(4), 2023, p. 493-507 (with P. Schmidt and M. Braun); "Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives", *Social Science Research*, 110, Article 102805 (with H. Leitgöb et al.)

[2] "Maternal Employment and Childcare Use from an Intersectional Perspective: Stratification along Class, Contractual and Gender Lines in Denmark, France, Germany, Italy, Sweden and the UK", *Social*

Katharina Meitinger

I work on a lot of data. The Programme for the International Assessment of Adult Competencies (PIAAC)for instance which has a very interesting mixed methods pre-test. I'm doing a lot of consultancy work also for researchers who want to conduct cross-national surveys, and I give my input on how to design such studies or how to assess comparability. And I am the methodological consultant of the ERC project, Capable (*'Enhancing Capabilities? Rethinking Work-life Policies and their Impact from a New Perspective'*). As you can see, my research interest, where my heart is, it's really assessing comparability, working with cross-national data, how can we be sure that we can compare it, how do we know that differences are differences and similarities are similarities?

I'm working on measurement invariance tests, this is more the quantitative approach to assess comparability of data. And I did my PhD on the method of web probing, which is a qualitative approach of assessing comparability, but which has the advantage that you can also understand why data is not comparable. I'm also working on perceived discrimination of LGBTs and the cross-national measurement of gender identification. So basically, this is my main substantive research area in the last year and the ongoing one as well.

The goal of the presentation today is really to give like a primer on working with cross-national data. Either you have an existing data set and you're wondering if the data is any good. Or you're thinking about collecting the cross-national data yourself. Both situations have rather the same quality criteria. These would be the points I would look at.

So why cross-national data? Why is this topic relevant? Because there has been a lot of cross-national data and cross-national data usage in the last decades: the European Social Survey, the European Value Study. Also a lot of surveys covering basically all countries in the world such as the World Value Survey or the PIAAC (Programme for the International Assessment of Adult Competencies) I mentioned. And we have

Politics, vol. 30(3), 2023, p. 871-902 (with Edoardo Magalini); "Comparative mainstreaming? Mapping the uses of the comparative method in social policy, sociology, and political science since the 1970s", *Journal of European Social Policy*, 33 (1), 2023, p. 132-147.2023(with Christopher Deeming).

surveys that are specialized on specific cultural contexts distinct from the Western hemisphere. This is relevant because most measures are either developed in the US or in the European context and often taken and transferred to other cultural contexts. Obviously if we have a very distinct culture, it's important to assess if we can actually take this measure and have it travel to a very different cultural context?

Since I'm by training a survey methodologist, I mostly think about studies in the total survey error framework, meaning you think about all the different conceptual steps that you have to take to ensure that you have a very good survey statistic. Survey statistic could be, for example, mean level of health in a specific population. Then you have the measurement side, what construct do I want to measure? How do I operationalize it? How can I make sure that I really collect the response that I want to collect, how do I reduce the measurement error? And how do I make sure that it is also recorded correctly? Then there is the representation side, defining a target population. Who do I want to actually study? It's actually not always as straightforward as you might think. Here it's not only the French population, it would be, for example, French adult population from 18 to 65 years etc. Then you select a sampling frame, you would draw a sample hoping that respondents actually are willing to open the door or click on the email invitation and participate. You describe also how you have collected the data (telephone, mail, waiting room surveys). There is the questionnaire design, the pre-testing.

This “total survey error” approach is very useful in a single country context, it has to be expanded when we talk about cross-national data, or 3MC data (multinational, multiregional and multicultural contexts), implying different cultural backgrounds. People might have very different associations or concepts in mind when they read the same question, even if it's perfectly translated. And you have the translation issue. So when we work with 3MC data, we are in high danger of potentially comparing apples and oranges. Therefore, we have to be a bit more careful and address this complexity. And here you have the additional step of data harmonization.

You have to think it over, because there are country-specific variables such as education or occupation and sometimes you cannot ask the same question, and you really have to think about what kind of procedure will make the variables comparable

When talking about cross-national data, there are two very important concepts that are equivalence and bias. Equivalence means basically comparability of data. However, an important note is that everyone can understand equivalence differently. In 1998, there was a study by Timothy Johnson[3] and he found more than 50 definitions of equivalence. So you can imagine nowadays, there are probably hundreds of different definitions.

When I talk about equivalence, I ask if there are there measures or scores equally valid and reliable in all countries. The bias is basically the opposite. It's the nuisance factor that has a deteriorating impact on equivalence. The higher the bias, the lower the equivalence.

We have to really think about equivalence because this is the absolute precondition that we can conduct some analysis and we can draw some substantive conclusions. And we really want to know what are the real differences and real similarities and how much or how large is our bias in our data.

There are different types of bias as shown by Fons van de Vijver. The most important thing in cross-national data is that the construct is existing in all of your countries. If you have construct bias, that means that respondents either don't know the concept or they have different understandings of the concept. Think about constructive patriotism, for instance. This is a concept developed in a German context. Germans probably have quite strong connotations with this concept, but it might not work in all countries across the world. So when you have construct bias, you are basically in serious trouble. Assuming the construct is there, you still have to be careful because there are method biases and item biases.

A method bias means that you have a bias due to the method that you're using or to the context of the measurement. Here we can distinguish between three different types of biases.

There's sample bias, that means different sampling procedures are used in the different countries. WEe are ending up with different selection procedures, but

[3] Johnson, T. P. (1998). Approaches to equivalence in cross-cultural and cross-national survey research. In J. Harkness (Ed.), Cross-cultural survey equivalence (pp. 1-40). Mannheim: Zentrum für Umfragen, Methoden und Analysen -ZUMA-. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-49730-6.

potentially also different populations that we are comparing. Then you have instrument bias, the instrument that you use creates this bias. For example, you use Likert scales. the only fact that you're using Likert scales can create something called response styles. And there are different response styles such as acquiescence or extreme responding, I will have an example afterwards about acquiescence response style. And we have administration bias, that is basically the mode of data collection differs across countries, in one you have a face-to-face survey, in the other a web survey. And depending on what you're studying, this might create comparability problems.
And finally, we have the item bias. The question itself can create comparability issues and there are various reasons for item bias.

For example, we have a bad translation, and this happens more often than you think, also in the large scale surveys. It's not only the translation that is problematic it can be the source questionnaire, the English questionnaire is often an issue. Or you have, for example, key terms that don't work across countries. For example, one question asked how much you endorse civil disobedience, and the notion works really well in the European context, but Canadian respondents had very different associations, civil disobedience was very connected to aggressive behavior, and then you have a key term that is not comparable across countries.

I hope this overview of biases makes a good point that it's highly complex to achieve comparable measures, and you must be careful at very different points.

I have a couple of examples of the different bias types, like construct bias. This is really where you are in big trouble, because the construct might not exist, or take different forms in the different countries. Take the example of filial piety, being a good son or daughter.
When you take a look at those two pictures, you can see that it's a depiction of a Chinese concept of being a good son or daughter. Here you see, for example, the son kneeling in front of the parents, or here on that picture you can see that really the son is taking care of the sick parent. The Chinese concept includes basically the role of caretaker of elderly persons, the parents, and also this obedience aspect, and this is a broader concept than in the Western hemisphere. If you take a Western measure of filial piety, then you would basically miss important aspects of this construct.

