Q: And...then we can start with the interview. (laughs) Okay, so, the first question would elaborate a little bit on your perspective on secondary data use as a data user, so to say. And the question would be how often you reused secondary data in the past. So, from your lab and also from other labs. And I would like to ask you to quantify your specification just by providing the relative frequency for reusing data compared to producing primary data. #00:00:40-7#

R: Yeah. Mm (thinks). So the kind of research I'm doing is methodological research, so I propose or I develop techniques for doing data analysis. So to evaluate the techniques that I propose, we often use simulation studies. #00:01:06-8#

Q: Okay. #00:01:05-7#

R: So we simulate our own data. #00:01:10-8#

Q: Okay. #00:01:08-2#

R: So we are not reusing data for that purpose, so each time we do a new simulation study. But at the same time, we also try to illustrate the techniques and therefore, yeah, well, one of the major research topics of mine is meta-analysis, so to illustrate the techniques, yeah, we use meta-analytic datasets that we find in the literature or we collect, yeah, we look in the literature for effect sizes or some - to do a meta-analysis ourself. #00:01:56-1#

Q: Ah, okay. #00:01:55-1#

R: That's (...) very often, we use existing data. #00:02:01-6#

Q: Okay. Mm (thinks). (...) would have to mention certain percentage (...) #00:02:13-4#

R: Yes (thinks). Yeah. But I'm also involved in some more applied projects. For that projects, yeah, there are different kinds of projects as well. #00:02:30-0#

Q: Yeah, okay (laughs). (...) #00:02:35-0#

R: (laughs) Yeah, for instance, I...I'm involved in a project in which we use data from the (TIMSS) and PISA studies, so that our international large-scale studies...therefore, we reanalyze data. In other studies, we collect data ourselves. So, for instance, I'm involved in a study with a whole research group in which we collect data over the years, so there are (two waves or even) four waves, yearly waves, but in that project, yeah, a bunch of PhD students are using these data. #00:03:18-6#

Q: Ah, okay. #00:03:20-6#

R: (So it's the same -) it's one large dataset that is collected by a lot of people but many people, I think, yeah, there will be four or five PhD students working on the same datasets. So it's not really a reuse of the dataset but...yeah, it's the use of the same dataset by several people. #00:03:41-7#

Q: Yeah, okay. #00:03:41-5#

R: And, yeah, I'm also involved in other (applied) projects in which we collect data. (ourselfs, only ones), so one person is collecting data for a specific study. These data are not reused. So but you asked for a percentage. (thinks) Yeah, so, if you include the simulation studies, mmh, so in which we (generate) ourself the data, so that's not the reuse of data, that's the data that we make ourselves, we simulate ourselves. #00:04:21-6#

Q: (...) #00:04:23-5#

R: Sorry? #00:04:25-7#

Q: That would be primary data also (...)? #00:04:29-4#

R: Yes, yeah, I would call these primary data, yes. So then I would say (...) about 80% of the studies we do, we use own data. #00:04:41-5#

Q: Okay. And resue of (…) data? (...)? #00:04:50-6#

R: Sorry? #00:04:52-6#

Q: Reuse of other datasets or from other labs, so for instance, for your illustrations...? #00:04:57-7#

R: Yes, so therefore we use... #00:05:01-7#

Q: That would be less or...? #00:05:02-0#

R: (...) a lot of meta-analysis and, and in such a meta-analysis, we look in the literature for primary studies (...) topic we are interested in. And we look in that report (in) primary study. So, for instance, in the research article (we look) for, yeah, summary statistics, for effect size or test statistics that we can use to calculate an effect size. And then in that meta-analysis we combine all these effect sizes over studies. So in a sense, we are collecting data, we are making a new dataset, the dataset consist of effect sizes that are calculated on data collected by other people. #00:05:55-8#

Q: Ah, okay (laughs). #00:05:57-4#

R: Yeah. #00:05:59-6#

Q: Okay. And how often do you do that approximately? So if you would take a timeframe of one year, for instance? #00:06:10-1#

