Legend

1. [REDACTED] means that the original word/fragment was deleted to ensure the anonymity of the participants.
2. [?] is a placeholder for words/fragments that could not be transcribed.
3. (?) means that the transcriber was not completely sure what the last word/fragment was, but had a guess.
4. Sentences that begin with "I:" were said by the interviewer
5. Sentences that begin with "P:" were said by the participant

Block 1: General Information

I: So, now we will start with the first block. The goal of this block is to get some general information about you. And the first question is: Are you a PhD Student?

P: No.

I: Ok. Then, how many years has it been since you got your PhD?

P: 8.

I: 8 years, ok. And what is your field within psychology? With field, we mean for instance social psychology or cognitive psychology?

P: Human factors.

I: Human factors, ok. And did you conduct any experiments including a Stroop task in your career?

P: Yes, I did. And I – oh, I did the Stroop task when I studied myself, so that was sort of a, what is it called, experimental practical. And I did a Stroop task study [REDACTED].

I: Ok.

P: You wanna hear more about that.

I: Yeah.

P: Now or later?

I: No, no, no, you can ...

P: [REDACTED].

I: Ok, thanks. I am still a bit sick, therefore I have to drink something in between. Is it still running? Yes. Ok. And which statistical analysis programs do you use at least once a week? Multiple answers are possible. For instance, SPSS, R, Stata, SAS, Matlab, Python, or any other?

P: For data analyses, I solely use - exclusively use R.

I: Ok. And how would you rate your knowledge of statistics relative to your peers on a scale from 1, extremely poor, to 10, excellent?

P: A 9.

I: 9, ok. And how confident are you that your fabricated data will go undetected as fabricated? Again on a scale from 1 to 10, where 1 means extremely insecure and 10 means extremely confident.

P: Well, I don't know what you do. I mean it is like you could – if you compare the data to the real data of a Stroop task, then it is probably lower. But I still think I did a good job here. So, I think it is an 8.

Block 2: Timeline of Data Fabrication Process (When?)

I: Ok. Then this is the end of the first block about general information. Now, we will start with the second block. The goal of this block is to get some information about the timeline of the data fabrication process. So, the first question is: Did you fabricate the data in one day or spread the data fabrication over several days?

P: I fabricated – actually I had that already, I already had fabricated Stroop data for education purposes, of course. So, it was only adapting that program, that function, and that took me 1 or 2 hours.

I: Ok and you did the adaption of the already existing data set on 1 day, right?

P: Yes. I think so.

I: Ok. And how much effort – or no, let me ask first: How long did it take you create the fabricated data set that you used here for education purposes when you created that one?

P: It was the first one I created, so - and I did a lot of modifications, so I improved it continuously. So, actually what you get from me is even a ripped down version of what I have. It is very difficult to say. So if you take the whole process of – well, actually it was the process of learning how to fabricate data in an efficient way – so then you will be in the range of days maybe. But if I would do it now for any other experiment, it would take me probably less than an hour.

I: Ok. And how much effort do you feel you investigated in fabricating the data on a scale from 1 (no effort at all) to 7 (a lot of effort)?

P: 3.

I: Ok. Did you prepare in any way before starting to fabricate the data?

P: Well, in a way yes. So not for this request here [REDACTED] I have like 15 simulated data sets with all sorts of scenarios and conditions. So if that is the preparation – yeah, that is the preparation, I guess.

I: Yeah.

P: So I was actually like – yeah, I got this request and I practically have that here already, almost ready to submit.

I: Yeah. Could you estimate the time or how much time you spent on this book that was sort of a preparation for this task?

P: Well, weeks. But I – I really cannot separate that. I mean, you are writing and you think of a scenario and then you think, yeah, that scenario is good to, I don't know, for example, explain interaction effects or explain random effects. And then you start writing a simulation function and then you refine the simulation function and you use it again. So, that is impossible to say.

I: Ok, and did you read literature on detecting data fabrication?

P: Well, I read the Stapel report, but – and well, I am very interested in this transparency movement, but it is not like I have read something on how to do it or I don't know maybe there is an inquiry somewhere how are people doing it.