Another example of construct bias is symptoms of depression in Zimbabwe. In Zimbabwe, you have somatic expressions of depression, for example, depressive people express their depression with somatic symptoms, for example, low mood, or fatigue, or difficulty concentrating. And those somatic expressions also appear, obviously, in Zimbabwe, headache, fatigue. But the issue is that people don't talk about depression, they would talk about thinking too much, supernatural causes or social stresses. So there, you have universal symptoms, but you don't have the construct in itself. You have to really be careful about your measurement strategy to really inquire about the symptoms and not about the construct.

Then there's method bias. And one of them is sample bias. This is a very important point that gets really often overlooked by substantive researchers, I see that a lot in my consultation. You have to really be careful, which sampling strategy has been used, and whether you have a mixture of sampling strategies in your country: the probability-based sampling, which you should aim for, the non-probability sampling, or respondents-driven sampling, the snowball sampling, convenience sampling. The issue here is, those sampling procedures often have self-selection bias in it. The respondents can decide if they want to participate in the study, versus here in the random procedures, probability-based procedures, they do not decide, they are asked to participate. So those sampling procedures have a far higher accuracy than the non-probability sampling procedures. Here, that would be my word of advice, if you look at your data, definitely check whether you have probability or non-probability data.

The second type of sample bias is, what target population do we have. You usually have slight variations in the target population. When you, for example, look at the data methods report of the ISSP, the International Social Survey Program, you see slight variation in age range. For example, in one of the countries, it's over 16 years old in other countries, over 18.

You also have to assess whether you include in the sample hard-to-reach populations or not. For example, respondents in the elderly home or people in prison? This might not matter much if you have a very general research question, but it might matter a lot if your research is about people in prison, or if you are studying the elderly population

across different countries.

So this is really a thing to think about, like, how does this relate to my research question?

Another method bias is the instrument bias, for example, the response style. Response style is problematic because respondents have a tendency to select a response category independently of the content of the question such an acquiescence response style. And the problem is that cultures differ in how likely they use a particular style. Acquiescence response style (or yes saying) is very often found in Asian countries, like in Japan, China. They have a higher tendency to agree. Whereas, for example, naysaying is the opposite, where respondents tend to disagree, or extreme responding, where they always choose the extreme points, and this has been found to be prevalent in Hispanic populations, for example.

Or sliders (respondents select a value within a range by moving a slider) I would never use because it has other issues, they're very imprecise. I have a strong opinion about sliders, I'll happy to discuss.

Also don't use Likert scales, because they run from completely disagree to completely agree, and basically Likert scales push this kind of response style. It's better to use item-specific response, where you ask, for example, "how proud are you"? And then you would say, "I'm very proud", that I agree with.

Another example of administration bias is when you have differences in mode of data collection, personnel face-to-face interviewing, telephone interviewing, you have it usually in those large-scale surveys. For example, ISSP has quite a mixture of modes, and this has to do often with the fact different countries have different traditions. For example, in Australia, it doesn't make sense to send interviewers around the country because of the huge distances. So different contexts push different modes, but you have obviously to be careful about them and to really think: does that have some impact on my data, on my analysis.

In particular, think about sensitive topics. If there's an interviewer in the room,

respondents are not that happy to report sensitive topics.

Here we have an example: how many times have you used marijuana?

So there, you have like a double issue because it's a sensitive topic, but then there is also cross-national variations of legalization of Marijuana. So like you have kind of like a wide hot pot of what could be going on when the respondents answer to this question.

So yes, variations of mode can create over or under reporting, but the good news is, you could also control for that.

And I have also an example of item bias. It's an example from religious involvement, a measurement of religious involvement in the ESS round two, and it was measured with three indicators regardless of whether you belong to a particular religion: how religious would you say you are, apart from special occasions such as weddings and funerals, how often do you attend religious services nowadays, and apart from when you are at religious services, how often at all do you pray. This measure was basically not comparable because it does not work well in Muslim majority countries, like Turkey, where it is customary that women do not attend religious services. So what would happen is that you would completely underestimate religious involvement for these women.

As you can see, there are a lot of different examples for like where bias could appear, and therefore you should be careful when you look at cross-national data. And this is basically my personal checklist to assess what data is there, would I be interested in this data etc. Basically, I would like to see that the question development was taking cultural context into account.
I would like in particular to know that the translation procedure was using translatability assessment and using the correct translation procedure. The question should be pre-tested, the sampling and data collection should be sufficiently comparable. If possible, equivalence should be assessed, and I want to have l a thorough methodological documentation. Basically, that goes a bit hand in hand, because usually you cannot check when you don't have the documentation.

Why is it so important to have the cultural context taken into account? Because it's often not done. As I mentioned already, measures are often developed in the US or in Europe. So it happens often, that good translation procedure is not used for other countries, and yet everything is like presented as differences or similarities from one country to another. And it's not often the case that questions travel well. So before you start translation, at the question development stage, if you collect your own data, first take a step back and really think about, do we need to adapt it to different cultural contexts.?

So before you have an existing set of questions, or when you have it, you want to critically assess whether translation would be sufficient. Usually, when you are in a cross-national project, you have already some resources to get some cultural expertise, think about your collaborators. If you have collaborators in multiple countries, then you can just ask them about or give them the existing measures and ask them, would that work in your country? And if they say no, then you really have to follow up and ask, why doesn't it work?

Another source, a very valuable source of cultural expertise, are translators. What I usually do, when I have a question translated, is also give some extra hours where they can give some cultural expertise additionally. Another very good way to assess whether a construct, for example, is existing in a country or not or which different dimension it has in a particular country, is to use focus groups. Not necessarily to assess the question, but before you draft a question, to really think about what themes the participants of the focus group have in mind, and then try to grasp all the dimensions in your questions.

A very easy tool to assess cultural appropriateness is a question-appraisal system (QAS). This is a quite short list of questions, which has to assess basically the question that you are developing, and it has also some questions regarding cross-cultural considerations, and then you have basically a way to find out that the knowledge may not exist.

Basically, the idea of the QAS is really you sit down, and then you go through all the different points, and you check each question, and it helps you to really critically assess

your question.

For instance ”Think about different sports, drugs, food, springs, activities, health system”.
You obviously would have to adapt everything in each country. And I think this one is an interesting example, how often do you watch football on television? Every respondent in Europe and in the US would understand this question, but they would answer something different, because American football is something quite different than European football. So this is basically a question that was developed in the US context, it would work very well in the US context, but it might be problematic elsewhere.
Another problematic question is about opioids. I'm not sure whether all French respondents would know what this is. Also, the same with the idea of “over-the-counter pain relief”. And “health professionals”. Here you can see this is a very vague term. It leaves a lot of space of interpretation. There are differences across countries, and obviously very specific medicine should might be adapted for each country.

So now let's assume you were developing your questions, you were doing focus groups, you reached out to your cultural expert, you have now a set of like a source questionnaire of which you were very proud, probably in English language, and you want to then translate it.