R: Mm (thinks)... #00:06:13-8#

Q: I'm not sure how long it takes to do such a thing. (laughs) #00:06:15-4#

R: Yes, it can take a lot of time. Yeah, I would say, between 3 months and half a year. Well, one meta-analysis, and if you...if I look at my research group or projects in which I am involved, I think, yeah, about five...we do five meta-analyses each year. #00:06:45-4#

Q: Okay. Yeah, I just need some quantification just to estimate how important the theme is in the field, right, that is why I'm always asking for numbers (laughs). #00:06:54-0#

R: Okay. #00:06:56-5#

Q: Okay, so the second question is on the purposes for which you use secondary data. We have already talked about this for the last minute. So that is why I would just be interested in what specific additional information on the data, so meta-data, you would need to optimize your work. #00:07:19-7#

R: (thinks) Mm, yeah so, for the meta-analysis, so if we are combining the results of several studies, what we need is at least one effect size for each study, so it can be correlation coefficients or it can be a standardized mean difference for instance, but we also need information on, yeah, on how precise the effect size is. So, for instance, a confidence interval or (...) standard or... #00:07:59-5#

Q: Yeah. #00:08:01-0#

R: That's what we need minimally. But, yeah, in the meta-analysis, we also try to...to take into account the quality of the study, so we also look for information on the quality, so for instance, yeah, what the reliability of the measurement is or...yeah, how (...) what the external validity is or the internal validity of the study. And what we also need or what we are also looking for is all kinds of characteristics of the study that can have an impact on the size of the effect. So now, to give an example, if we are interested in the effect of a specific intervention, that intervention has been studied by, say, thirty studies, probably the intervention was not implemented in exactly the same way in all these thirty studies. So maybe in some studies the intervention took somewhat longer or (several?) instructions were given or sometimes the population is not completely comparable over studies. So we need information on all data, on as many characteristics as possible... to loof afterwards afterwards when we see that the effect that was observed in the different studies, if that effect varies over studies, that we can try to explain why that is. So is the size of the effect related to characteristics of these studies. #00:09:44-1#

Q: Okay. So you need a lot of the procedural information, right? #00:09:49-3#

R: Yes. Yeah, on the procedure but also on the population from which it was (...), yeah, and information, yeah, on the intervention. And that's often a problem. Very often, that are not very transparent about how exactly the study was done, how exactly they sampled participants, how the population looked like, how exactly variables were measured, and so on. #00:10:27-2#

Q: Do you think this is surprising given that we have something like the JARS, so the Journal Article Reporting Standards from the APA? Because there everything is standardized, and, yeah...normally, we would expect that...that researchers write everything. #00:10:47-4#

R: Yes. We would expect that but that's unfortunately not the case. So there is some improvement, so for instance, yeah, if you look at old studies, it's...yeah, there is sometimes you simply say/see sometimes in the best case, you get an effect size but (it is said it is statistically) not significant without peer review or without test statistics. Where you do not know how precise the estimate was exactly. So nowadays, very often you will find information on effect size and on the precision of the effect size but information on how the study exactly was done is very often lacking. #00:11:32-8#

Q: Mhm (agreeing). And would you prefer that this information is provided within the article or in an additional file which is provided together with the data? #00:11:48-1#

R: I think that is the problem or a problem. A journal article, yeah, very often (...) have a work limit and, yeah, you simply cannot be very transparent on your study within that (work,) you cannot give all required information. So I think that's a good tendency nowadays that you can provide additional information in supplementary documents. So for meta-analysis, (yeah as long as) we can find that information, that's great. (laughs) (...) in the article itself, if it is available somewhere else, that's great. What we also sometimes do if we do not have enough information, we sometimes contact authors. #00:12:47-0#

Q: Yeah. Yeah. #00:12:48-9#

R: To ask, yeah, for instance, if...yeah, if you have an effect size but we do not have information about the exact (b/p value(a)s??) or if you do not know, yeah, some characteristics of the study, we contact the authors but in our...in our experience, very often authors do not answer. But also often they answer but they answer that they do not have the time or that they do not find the information anymore or that...yeah. #00:13:23-3#