I: Ok. So in the Stapel report or like in other things that you have read can you describe or name any methods you read about?

P: Well, as far as I remember, he did it – he just filled out the questionnaire. I can't remember that he was using some sort of random simulation or model to fabricate the data. So, quite amateur-ish if you ask me. And well, that's it. I mean, in the Stapel – well, I heard about another case that was a study on . . . [REDACTED]. It was one of these notorious embodied cognition studies, but that is about it. So, I am more a bit of an autodidact, I would say.

I: Ok, but could you describe in a bit more detail what you learned throughout this [REDACTED] about . . . ?

P: Data fabrication?

I: Like about the methods or like . . .

P: Oh no, that was not of course not disclosed. So, they just discovered that the same author has done the experiment three times and that surprisingly it was in practical almost – precisely the same effect. So, what I learned – or what I got sort of how do you call that approved is that if you fabricate data it is probably - doing it well is probably about making it imperfect.

I: Ok, thank you. Ok, would you say that you prepared in any other ways or do some things come to your mind that you think influenced your approach to fabricating data?

P: Oh yeah, many things. Yeah, so let me check. The first thing, of course, is knowing advanced statistical models. So, I guess that if someone would do it very naively, this person would probably forget individual differences, that people differ in having an intercept, random effects on how quickly they respond. So if you know how to – if you know what a mixed effects model is, then you can also sort of – then you know how to set up a simulation for the data generating process. So, that is the one part: understanding sort of the – understanding the structure. And the second is knowing about the distributions for response times. So, I guess, someone trying doing it naively would model the response times as a normal distribution which, well, is strictly not – cannot be. So, most response time distributions are skewed. And so, I know - as a statistician, I know a lot of about statistical distributions, how they look like, where they arise, these sorts of things. And the third is, yeah, of course, R programming and in particular data modelling.

I: Ok, thank you. Then this is the end of the second block. Do you have any other comments about the timeline of the data fabrication process that come to your mind and you think could be interesting for us to know?

P: About the timeline?

I: Yeah.

P: No, I don't think so.

Block 3: Broad Framework of Data Fabrication Process (What?)

I: Ok, thank you. Then, we will start now with the third block. The goal of this block is to get some information about the broad framework of the data fabrication process. So, the first question is: Could you name specific characteristics that would make data look more fabricated in your opinion?

P: Yes, any signs of perfection. So that – I mean to be honest I was a bit surprised of the format you are asking me or asking us to hand over the data, because what I did is I modelled the data on the response level. And so in fabricated data I would expect a certain – yeah too much perfection. So, for example, all participants being on the same level of reaction time that would be one thing. A second thing would be no outliers. So, I mean, in these experiments it just happens some times that people are distracted and then you suddenly have a response time of I don't know 1.5 seconds, although the normal range is more below 500 milliseconds. Missing data, of course. Or the sequences in – I mean what you also don't have here is success or failure. So, I would actually – I didn't model it here, but if you would have asked me to hand over data on the individual observation level, then I would have modeled in some accuracy-speed-tradeoff, right. That people who would do it very quick they tend to have more false responses. The third thing is that – well many researchers still think that the normal distribution is the normal thing to happen which as I said is absolutely not the case for reaction times. So, I would – I would check for is the error term – is it – are the residuals perfectly normally distributed. And I would not check that on the aggregated level, but I would check that on the participant level. And then I would – if I would see a perfect Bell curve, I would assume that, yeah, someone was using a random number generator or