Before you translate, you should take a step back and be very critical of what your source questionnaire is, because usually a lot of translation issues appear because that source questionnaire is problematic. Once you have fielded it to a lot of different translators, this becomes problematic and very complicated because then you have to go in a second round of translation. So better have a very good source questionnaire from the start that can be very smoothly translated. The role of translation is often underestimated. If you ask, for example, for research grants, for money, I highly recommend to provide for a proper translation procedure, because translation, gold standard translation procedures, are more expensive than the traditional or frequently used ones.

What you often find is that you have a single translation with basically some person

doing one translation. And it's often, a colleague or collaborator. The next step is the back translation, basically you take the English source questionnaire, translate it into French, and then you translate the French version into English. And then you check whether the source and the final translation are roughly the same. This can show a variety of errors.

And if you are very interested in this one, there's a very nice article on that by Dorothée Behr[4]. So better first do the translatability assessment and then do the TRAPD (*Translation, Review, Adjudication, Pretest, Documentation) procedure of translation. The translatability assessment is basically critically going through your questionnaire and thinking, are there some other things that are very difficult to translate in different languages? Here you can also use country experts. First, send your source questionnaire to collaborators and ask, do you think it' can easily be translated or not? And if not, why? This would be a precise input. But then QAS is also helpful here because it points at what to look out for. What is really difficult for translation are idioms, for example, here you have one, “that you could not shake off the piece”. Very nice in English, but how do you translate it to Polish or Chinese? Even French.

Also, the use of acronyms is problematic and probably not everyone would understand what ECLSB is. And if it's very important, you should add explanations about what you're actually asking. However, if you have to provide a definition, you make the question more complex.
So that's also a caveat. Also in English, mind the unclear use of the term “you”. Rather tricky because it can be singular, plural, it can be masculine, feminine, formal, informal. This leaves a lot of ambiguity for question translation. You should be very careful here. Also from my own work, I use an American scale that is about also the medication: You can see how problematic it would be, for example, to translate “be good” or “get high” or “because I am hooked”.

Another thing you should be careful about is the response scale. They're often overlooked and particularly if they are very vague. What is “a little”, “a lot”, or

[4] Behr, Dorothée. 2017. "Assessing the use of back translation: the shortcomings of back translation as a quality testing method." International Journal of Social Research Methodology 20 (6): 573-584. doi: https://doi.org/10.1080/13645579.2016.1252188.

“somewhere in between a little and a lot”. Different cultures will understand in very different ways what means “a little”, right?
You could ask about more precise references, for example, or like “how often a week” or so.

Once you have done this translatability assessment, you can continue with the translation and the gold standard of translation is something called TRAPD procedure. It basically means you have like a parallel translation. You have two translators translating the same questionnaire. There is translation one and translation two, and then they meet up with a third person involved like in a review discussion. These discussions usually take a lot of hours and then they really just look at the two different versions and they combine them and try to find the best version. Sometimes what happens is that translation one and translation two are merged into like a translation three, but it gives a lot of inspiration of what would be a good translation.
If in the review, you can't come up with a final version, then you would take additional information or additional input to come up with a final version.

Basically questions are developed, they are translated. What you still want to know is : do the question work for the population that you want to study? There can be a big difference between what the researcher thinks is a good question versus what some very old illiterate person in the rural countryside will think. Therefore you should really think about a pre-testing stage. You want to really assess the equivalence of the translation and verify that all key terms work equally well in all countries. A possible pre-testing approach, for example, is a quantitative pilot field test, like a sub-sample of 100 respondents just getting the final questionnaire. The no-answer rate gives you some information, about questions that might not perform well. You can also do interviewer or respondent debriefing, you both after filling out a questionnaire, were there some issues? And you could also do qualitative approaches, for example, cognitive interviewing or web probing. As I mentioned, I'm expert in web probing, so if you ever have a question, I'm happy to share.

Cognitive interviewing is the idea that the interviewee would answer the target question and then the interviewer asks follow up questions like, for example, why did you select - completely agree at the previous question, or what do you understand under the term

“civil disobedience”. Web probing is face-to-face approach in a web survey, basically here respondents get a target question, and after they give an answer, they get an open-ended question with a probe.

This gives you the qualitative insights about the cognitive question answer process, how do respondents understand, do they have similar perspectives. It also detects silent misunderstandings, for example, the respondents give an answer but they actually understood something else, misunderstood the question. So web probing or cognitive interviewing tells you if respondents provide similar reasons when responding, when choosing a response category, does it mean the same, when, for example, you ask “how proud are you of being American”, and someone says “not proud at all”, to a German that says “*behaupten stolz*”, that is basically also “not proud at all”. Because in Germany, you have a lot of different issues, , and it's not acceptable to be proud of being German.

Or respondents may adopt very different perspectives when responding. For instance “how proud are you of America with regard to the social security system”, it was translated to “sozialstaatliche Leistungen”, and Systema de Seguridad Social, in Mexico, and basically what happens is that Americans think mostly about the pension system because that has like a very similar labelling in the US. While Seguridad Social in Mexico, they think about social, like the safety on the streets, like the Seguridad instead of Seguridad Social, that's not safe to go on the street because there's crime and robbery.

So if you're interested in web-probing, there's a brand new book chapter that really summarizes the state of art of web-probing, in the International Handbook of Behavioral Health Assessments, if you're interested, I recommend to read that one[5].

And obviously also you need to take a look at the sampling and data collection procedures, what I was discussing before is all about the measurement side, you also should be a bit concerned about the representativeness of the data collection. Here my really, really quick check is like with regard to sampling, it's important to understand

[5] Meitinger, K., Neuert, C., & Behr, D. (2023). Cross-Cultural Web Probing. In *International Handbook of Behavioral Health Assessment* (pp. 1-20). Cham: Springer International Publishing.

that if we have cross-national data, there are by nature differences in sampling. This has to do with country context, for example the availability of registered data in some country or not. But what I would check first is whether we have always probability-based approaches versus non-probability-based approaches? Do we have a very clear target population in each country? Are there huge discrepancies or is it not reported at all? And with regard to the modes, you have to be really careful because these can really blur the differences in your data collection. But it's also natural that you have some variations, and some mode combinations are also not that problematic.

Think also about interview involvement, you have different modes like telephone and face-to-face interviewing. If you think it's important that the respondent actually sees the response categories or the slides, then you should really think about possible variations in the visual representation. For example, in telephone interviewing the respondents by definition cannot see any scales, versus in web mode, the respondent can see it very well or like face-to-face interviewing allows for showcase.

And finally, equivalence of data should be assessed if it's possible and that is also a point that I really want to make. If you want to collect cross-national data, you have to think about equivalence assessment before you collect the data because equivalence assessment measurement and variance testing presupposes that you have a construct that is measured with multiple items and then you should draft your questionnaire accordingly so you can do this quality check afterwards.

When you have this construct, let's assume that it is patriotism, that is measured with multiple items and the idea of measurement and variance testing is basically that we test different aspects of the measurement model. For example, if we have the equal factor structure, basically all items have a relationship with the construct somehow in all countries and the metric measurement and variance tests for equal factor loading, so this relationship is kind of equally strong in all countries and the scalar measurement and variance tests for equal intercepts, so the origin of the latent scale is the same in all countries. Configuring measurement and variance, if you don't find it, you have a big problem because if there is no factor structure at all, there is not much you can do.