Q: Ah, okay. So they did not document their research very well? (laughs) #00:13:32-4#

R: No. #00:13:30-7#

Q: Okay. Yeah, I know about these problems. (laughs) #00:13:36-8#

R: Yes. (laughs) #00:13:36-5#

Q: I also had them in the past. Okay. Are there other methods you know but haven't used on your own, so regarding secondary data use, which would require other metadata than the ones you already mentioned? #00:13:56-3#

R: Erm (reflects). Yeah, that's not very clear to me what you mean. #00:14:00-3#

Q: Yeah, you already mentioned that you are doing meta-analysis and also some kind of re-analysis and illustration. #00:14:09-7#

R: Yes. #00:14:09-7#

Q: Are there (above that -) beyond that other...also other methods which would require secondary data use but are not part of a meta-analysis or (...) that you already mentioned. #00:14:22-3#

R: Mhm, okay. Yeah, so I...I already mentioned that...that in a project we are re-analyzing data from PISA, TIMSS. (So that are,) yeah, they collect huge datasets with a lot of variables. And #00:14:49-1#

Q: The TIMSS is what? #00:14:50 –#

R: Sorry? #00:14:50 -1#

Q: TIMSS dataset, I don't know that. #00:14:57-5#

R: Okay, I...so that are...TIMSS and PISA are international studies. #00:15:00-3#

Q: Yeah, PISA I know. Yeah. #00:15:06-6#

R: Yeah, okay, yeah. TIMSS is very similar. Yeah. #00:15:11-0#

Q: Ah. Okay, good. #00:15:13-8#

R: Yeah. (…) in/with these studies, yeah, a lot of variables are collected. A lot of information is collected and...and, yeah, so (thinks..) sometimes we are interested in questions that are not yet studied with these datasets, so there much more information in these datasets than...than was retrieved from the datasets. #00:15:42-1#

Q: Yeah, so (laughs). #00:15:45-7#

R: Yeah. #00:15:44-5#

Q: Yeah. #00:15:46-6#

R: So, therefore we sometimes use these datasets to...to study further research questions. #00:15:51-8#

Q: Ah, okay. Mhm (agreeing). And these datasets are sufficient for...for doing this research? So, doing this further research (...) I have heard from another researcher that PISA has a really extensive codebook and... #00:16:19-7#

R: Yes. #00:16:16-6#

Q: ...that he really likes it. Would you say this is sufficient to do your research or is it even too much? (lauughs) #00:16:25-1#

R: Well, of course it...it has limitations. So, sometimes you want to use variables that are measured in, yeah, not really measured in a very detailed way, so that are measured, for instance, by one single question with categories as as possible responses. So such a question does not give a lot of information. #00:16:56-7#

Q: True. #00:16:55-2#

R: Though, if you would have set up a study ourselves, to answer that question, to answer...yeah, a research question, you would have measured that variable in a more detailed way. #00:17:09-5#

Q: Mhm (agreeing). #00:17:11-7#

R: So using several items and, yeah, not only categorical answers but also quantitative answers, so... sometimes, we are limited by the way data were collected. #00:17:24-3#

Q: Yeah, mhm (agreeing). #00:17:26-2#

R: Or sometimes...yeah, so now, yeah, recently, we were comparing different types of schools in *[area 1*], so in the *[area 1]* region of *[country 1]*... #00:17:42-2#

Q: Yeah. #00:17:44-1#

R: And, yeah, using TIMSS data and for some of the...well, we were comparing three types, and for one of these three types, only a few schools were included in the study. #00:17:57-8#

Q: Oh. Okay. #00:17:56-2#

R: In fact, the number of schools was too limited to say a lot about that specific type. #00:18:04-9#

Q: Yeah, okay. #00:18:05-7#

R: But it's true that the dataset is...the datasets are very well documented. That's very helpful and...yeah, of course it's (thinks)...it's...it's logical that if you want to study questions that were not the initital purpose of the TIMSS study or the PISA study, then it's logical that the data were also not collected from the perspective of these research questions, so...yeah. #00:18:36-9#