– however, I think we know that people – so, I mean what is more tricky is I think if someone does it without a simu-, without a program. So if someone tries to be their own random number generator. So the only thing I know is that people are particularly bad at it. But I am not entirely sure how to check it. I mean because I don't know precisely what the biases are. Yes, a few I know. It's like if you have – if someone does it – so tries to generate random sequence of random numbers, then I think there - then often too large jumps happen. Because people think like, well, I here entered 230 milliseconds, it is really not likely that the next response is almost in the same, so I make a big jump to make it more – to make it look more random. So, I would look for that. How often do very similar values appear together and is that something you would expect in the real process. Then I would look for other features that people might forget. For example, again on the level of individual responses, I would model in or I would check for, depending on what side of the table I am sitting, for learning and fatigue effects. So is there sort of a trend over time? In a real experiment, you would always expect that or you would see something like that. Or (?) in the end, the errors get more often because people get tired. Things like that I would look for. And then, yeah, finally, the, yeah, the distribution I named already, outliers, distribution, individual differences, also individual differences not just in the intercept, so what you would naively do is like – I mean halfway naively – is set a sort of an intercept random effect, so saying people just differ in their overall speed, but I would also – if I would do it very consciously, very precise – I would also let the difference between the two conditions vary between participants and then you might go into further things like, so, for example, I would also expect, maybe expect some interaction effects between participants, so some – no sorry, some correlations between intercepts, random effects, because if you have participants who are doing it very slow in the congruent condition already, it is less like that they get a lot slower in the incongruent condition. Because if they are slow it means they are – well, let's, if we take for example the Kahneman System 1 System 2 model if they are very slow that indicates that they are doing it very consciously with a lot of attention, and then - well then of course the difference to the incongruent condition is less pronounced. So, I would look or model these sorts of things.

I: Ok, thank you.

P: May I ask the question why you decided to ask for aggregated data because all these things are no longer visible? Maybe after the interview.

I: Yeah. That is a good suggestion, I think. But you are happy to ask – you are welcome to ask all of these questions that you have after the interview.

P: Ok.

I: And could you name specific characteristics in addition to those you already named that would make data more genuine in your opinion?

P: I think, I named a lot of things, right. So, cannot think of any further . . . Well maybe yes. There is this - I am not sure about it. But it is more from - it comes more from introspection, of sometimes also being participant or testing these experiments that if someone is - if the participant had an error, so pressed the wrong button, then I think there is sort of a startling moment and there is sort of a carry-over to the next trial. So, this is just an assumption, one had to check that first on real data, I guess.

I: Ok, thank you. The next question is: Did you take these characteristics you just mentioned into account when fabricating the data?

P: Let me see. So, I can tell you precisely what I took into account.

I: We may want to - so like I don't want you to be too far away from the microphone. So like ...

P: Maybe we place it here for the moment and then I can do both.

I: Yeah.

P: So, I modelled individual differences in speed, that I did. It is actually very simple what I did here. Yeah, I actually - so for this, because you asked me for aggregated data all this fancy stuff I just mentioned is actually not visible. So, I think the only thing you can do is - you can see is there are, well, too much regularity in the participant aggregated data and, well, they said to take care of - wait - well, the first thing is I, yeah, I did individual differences, so they differ in overall speed. And the second thing is I used the exponential Gaussian distribution for the response times so that is the residual. And that has two consequences. The first consequence is that the response times here are - have a left skew. And the second is that with longer - with overall longer response times, you also get more variance. So, I mean this is also what - that is, the residual-variance structure. Well (?), people coming from ANOVA they think that homoscedasticity is the norm, but if you think about the process this is not realistic. If times get longer, then typically also the errors get longer. And this is something this distribution does. So, if the mean increases, then the variance - the residual variance also increases. So, these are actually the two features I modelled into this data set. But, yeah, as I said this is more or less a ripped down version of a simulation I did earlier where I also did some time trends and incorrect responses and all kinds of other stuff, but I don't remember exactly what it was, I had to look at the program. But here you get, yeah, the individual differences and the skewed distribution and the mean-variance structure.

I: Ok. Maybe we can put it here again. Ok, so did you take into consideration relations in the data other than the Stroop effect itself?

P: Relations in the ... yes, you mean structures?

I: Yeah, ...

P: I mean there is only two variables here, right. So ...

I: So, for example, the distribution of the scores or other aspects that could be inspected with the data set.

P: Well, you can't inspect the skew of the data because you are asking for aggregated data. The skew of the data you only see on the individual observation level.

I: Ok.