With metric measurement and variance, you can already start interpreting

unstandardized regression coefficients across countries and for scalar measurement and variance you could for example compare the latent means of the construct across countries. This is important if you want to do multi-level analysis or ranking of countries, that's basically what everyone is aiming for. And if you're interested in measurement and variance, this is also a state-of-the-art article of measurement and variance that I was writing a chapter on, it gives a very nice overview of like what is like the history[6].

Maybe a last point, methodological documentation, and here I had the example of the ISSP, how it could look like.

So to conclude doing cross-national analysis is very important because this can give really crucial insights on how cultures differ or are similar across the world. It's basically a very interesting thing to know, but you have to be really careful because 3MC data is more complex than just a single country analysis and you have to be careful when you work with 3MC data or you collect 3MC data, you have to be very strategic about the analysis because you have to plan for this comparability before you develop the data, before the data collection, you have to be very careful during the data collection and also after the data collection to assess the comparability.

So you have to reconsider various aspects. Assess the data collection procedure, was it any good, assess the question development procedure, the translation procedures, following guidelines or like gold standard procedures, was there some pre-testing, is there a good reflection on how we do sampling and selection of modes of data collection, is here any consideration about equivalence there and is there a very good documentation.

What you usually find is that surveys that have achieved comparability and of a very good high standard, have usually very extensive documentation, so the amount of documentation already is like a very good l primer on the interest too look further into this data.

[6] Leitgöb H, Seddig D, Asparouhov T, Behr D, Davidov E, De Roover K, Jak S, Meitinger K, Menold N, Muthén B, Rudnev M, Schmidt P, van de Schoot R. Measurement invariance in the social sciences: Historical development, methodological challenges, state of the art, and future perspectives. Soc Sci Res. 2023 Feb;110:102805. doi: 10.1016/j.ssresearch.2022.102805. Epub 2022 Oct 31. PMID: 36796989.

Emanuele Ferragina

I thought of some conceptual issues that can be important at this stage because there are people connected. The general research question of the presentation is, for me, very fascinating. Can questions travel successfully across national and cultural borders? We can approach this question in two ways. One is very destructive and the answer would be no. The other one is to actually find ways of doing this work well and actually understanding what the limitations are, and in this Katharina's work is very, very interesting.

So first of all, should we accept the possibility to have perfectly comparable survey questions?

Perfect comparability in comparative research is impossible. It's like when you look at the welfare state, when people criticize the inefficiency of the welfare state, well, we need to have a part of inefficiency in the system to fight against poverty, for example, and I think the same applies to comparative research. We need to accept that there is a part of things that cannot be compared, or they're lost in whatever translation, a bit like in a movie by Kurosawa about the Rochman effects, and the way from where you look at things does have an impact on what you see. But it is very important also in the ambition we have for comparison.

At some point we had this idea, you know, think about the World Value Survey, that we can compare everything. But this is, of course, highly problematic from an epistemological perspective and from a measurement perspective. Sartori already in the 1970s spoke about conceptual stretching. This, you know might happen when the ambition overtakes the capacity, the means that we have conceptually. Knowing the limits can help properly discuss these things. More specifically, I think there are some other themes than those Katharina has been putting on the table that we can approach in the discussion.

First, the idea of functional equivalent. We go about translation of words, but what we're looking for here is meaning. You had the example about football. I remember when there were all these discussions about social capital, what are the activities that people do? Well, for the Finnish, you know, playing pétanque or whatever is very important, but for other people it might not be.

In a sense, if we want to compare, and this you showed quite well, but it's important to be emphasized, literal translation of things is not the good way of comparing. So if we want to compare societies, we need to understand them and actually compare activities that have functional equivalence in the process of socialization. I will go even further. If people want to compare, they should compare social phenomena they can comparatively understand.
My main experience with all these issues of translations comes from the study of social capital. That's why I rely on this but I'm sure you have thousands of examples that can come from other fields.

Then there is another issue that has to do with financing of research and how we do research.
The incentive to do all these things in comparative survey research is very little. Nobody asks you for these things. Either you have people who don't believe in these things, they will tell you whatever you do about attitudes anyway doesn't work, because attitudes cannot be properly measured and compared, or you have people who bypass this. And I link that to what you were saying about measurement invariance, because it used to be one of my passions, structural equation modelling, confirmatory factor analysis.

There is a strong aversion in the social science field against descriptive techniques and actually against causation. And very often descriptive techniques are not considered as inferential technique, which is, of course, a mistake.

So what you show about using confirmatory factor analysis, for me, is the standard. You want to measure it in concept. Again, the example of social capital, what you first do is measure different models with different items, and you see if they give you similar ordering across countries or they give you similar effects across cross-national research. This is something that can be easily done even without having all the armaments that you propose.

There are some cases where it becomes very hard to come down to certain conclusions. So sometimes we should do a lot more of the current service we have. =on this, I think I did my fair share in the past when I used to work with all these things.

But I think it's also a way to interestingly understand more about your concept. If we move the questions away from methods and we bring it to substance, trying different models that capture a similar latent concept is an interesting exercise because it allows you to reflect more on the concept you're trying to measure.

We're not a bunch of plumbers who are just there to operate on things, we want to understand the concept we measure. And therefore, I think that the idea of looking at measurement in balance is very important. Even when it varies, trying to understand why, if it's part of a specific bias of the concept, if it's because actually adding up new items to a specific construct changes the nature of certain things and allows us to reflect back on what we're doing.

Another issue is the division between attitudinal questions and questions that relate more to factual things. Very often, research has implied that attitudinal questions are a lot more problematic than factual questions. I give you the example of time use.
I actually think that there is not so much distinction between the two because people don't remember these things very well or they reply in a way that is conformed to social norms.

So once again, I think these procedures are not only very useful for attitudinal research and attitudes, but they're also very useful for research and more, between brackets, factual questions.

To conclude, I think that the things that should be retained, - but then, of course, we can discuss about many more- are ambition, compare things that we can compare, bias. And here, Katharina, I might have some slightly different view from you.

Some of the biases you're talking about need to be fixed, but some of these biases are the essence of comparative research in the sense that a concept that cannot travel in a certain culture, and we observe that, it's a matter of interpretation. So, for example, we ask the questions about football, or we ask the questions about how people spend their time, or socializations.

The very fact that we use a question that is standard and it doesn't work, makes it's

very interesting to understand where countries are positioned. So I mean, it's not that I'm in disagreement with you. How can one be in disagreement with what you say? But the fact is that sometimes a bias, is part of the interesting interpretation of the comparative context that we have. So I would actually say that when this bias is controlled, under the limits of certain procedure, then it's something that can be useful for the analysis. And finally, I would like to re-emphasize, if we want to use probabilistic survey research, we need to work harder on descriptions, and this is something we don't do anymore. And personally, it makes me quite disappointed, because I think it's a very important step, but I'm very happy to discuss more, and thanks a lot.