Q: Yeah. True. Mhm (agreeing). Okay. Last question. So, from your perspective as a data user (laughs), would be which kind of data are you using generally? So you already mentioned PISA and TIMSS, these are mainly questionnaire data, I think, and behavioral data, largely. Are there other data types? So, for instance, genetic data or physiological data? #00:19:14-3#

R: Well, I'm also...yeah, another research line of mine is e-learning, so do you...when, when students or persons are making use of digital learning environments that generate a lot of data and so I'm also developing techniques and models to analyze these data, to say something about these students but also to say something about these learning environments, and to try to personalize the learning environments and optimize the learning environments. #00:19:58-2#

Q: Oh, okay. #00:20:00-1#

R: So there's a lot of kind of data, so learning data. So, for instance, if people are making use of MOOCs...you know MOOCs? #00:20:05-5#

Q: Yeah, mhm. #00:20:07-1#

R: Massive Open Online Courses. #00:20:12-3#

Q: Yeah. #00:20:10-5#

R: Very often, they...yeah, they click on things or they answer questions. The old data are tracked, are logged. We are looking at how we can analyze these data. #00:20:21-5#

Q: Mhm. But these are also mainly behavioral data, right? #00:20:32-5#

R: Yes. Yes, that are behavioral data indeed. #00:20:36-5#

Q: Mhm. Okay. And (thinks)...do you perceive any differences between the different kind of datas you are using? So also within the class of behavioral data, for instance? Are there any differences in documentation quality? #00:20:58-0#

R: Mm (thinks). (...) these data, yeah, that are...that are data that we...that we often collect ourselves, so that's (...) #00:21:15-3#

Q: Yeah, if you just refer to the data that you reuse. I think that could be also behavioral data. We already talked about the PISA, which is very well documented in...in your eyes. And are there other behavioral data which you use but that are not that well documented? #00:21:34-7#

R: (thinks) No. So (...) I said that I am doing a lot of meta-analysis … in different domains. #00:21:46-1#

Q: Yeah, okay. #00:21:47-6#

R: And...so, not only...not only in the domain of behavioral sciences but also in the bio-medical sciences. And I do not really see a difference there. #00:21:58-2#

Q: Okay. #00:21:59-9#

R: I think it's a common problem that (laughs) sometimes it's hard to...to find the necessary information. #00:22:03-3#

Q: Yeah, true. Good. Then we switch to the area of data sharing. And, yeah, I would be interested in what sorts of metadata do you generally provide about a dataset when you upload to a reposity? I'm not sure whether you upload datasets on you own (laughs). If not, then we can just skip this question. #00:22:36-5#

R: Yeah. No, I have to say, no. Maybe we could try to, of course, to document our data for (...) people ask afterwards how did you do this or you want to re-analyze datasets. Even with (...), it's rather for our own use, so it's documented for our own, so that afterwards (...) #00:23:03-3#

Q: Oh sorry, now the (...) #00:23:08-2#

R: We can remember or we can retrieve (...) Sorry? #00:23:12-4#

Q: Can you just repeat the last two sentences? The connection was a little bit...broken. #00:23:19-9#

R: Yes, okay. So...we tried to document data that we collected. #00:23:27-9#

Q: Yeah. #00:23:29-1#

R: But not...yeah, for instance, to share with others. It's rather for our own. #00:23:36-3#

Q: Yeah. #00:23:37-6#

R: So that afterwards, if you want to look back at our data, we still...or if...if we get a question from somebody, how did you come to that result, that we can re-analyze the data or that we can reconstruct what we did. But even then, I have to admit that sometimes...that...that we or that I, that we are sometimes sloppy. #00:24:01-9#