P: I mean, you also don't see like - you were, I assume, yeah you asked for the standard deviation, which means that is one value. But if you have a skewed distribution, then for example if you would have asked for the confidence limit or the 95 quintile then you would get an asymmetric interval around the mean, but you don't see that here. So, the only thing, I guess, you see in my fabricated data is the individual differences.

I: Ok. And what criteria did you use to determine whether you thought your fabricated data would go undetected?

P: I went to the Wikipedia page on the Stroop experiment and they have a plot of the response times and so I did the same plot and then adjusted the parameters so that it would look like that. And in this plot, you also see the skew of the residuals.

I: Ok. Did you use different criteria for the means and standard deviations?

P: For the means, yes, that is the individual differences. So, I think I was lazy and just sampled that from a normal distribution. I mean that is something to argue about. Maybe it is also - maybe it is gamma distributed or something, I don't know. And yeah, you will see a relation between mean and standard deviation, so, when the mean increases, the variances are (?) also larger.

I: Ok. And now, in hindsight, are there things you think you should have paid specific attention to while fabricating the data?

P: Well, not on this level. But as I said earlier if you had asked me for individual responses, so on the - also the raw data actually, I would have - I would probably have done all this fancy stuff. Or as I said, I did that already most of it.

I: Ok, then this is the end of the third block. Do you have any other comments about the broad framework of the data fabrication process that you think could be interesting for us to know?

P: No.

Block 4: Specific Steps of Data Fabrication Process (How?)

I: Ok. Then, we will now start with the fourth block. The goal of this block is to get some information about the specific steps of the data fabrication process. Just checking whether it is still running - ok. Could you indicate what steps - what specific steps you took to fabricate the means for the participants?

P: Yes. To look at my program here and walking through it. Right. So, I programmed that as a function in R. So, a function that is in principle parametrized, so I can with this function produce now any Stroop data like data set I like, so with more participants, more stimuli, or more pronounced effects whatsoever. So, these parameters have been - I have tweaked to look like this plot on Wikipedia. So, what I typically do in these data simulation programs is I sample first from the participants. So, I - generally I assume that participants can vary on any factor in the experimental design. So here, that is 2, 2 parameters, namely the intercept or the congruent condition and the difference between the congruent and the incongruent condition. And yes, I sampled them - let me briefly check - yeah, so, what I do first is I create a table for the participants, where there are participant level effects, so that is - these are practically random effects, so deviations from the population level effect. Then, I create a table for the stimuli although that - here, I took that very easy because there is not a lot of variation in the Stroop stimuli. So if I would for example simulate data for a - for the semantic Stroop task [REDACTED], I guess there is a lot of variation between the pictures, so the stimuli, then I regard the stimuli actually as a population from which I draw. So, then you have some pictures which are very triggering for associations, and some do that less. But here, I assume they are all the same, so the words, the color-words (?). Well, and then I create a third table that merges the participant table and the - actually, technically, it is not merging, but joining (?) - so that is a data modelling procedure, so I am joining both tables into one table for observations. So, in this table, I then have the fixed effects for the stimuli and the random effects for, well, how participants vary in their response to the conditions and then, well, I use a linear model to calculate the expected value for every participant and, well, I then used a random number generator for the Exponential Gaussian distribution to create the noise, so to add the noise, and that's it. Well, and then of course, I am doing the aggregation steps, but that I think is trivial.

I: Ok. And could you indicate what steps you took to fabricate the standard deviations for the participants?

P: Well, the standard deviations they come from the noise part. So, actually, the residuals are exponential Gaussian distributed which is - I don't know if you know this distribution . . .

I: Yeah.

P: It is - so, it is skewed, but the difference to the gamma distribution which they use a lot in reliability engineering for example is that you have an offset, because with the gamma distribution you would get some participants or you would expect some participants to have response times close to 0, which is just not realistic. So, there is this offset. And then, yeah, I just computed the standard deviation per participant and per condition, that's it.

I: Ok. And did you repeatedly fabricate data until you were satisfied with the results?