Katherina Meitinger

I'm not sure we're different there, because I fully acknowledge that measurement, like this whole bias typology, should not just make one-form–fits-all, and we force it on all different countries.

But these nuances that you're talking about, like which sport is working in different countries, obviously, you need a good approach, but you'd also have to keep the complexity of the data.

And for example, regarding measurement and variance testing, I think, and this point is also related, like a lack of comparability, I would also argue, similar to you, it's not the end of the world. It should happen that countries are sometimes not comparable, because they are so specific. And then you could actually go a step further and try to understand why they are specific, and you could do a Bayesian structural equation model, or like a MIMIC (multiple indicator multiple cause model)model, and really try to capture the country or culture specificities, and use basically a lack of measurement and variance as a starting point for your research, and really delve into the cultural particularities. So, I'm not sure whether we disagree so much there.

Questions

Noemi Piollat (M2 Sciences Po)

You mentioned many cross-sectional surveys, but which ones are the best following your criteria? I was wondering, because I worked with EVS data, and I didn't check all these things, because some I just took for granted, which is probably my mistake. I was wondering if there is somewhere a list of those that are safe, that you can use.

How do you then very concretely choose, if there is no other way to measure what you want to measure, because these questions are only asked in this survey, but this survey has many limitations, is it enough just to assess and to write a paragraph saying, yes, I'm aware of these limitations, but here you go, and here are my conclusions. I'm thinking about very concrete things when you're doing research, and how do you deal with them. So basically, I have two main questions, right, which are like the good surveys to start with.

Katharina Meitinger

So I think, basically, a good indication of what is a good survey is the existence of an extensive documentation. Another indication is if they have their own methodological research going on, if they're really focused on that one. For example, the European Social Survey Program, it's a very, very well-funded program that has an extensive methodological program to assess all different aspects.

So they do, for example, research on including institutionalized respondents or not, for instance. Had. I have a reference, the article by Lukas Schanze[7] .

Another indication is how it is structured, is there a centralized or decentralized core team doing the study. For example, the ESS has also a centralized secretary with a core team and country team, so they have a very good structure. However, their disadvantage is, obviously, a limited geographical reach. You have Europe, and I'm not sure if Russia is still in there, and Israel used to be in there. But basically, you're very Western-focused, and that's maybe not what you're interested in, right? So then you have, for example, the World Value Survey or the International Social Survey Program, the ISSP. ISSP also has a very strong methodological approach and very extensive documentation. But it’s a very decentralized structure, and the funding is not as strong as for the ESS.

And the other aspect i “I don't find the question that I need to have in the questionnaire” is a tricky one, often you have to then collect your own data, but you might not have

[7] Schanze, Jan-Lucas. 2023. "Hard-to-Survey and Negligible? The Institutionalized Population in Europe." *Survey Research Methods* 17 (1): 91-109. doi: https://doi.org/10.18148/srm/2023.v17i1.7830.

the resources or the funding.

Another thing is if you do this research and you find a very particular question, like, typically, national identity. One question you find in all the programs is the National Pride question. But if you would be, for example, interested in, I don't know, some kind of nationalism or a very specific concept, I would be very careful about just taking the National Pride item and then, labeling it nationalism. I think you have to be very careful when you go in the questionnaires very realistic what you're measuring, and is it actually a good indicator, or is it only that it's available everywhere? That's a big challenge that you have a specific construct in mind, but, you don't have enough items, or it's not exactly measured as you want to. This is basically the huge issue of secondary data use.

Noémie Piollat

I think that ultimately, this checklist is extremely important, and especially because as master's students, we had a quantitative methods class, and we really never, you know, reflected on this part of the work, which is probably the most important part, and very essential to understand what you are actually measuring and comparing.

But then, checking these six items, realistically, I don't know how many surveys could actually check all the boxes, and also, I think that, it's kind of narrowing down all the accessible data we have. Also, I believe that just checking and having experts for all the countries and translating everything is probably really, really costly. And it's also something I was thinking about, it's not really a question, but about doing comparative research with no money or less money.

Katharina Meitinger

First of all, this checklist is like gold standard checklist, that's what you should strive for, it's like the criteria that I would like to see, but then there are also variations. With regard to your money, or "do I have the resources", I have two points. The first one is, if you don't have a lot of funding, why do you need to do a study in 36 countries with 40 languages? That is, what kind of data do you get, probably not good data, and how can you trust the results, then probably not much. So the question is, like, what kind of realistic research design can you achieve with the money you have? It might be you

don't need 36 countries, if you are actually interested in a policy comparison let’s say four countries would be enough, for each type of policy, you select one country, and you make a very good reasoning why you select this country? So you boil down, you reduce the complexity of the cross-national data.

And then the second is, I think this checklist can also inspire you to find creative ways of funding. Think about, you might want to put your money at a probability-based data collection somehow, but like with the translation procedure, you could think about, using your country collaborators and two of the teams, and then you make parallel translations, right? So does it have to be a professional translator every time, or do the research collaborators have sufficient language skills that they could translate the questionnaire? Probably they have, and also expertise about the topics, so this might be even more helpful. So think about good ways where you should focus, where you should really put the money, and where you could have cost-reducing strategies.

George Marcus (emeritus professor, Williams College)

I have a fairly long list, because I read the nationalism, populism, patriotism, stuff, and that got me thinking about it. So why don't we do this step-by-step, if you want, I'll give a little information, and if you think I need to know more, just ask me. One thing you didn't mention in your checklist is, are you a scholar using the data for parameter estimation or for hypothesis testing? Because some of the criteria for the data change, right? If I'm hypothesis testing, I don't really care whether the parameters are being accurately measured. I want to just hopefully have some mechanism for causal attribution and testing. That may be a consideration on whether the data is suitable or not.

And then you mentioned, getting ready for the survey and using focus groups.

My own experience with them is they're less valuable and I would propose an alternative, which is Q-Sort. Do you know about Q-Sort? Very briefly, and I can give you citations, but it's fairly easy to find. The idea is to create as wide and diverse a sampling of statements about some phenomena, like sexism or patriotism. You can do that either looking at literature, at novels or stories or whatever, and get as many as you can. The Q-Sort is a technique where each statement gets its own separate card. And you only need a very small sample of people to organize it. Out of that, you can

extract the underlying themes that these statements together identify coherent groups, and which ones are important to different people.

There's a very good paper by my research colleague, John Sullivan[8]. In the 1988 U.S. national presidential election, that was George H.W. Bush against the governor of Massachusetts, Michael Dukakis. Bush ran a heavy-duty pro-patriotism element in his campaign. And John said, wait, that's not what I think about as patriotism. That led him to do this study. He found that actually Americans have at least seven discernible different views about patriotism. And he was ticked off that the Democratic Dukakis didn't respond by hearkening people to these other views of what love of country means.

So for almost any concept you can imagine the same, for instance what do people make of sexism? You'll come up with 50, 60, or so statements, whatever number. You could also, obviously, do that country by country, because this is a very low-investment process.

It takes a little bit of deciding what kinds of materials, but again, the literature on QSort is vast there is a nice book by Steve Brown that gives you examples of how to identify and collect these[9]. By small sample, I mean under 100 people. And you don't need random sampling.