Q: Ah, okay. #00:24:01-3#

R: So sometimes, yeah, so sometimes I get a question. So, for instance, or I want to look back because I have another...I have to do another analysis and I want to look back, how did I...did this similar analysis five years ago, and sometimes, I have to admit that it is hard for me to reconstrct what exactly is the last dataset I used and how exactly did I do the analysis. So I'm...myself, I'm also sometimes sloppy in...in (thinks), yeah, in describing a dataset and in describing the analysis that I have done. #00:24:55-4#

Q: Mhm (agrees). And it is because it is too time-consuming or...what's the reason? #00:25:02-0#

R: I think that's the reason, yes, indeed. Yeah. #00:25:04-5#

Q: Mhm. #00:25:06-8#

R: And also we are working on the dataset and you are doing the analysis and then it is published and then you think: Okay, now (laughs) I'm going to the next study. #00:25:15-9#

Q: Yeah, true. #00:25:13-4#

R: But then it's time to...to write out exactly what you did and then (what the) final analysis were. #00:25:23-0#

Q: Yeah. Would you presume that researchers are like you? But also many others would spend more time on documentation if it would be rewarded more? So, for instance, by funding agencies or by the journals. #00:25:49-5#

R: (thinks) Yes. I think that...that would be a good idea. So I Iearned that some journals now also explicitly ask to make datasets available (...), maybe not publicly available but at least that they can see the dataset. So I think that that would be a good solution, so that...that also journals and especially funding agencies, that they require that datasets are available and well documented. #00:26:18-5#

Q: Yeah. Yeah, for these reasons, we would need a standard, right? (laughs) #00:26:23-0#

R: Yeah. #00:26:22-1#

Q: So that researchers know how to do it best. Yeah. Okay. Mm (thinks). So I think the next question we can skip because you already answered it. That your documentation is oftentimes not sufficient (laughs) for re-using your datasets. Have you ever used certain metadata standards for annotating you data? Or do you know about these standards? #00:26:54-2#

R: Mm (considers), no. No, we never used these standards. #00:27:04-6#

Q: Okay. Then the last question would be: In a perfect world (laughs): If you would have to create such a standard, which information should be included in such a standard? Perhaps you can think about this question in terms of the categories mentioned in the JARS. So the categories often used or always used in a research article. #00:27:32-2#

R: (considers) Mm, so you mean for... #00:27:35-1#

Q: For data documentation. #00:27:38-9#

R: Metadata (...). Yeah, well, a clear description of how the dataset was collected. So, for instance, how participants were sampled. (thinks) So, yeah, the procedures. Also whether there were missing data, how missing data were handled, and therefore...yeah, a description of each of the variables. Very often, variables are also transformed, so for instance, sometimes variables are centered or standardized, or different items are used to construct one new variable. So an illustration, a documentation of what transformations were done. I think it would be intersting to have the original data. #00:28:39-4#

Q: So raw data, yeah. #00:28:39-4#

R: Yeah, the raw data as well...as the transformed variables. I think it would also be intersting to have the codes that were used or that you have the syntax that was used. #00:28:56-4#

Q: Yeah. #00:28:55-4#

R: To analyze these datasets. #00:28:58-9#

Q: Mhm. For...for analyzing and also for data preparations or...for instance... #00:29:05-2#

R: Yes. Yeah. Indeed. Mm (considers). Yeah, I think that's it. #00:29:17-0#

Q: Okay. Good. Then I thank you very much (laughs). #00:29:21-0#

R: You are welcomoe. Right. #00:29:22-8#

Q: And do you have any further questions, concerns, ideas? #00:29:32-5#

R: I don't. #00:29:31-4#

Q: Okay. Wonderful. Then I wish you a nice weekend. #00:29:32-2#

R: Okay. Thank you, you too. Good luck with your study. #00:29:37-8#

Q: Yeah, thank you very much. And perhaps we see you in *[city 1]* (laughs) next time. #00:29:39-8#

R: Ah, yeah, yeah. Yeah, that would be nice. #00:29:42-5#

Q: Okay. #00:29:45-9#

R: Okay. #00:29:47-0#

Q: Bye! #00:29:45-5#

R: Bye, *[name of the interviewer]*! #00:29:48-7#