P: Yeah, I wrote this simulation function parametrized, so I can modify the fixed effects, the standard deviation of the random effects, and also the parameters of the exponential Gaussian. And I - yeah, I wrote this function and then I tweaked it so that my plot would look very similar to the one found on Wikipedia from the original article of Stroop.

I: Ok. And so you said that you used the Wikepdia article to compare the images there, were there any others ways that you determined whether you were satisfied with the fabricated data or that they needed to be adjusted or . . . ?

P: More or less gut feeling, knowing approximately where these response times typically are.

I: Ok and I think, yeah, we covered . . . well, I will ask it anyway. Did you try to inspect whether the fabricated data looked weird?

P: Yeah, I wanted it to look weird, because I think it is too much perfection that makes data suspicious.

I: Or did you try to inspect whether the fabricated data looked fabricated in like more things that you have already mentioned so far?

P: Oh yeah, I also did these density plots on participant level to check whether this is - well, whether this for me looks like typical outcome of an experiment. But that is more, as I said, more gut feeling than - more experience than something I could pin down.

I: Ok. And did you try to inspect whether the fabricated data looked genuine?

P: Yeah, with the Wikipedia article.

I: Ok, and how many different mean-sd combinations did you fabricate before getting to the final fabricated dataset?

P: You mean like how often I ran this - I did this simulation function or . . . ?

I: Yeah like how often you changed the final set of the . . . ?

P: Parameters. I don't remember exactly. Maybe between 5 and 8 times, I guess.

I: Ok. And besides the supplied spreadsheet, did you use any other computer programs to fabricate the data?

P: Yes, a R program.

I: Did you use a random number generator to simulate data during this study?

P: Yes.

I: And did you use real data during the fabrication process?

P: No, just this one plot from Wikipedia.

I: Ok, then this is the end of the fourth block. Do you have any other comments about the specific steps of the data fabrication process that you think could be interesting for us to know?

P: No.

Block 5: Underlying Rationale of Data Fabrication Process (Why?)

I: Ok. Then, we will start with the fifth and final block. The goal of this block is to get some information about the underlying rationale of the data fabrication process. So, the first question is: Did you consider fabricating these data a difficult task to complete?

P: No.

I: Ok. Do you think your approach to data fabrication will be difficult to detect as fabricated?

P: Yes.

I: And why?

P: Yeah, first of all, because you are asking for aggregated data, so all these things that make data suspicious are lost. So, maybe, you will - so, of course, if you compare this data to the original Stroop task, you will find, well, different mean response times and stuff, and you might say, typically - so I didn't check, like maybe there is a meta-study somewhere this article of MacLeod, 25 years of the Stroop task or something, which is now also 25 years old, so I didn't check that if the response times - or if you compare that to the real data probably you will just see, yeah, here, we have a, what is it, a difference of 100 milliseconds, but for the Stroop task, the difference is typically 150. And then, of course, then I am screwed, because I didn't check these things, but from sort of the structure in the data I think you will have a hard time.

I: Ok.

P: On this aggregated level.

I: Ok. And why did you decide to participate in this study?

P: Oh, because first of all because I find that very important, very important issue to research. I mean, it is - currently, it is embarrassing to be a psychology researcher, so because of all that - I am trying to find another word for this word - because of all this p-hacking and obvious data hacking and [?], it all goes down the drain, right. I go here (?) - if I just follow this hallway here and then now I think every second or third researcher is currently cheating on the scientific process and then I am very angry of course about the, well, the level of statistics that people are doing. So, this whole null hypothesis testing framework. So, I am very angry about this situation. And this is why I had a strong motivation to support this project and, of course, because I have done that many times for the classes that I give. It was a piece of cake to do. So for me, it is just a routine task.

I: Ok. And did you discuss this study or the fabrication of the dataset for this study with other people?

P: I told [REDACTED] and I briefly discussed it with one colleague.

I: And did these people help you to fabricate the data?

P: No.

I: Ok. Then this is the end of the fifth block. Do you have any other comments about the underlying rationale of the data fabrication process that you think could be interesting for us to know?

P: No.

I: Ok, then this is the end of the interview. Is there anything else you can recall about the data fabrication that you think is worth mentioning?

P: No.