All you want is diversity. If you have 100 different people from a country, you'll find which themes. This person likes A and C. That person likes four of them. Another person likes three different ones, and how they've clustered them. So that would be my recommendation. It's a little systematizing, but it's probably enough. But it's much better, because you don't understand.

Katharina Meitinger

Why is it much better?

George Marcus

Because we're probing people who differ in terms of how valuable they are, and how willing they are to write, and how competent they are to write. And you're asking them

[8] J. Sullivan, A. Fried, Mary G. Dietz, Patriotism, Politics, and the Presidential Election of 1988 *American Journal of Political Science*, 36(1), 200, 1992.

[9] Steven R. Brown, A primer on Q Methodology, *Operant Subjectivity*, 16(3-4), p. 91-138.

to generate verbal response to things they may be quite inarticulate about. If you ask people, my favorite example is, what is good behavior in an elevator? They look at you like, what? There are at least 20 or 30 different rules about being on an elevator. Where you stand, where you look, how close you are to other people, what you make of people, how do you identify them as a couple, or a group, or a family. It's very complicated. And we do all those things without thinking about them.

Which brings me to another theme you might want to suggest people look at when they're evaluating either a large group of questions, or creating their own instruction set bias. This was done largely in the 80s. A lot of scholars think that by posing a question in the following fashion, they will get more accurate responses. "So, when you think about patriotism, how important is it to you?" There's been research on that, it introduces a bias. Asking either, how do you feel about patriotism, versus how do you think about patriotism, or any other topic, you get a difference. Now the question is, which is the valid response? It's not the thinking one. Not only does it introduce the bias, it produces a social desirability effect on what people think is an acceptable answer.

So the way the ANES (American National Election study) deals with parties' identification is done is using a "think" instruction, which overestimates by about 5% to 7% how many Democrats there are in the country. And they will not change it, because their money comes on continuity. We're getting the same things we measured in the 50s, 60, 70, 80 years later. And that's why you should keep this going. If you're not acquainted with the instruction set literature, by the way, I'm happy to send out, I have most of these papers online, so I recommend it.[10]

The other factor to pay attention to is what discipline are you in, because that dictates the articles and books you will read and value. And some disciplines are heavily walled, and never read outside their own discipline. Others are much more critical. And there are walls even within disciplines, too, into subcategories. If you don't try to find out what these other streams are, you're going to be stuck with unstated presumptions that I've

[10] Neely, F. (2007). Party Identification in Emotional and Political Context: A Replication. Political Psychology, 28(6), 667–688. Burden, B. C., & Klofstad, C. A. (2005). Affect and Cognition in Party Identification. Political Psychology, 26(6), 869–886.

looked at the literature, and here's what it says. And someone from another discipline will look at it, and say what are you talking about? That's not the way we do it. So you will find yourself, it's much better to fix issues before you invest your resources than after you've spent them.

That's the point about bad data. In particular, I'm not sure if you meant it this way, but if you ask people about perceptions, they're often very good at answering that. But the literature I'm familiar with says, yes, but that doesn't dictate what they do. And there's a lot of data on this now. If you ask people, how strong do you think Russia is, they can tell you. But that won't influence what their position is in the Ukraine war, or what their position is in the world, or whatever, because human beings are normative creatures, not accurate perception creatures.
So that's another thing to be worried about. If you find an existing survey that has data, if it's heavily loaded on perceptions, it's probably accurate, but it won't give you much in the way of causal interpretation power.[11] I'm a big fan of multi-method.

So surveys are great, but if you want causal imputation, having experiments is good, but experiments have terrible external validity, because it's so sort of bounded within a wall. No one acts in an experiment the way they do in real life. That's the strength of an experiment, but it's also its weakness. But I don't know if it's prominent in Europe yet, but a number of us got active in investing resources in survey experiments. So in the US, there's an organization called TESS, T-E-S-S, (Time Sharing Experiments for the Social Sciences) and they welcome it, and you get free data. You just make a proposal, if they like it, they'll put your experiment in, and you'll get the collective sort of baseline data that all surveys want to know, rich, poor, male, female, et cetera, et cetera. So you just have to create a module that has an experiment. I think they are open to non-Americans. So if anybody's interested in exploring that line, I can get in touch with whomever.

One more thing, and this comes out of your study on patriotism. You use the term

[11] Friedman, J. A. (2019). Priorities for Preventive Action: Explaining Americans' Divergent Reactions to 100 Public Risks. American Journal of Political Science, 63(1), 181–196. Marcus, G. E., Wood, S. L., & Theiss-Morse, E. (1998). Linking Neuroscience to Political Intolerance and Political Judgment. Politics and the Life Sciences, 17(2), 165–178.

negative emotion as one of the distinctions between positive emotion and negative emotion. Negative emotion is going to go away in the wind, because there' is no such thing as a scientific concept. The basic components of negative emotion are anger and fear. They have different neural substrates. They have different antecedents. They work in parallel, and they have different downstream consequences. They don't share anything. One is mobilizing. One is inhibiting. And there's a bunch of other differences. So that's, again, a point for reading the literature.

There are still people using valence. There are still people using positive and negative emotion. I can give you any number of papers published just in the last six months that do that, because they've never run into someone to question and get them thinking about it.[12]

And then let's see. Remember that if you're creating a research program, your advice about thinking about your own data rather than using lousy data is an important point. Even if it's a small survey in a particular locale, maybe teamed up, for example, if you go to conferences and find that someone from Romania has a similar interest, maybe they can put together a sample of 100 or so, and you put a sample of 100 together. But you get complete control over the data that way. I don't mean control in the sense of we have it, and they don't. But you don't have to do these trade-offs between good data and lousy data. That’s why I've never been willing to use collective data because of how badly so many things are measured. And I don't have to make excuses for that.

Katharina Meitinger

Thank you so much for all your points, really interesting. But I also want to say, that was my caveat at the beginning of the presentation, obviously, this list is not exhaustive. So I will definitely look into Qsort. This is highly interesting.

Emmanuele Ferragina

Before you were saying about time invariance, I had this point, but I didn't bring it up in

[12] Marcus, G. E. (2023). Evaluating the Status of Theories of Emotion in Political Science and Psychology. *Frontiers in Political Science*, *4*(1090884), 1–20.

the comments. One of my doctoral students does research with cohorts and uses ISSP over the long run. And we keep a lot of bad questions to keep over time consistency, I mean, this is, also a big issue that probably we have to discuss, what can we do about this?

This is, you know, something very, very complicated as well as the other stuff about artificial intelligence. We're going to have more and more data and we're going to live in a world in which data analysis is going to be very easy to do. I mean, I go very extreme, but perhaps we might not need doctors anymore to really visit people but just to collect data. So the best way of taking track of people's health would be to put data in the machine and then ask the algorithm, what is the health of the person?

But the issue won't be any more about data treatment the issue is going to be about the quality of the data that we feed in order to have these things to run. And I think we're not ready at all for this, in the sense that most of the research is totally unbalanced. I never in my life, I've never had a class in which people were talking about these things. It's very rare. And actually it's going to be the main problem in the sense that doing data analysis is going to became increasingly easy. How we collect the data, what do we do with this data? How do we check for the relevancy of this data? And a lot of things will probably be done with artificial intelligence in terms of translations, opening up to a lot of the issues that we have discussed. So I know, I mean, nobody has answers to that probably, but I think this is a theme that is interesting in itself for, for future discussion.

Katharina Meitinger

I think I want to first react on your first point about basically longitudinal measurement invariance. And we keep on having the same like problematic measures. I think also, the burden falls a tiny bit also on the researcher in the sense that you have to be realistic in what the data can do. So I think my main point when I talk about this topic is you have to assess your data and then what conclusions can we draw from it. What kind of interpretation can we take?

Can we make strong recommendations, recommendations based on this, or do we have to be very careful in the interpretation and take our analysis and results with a big grain of salt?

And I think this is a really important point, that probably these data are measuring something, but what do they measure?

For example, if you think about sexism measures, they might have worked very well in the seventies or eighties, but then society is changing and let's say the sexism level stays the same, but what does that measure, right? Is it differences in understandings or is the attitude stable?
And I think if you are a very good researcher, you should have a good knowledge of what is happening and should be able to put this kind of results into context.
So like measurement invariance tests are existing for longitudinal panel data, they exist also for longitudinal cross-sectional data, there is a, I think a very nice article by Daniel Seddig[13]about cross-sectional measurement invariance.

Nonna Mayer
But what you are telling us here, it's not only for comparative data, it's for anybody who's working with surveys, it's really the things we should work with. And I like the idea of experimental surveys. It's one way out. Her we worked a lot with Paul Sniderman and he loved splitting randomly the sample in four, five, 10 sub-samples and just change one feature of each story proposed r. It allows you to pinpoint what factor really makes the difference in the answer.
That's something that I'd really use. Also, I'd like projective tests, such as photographs, online surveys are useful for that. And you get really things, people haven't the time to reason, they just react to the photo, you suppress the social desirability bias, and on racism, it works fantastically.
And I love back translation. We've done it many times in comparative surveys on racism, the full catastrophe, often showing that the translator hasn't understood at all what you meant in the beginning. That's easy and a very good exercise we can all do when we translate or import questions.

Last, what I really would like to ask you more about, is how do you consider the different

[13] Seddig, D., & Leitgöb, H. (2018, April). Approximate measurement invariance and longitudinal confirmatory factor analysis: Concept and application with panel data. In *Survey Research Methods* (Vol. 12, No. 1, pp. 29-41).

modes of conducting surveys? We coordinate the National Commission for Human Rights on Racism, and we've managed to get fund exactly the same survey, with the same questionnaire, online and face-to-face. And the results are striking, sometimes you have 25 percentage points difference. The structure of the answers is exactly the same, but not the level of racism. And we made an experiment, it's not the fact that the person all alone in front of its tablet answers more freely than facing the interviewer. It's the sampling. What really makes the difference between the online survey and the face-to-face survey is that it's not the same people who are interviewed in both. The face-to-face sample is socially, ethnically, culturally more diverse, closer to the population at large, with more respondents who are not educated, who are not registered on the electoral list, who are not French, than the online sample. Because of the digital fracture. Did you work a little more on that? We did an experiment where at the middle of the questionnaire, during the face-to-face survey, to a third of the sample we gave a tablet so they could answer alone the remaining questions. And when they answer alone, without any social desirability bias, their answers are practically the same than those of the other respondents, who had no tablet. And both were fr more tolerant than the online respondents

Katherina Meitinger

So you mean that if you restrict the sample of the online survey to be. if you, for example, modify a little bit the sample of the survey that you've done in person to really match the one online, you get the same results?

Nonna Mayer

I didn't match anything. I used the face-to-face survey sample

Katharina Meitinger

but if you match the sample, what happens if you match the sample of the physical survey with the one online?

Nonna Mayer

I pooled the two surveys, and checked with a logistic regression on tolerance to immigrants the impact of the mode of survey controlling for sociodemographic variables. It's not going online or face-to-face that matters.

George Marcus
Obviously, lots of other countries are wrestling with the same issue. So the American National Election Survey had been wrestling to do this by telephone, they mixed the methods, yes.

And so they had drawn various conclusions. Since I don't rely on them very much, I'm not sure you're my friend, but, anyway, so are you comparing, are you- not specifically, but to the extent that organizations, sampling organizations, are they conversing with each other about what they're concluding about best practices and coming up with conclusions that seem valid across the different institutions?

Nonna Mayer

Yes. ESS for instance has tested different modes and came out with a lot of publications trying to see what's best.

George Marcus
Is it better to do face-to-face?

Nonna Mayer
At least try to mix different data and compare.

Katharina
So they have done it. it's interesting, because what they put forward is still face-to-face, in spite of all the problems. So like the ESS is now starting this Cronos panel, right? This is an online panel. So we'll see what happens. But they have been trying for quite a while in several countries, they've done experiments.

We have done it with- it was very interesting to recall that we had this experience with shared data. So that basically we had- we were measuring mental well-being and subjective well-being over time. And whatever we did, both with elites of the general population, then with the shared size of people above 50, we noticed that there was an upward trend in mental well-being at the beginning of the- of COVID, of basically

lockdowns, and so like, you know, the first things you guys say, well, it's about the matters of the interview.

And the second thing is, you know, how we can go about it, especially about the population that have been excluded. Because this is one of the main issues with elites, for example, in France. We don't reach a certain population. And it's really important, because now the majority of the surveys are done online, because it costs four to six times less.
They are very useful for some things, but it's a catastrophe for others- And I think when you compare modes, I think it's always important to, when you think about, for example, web mode, to take a step back and think about how were respondents recruited.

Nonna Mayer
But it's the access panels. That's the big problem. When you work with web data, first thing I would look at, how was the recruitment process?
And it should never be online, it should never be an access panel.
But the commercial surveys, they are mostly online and based on access panels.
Some institutes have one million people in their access panel. And they claim to be able to draw representative samples. But still, they're not the same people than in real life. The blue collar who is on the access panel, who accepts to answer survey questions three or four times a month, is not the ordinary blue collar. No more than the little old lady online is representative. Last there are socialization effects of the access panels. A journalist of Le Monde, Luc Bronner, played the game, was recruited, and he could answer something like 20 or 30 surveys a month[14].You necessarily change, you are a different person after that.

George Marcus
It is another dimension that there are almost no follow-up studies. The best one I know of, I suspect no one knows about this particular study, but in the late 1940s, there was a social development program to help young males, just going to school to get help so they could have a better life. And the area, Cambridge and Somerville, is the area right

[14] Luc Bronner, « Dans la fabrique opaque des sondages », *Le Monde* 6 November 2023, https://www.lemonde.fr/politique/article/2021/11/04/dans-la-fabrique-opaque-des-sondages_6100879_823448.html

around Harvard University.

But those are working class families, they're the people whose parents might work on the docks, they might work staffing cleaning services at a hospital or whatever. So, a group, a couple of Harvard social scientists said, let's study this to see if it really works. And they identified 200 people who were eligible, young boys, and randomly assigned them to an acceptance program.

And at the end of it, they did comparable assessments, and at the end of the journey, years later, there's the study itself. Then a scholar came back to it and looked at every public record of these people things like mental issues, hospitalization, criminal violence, continued violence, and so forth, about a dozen years later. And the idea, was to see what happened between those who participated in the study and those who didn't. Well, I'm telling you the story, because you should be able to guess this. Those who participated in the program had worse luck than those who didn't. No one will talk about that story.[15]

Nonna Mayer
You're not going to explain that?

George Marcus
Oh, that's a good example of what's called scientific doubt, or motivated reasoning. Because if you're a liberal progressive, it takes a village to raise a child. And here, it's sort of a community of resources to help children, which are better kind of treated. So if it is true, it has dramatic consequences. It means that almost every social program for young people, is having malevolent effects. So, well, a number of things. We can speculate, but my speculation would be, these children were set up with high expectations of themselves, because they were getting all this debunked and resourced. Some can't. Counseling for their families, if they're stressing their families. Academic counseling, so if you notice, they're being more aggressive than they are in

[15] McCord, J. (1978). A Thirty-year Follow-up Study of Treatment Effects. *American Psychologist*, *33*(3), 284–289.

Powers, E., & Witmer, H. L. (1951). *An experiment in the prevention of delinquency: the Cambridge-Somerville Youth Study*. Columbia University Press.

real world, and then all of a sudden, it's better. That’s sad, for public policies and persons that do. You know what I mean.

My mentor in grad school was a doctor, Campbell D.T, he has a famous article called “Reforms as Experiments” .So everyone sells a policy change as a reform, which is a word that carries with it the presumption that things will be better. We're going to reform our policy about poverty, which means we're not going to use it anymore. Drugs won't work, etc. And what he says is, if we think about reforms as experiments, how do we tell them that? He developed a whole science of policy analysis, basically. About what kind of data do you collect, and how do you analyze it, to sort of get a sense of where you're going in the region.
So, social sciences are not going to help with that, by and large. No one's going to do a career saying, I'm going to do this, and we'll come back in 30 years and see how things are.
Except if you're in France today, the word “reform” is not very well connoted. Macron has made it very different. He wants to add something to what we should do in the future.

From the audience
Let me ask a question, because you would know, right? Is there a public site to cover things like the instruction set? So, question design, those kinds of things? Like recommendations, or guidelines? Literature that supports?

Katherina Meitinger
You can use an instruction set, but that's a way to measure bias if you want to study that, but it's not a best practice if you're looking for actually getting a comprehensive response.
So, as a first starter, I would recommend the cross-cultural survey guidelines[16] and there you have a lot of references, and they are being updated also right now. So, here at Sciences Po, when you teach methodology in this lab, do you make this available and have the sources available?I'm afraid they don't, and it's a pity. That's exactly what Noemi was saying in the beginninWe don't, we don't, we don't, people are afraid of

[16] https://ccsg.isr.umich.edu/

these things.

George Marcus

So, to give another example of the kinds of things that I'm talking about, to make some sense of it, questions that ask," have you ever in your life felt angry in a way?" The research has been done, and you know how people answer that question, they use their current feeling, you're not getting what they felt when their parents got divorced, you're getting it back when your parents got divorced.

So, that's fine if you want that, but if you don't know that, if you think you're stating something about what your past was, same thing about the future of masturbation, how, do you think of war in your brain? It depends on how you frame it, if it's a thinking structure, you'll get a different response if you react to that.

Also, to make it more complex, this kind of culture differ in temporal perception, like whether I think more in our days, and I think, can easily think about future terms, so like they also perform differently across cultures. Yes, the very idea of projecting yourself in the future is very different from one culture to another.

Nonna

It makes me think of the debate that you had at the beginning of the ESS survey, because they had the questions of, there are countries where you don't use the term of God, so how did you do it?

Duarte Amaro (master Sciences Po)

I'm Duarte, I'm a second year master's student, and first of all, just seconding the point that we're not taught any of this, that I learned about the issues with translation in a summer school, and there we had a guest lecturer, it's Marta Kolczynska, who's at the University of Warsaw, and I think she's written a lot on this, and she discussed the issues with the WES, and how they translated stuff like military government in so many different ways, that any cross-national assessment of people's support for dictatorship was basically null, and I really wish that we would have learned this in this program here.

About the nationalism aspect, I found that really interesting, and I was thinking back to Sartori, and I was also thinking about European integration, which is my focus, because during the Euro crisis, in Southern Europe we saw a much bigger shift in support for European integration than we saw in the North of Europe. This doesn't necessarily mean that people's attitudes are more fluid, it just meant that they saw the European Union in much more instrumental economic utilitarian terms, rather than cultural identitarian ones.

But the question was phrased in terms of support of European integration, so a much higher level of abstraction, which meant that, because the predominant understanding differed within the region, this is still within Europe, where we would assume that cultural context is closer, it still meant that people understood the question differently. So I was wondering about the point about nationalism and being a patriot in the U.S. and being a patriot in Germany. Rather than the question itself being phrased poorly, could it just be that we're asking too many questions, being lazy and just asking one, whereas really we should be asking several? “Are you proud of your country's history?” And then maybe Germans would answer not so positively, whereas “do you enjoy living in Germany” might capture some of what we actually mean by it.
Wondering if it's not just about translation or one particular word, but just asking more detailed, fine-grained questions, which goes back to the question of just get your own data, which you can tailor to your own uses. Actually, the ISSP has a whole module on national identity, and they have more specific national pride measures. But then the question is, do they always capture the construct that you want to measure? Also then there's the whole issue of like, is it exhaustive enough? For example, there's not pride in cuisine, and it's like that in France, also in Mexico. By the way, the things about the EU, it's not only that people are lazy, it's just that they don't want to measure certain concepts.

When we talk about the EU, you have clearly a different discussion that goes on at the cultural level, a different discussion that goes on at the economic level. Personally, I'm very critical of how the EU works. And when I discuss with my colleagues, you always have these feelings that the cultural values of the EU need to be put forward. So if the data you use comes from the Eurobarometer or whatever, one needs to be aware. And so this goes back to the point of how you collect your own data, whatever. But already

the fact of having this knowledge, I think, is very interesting in how you interpret the data. What you say is actually a very salient thing. But it's not only laziness. I mean, sometimes it's mistakes. We can take it to the positive side. But sometimes it's a substantive issue of how we look at things and what is the predominant view on things.

We just don't look at the same thing. Do you want to switch from, “do you want your country to be in the EU?”. They dropped that question, right? And they replaced it with trust, which is not the same thing.

Question from the audience

Do you think the social sciences at St. Petersburg have a course on methodology of crime? All of them. The problem is, it depends what you put in methodology. I don't think the jurists do methodology. But they think they do. There's methodology in law. And all the schools have methodology.

In methods courses, it starts with research design. Then it goes a lot more into just the basics of OLS and extensions of that. And it's much more focused on “how do we implement that?

Rather than “how do we think about what they're doing?”

It goes beyond the issue of survey data. It is about thinking about concepts. And so many times, we take concepts at face value. Because there is an interest around these concepts. And we continue to manage them. Especially social capital. This is one of the reasons why I stopped doing work on social capital.