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 participat

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. So, the first question is: Are you a PhD Student?

P: No.

I: And how many years has it been since you got your PhD?

P: 6 years.

I: 6 years, ok. And what is your field within ...

P: Oh wait, I am sorry. It's 8 years. I am not that good in math. Nono, I got - it is a bit confusing. I got my PhD in 2009. So, 8 years this year.

I: No problem. What is your field within psychology? With field, we mean for instance social psychology, cognitive psychology, and so on?

P: Cognitive neuroscience. So, it is cognitive psychology with an emphasis on neuro processes.

I: Ok. And did you conduct any experiments including a Stroop task in your career so far?

P: Yes.

I: Ok. Could you describe how many or like in a bit more detail like what your knowledge and experience with these experiments is?

P: We did one Stroop task where [REDACTED]. That data set was in the end not published because the experimental protocols or the description of the experimental protocols was not good enough so we could not make sure that the data was actually accurate. I did several other Stroop demonstrations which are basically – and is (?) one study for demonstration purposes. One was with [REDACTED]. Plus one demonstration where we [REDACTED]. I got one publication on the Stroop effect in which [REDACTED].

I: Ok and like were you involved in all parts of the study? Like did you like do the study and also the analysis of the data and so on?

P: For the study with [REDACTED] - so that is the study for which we could not guarantee data integrity I was only involved as a supervisor which is also the reason I cannot guarantee data integrity. For the demonstration effects, I have done the entire - well - writing of the stimulation script, carrying out the experiment and analysis of the data. And for the other publication where [REDACTED] I did part of the data analysis and part of writing of the paper.

I: Ok, thank you. 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 program?

P: At least once a week, Matlab and 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: Compared to my peers?

I: Yeah, with peers we mean other researchers or scientists in your field.

P: Oh, then I am very average. So, 5.5.

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

P: 7.5 to 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. Did you fabricate the data in one day or spread the data fabrication over several days?

P: Two days.

I: Two days, ok. And how much time do you estimate that it took you to fabricate the data in their entirety?

P: Ehm, let me think. That would be including all the sort of tinkering and eh - about 6 to 8 hours, I guess.

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

P: Well, I really tried. So, I would give this definitely a - at least a 5.

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

P: Yes, yes. I read up a bit on models of the Stroop effect and - yes, so preparation was involved.

I: Ok. And how much time do you estimate you spent on preparing?

P: Oh about 2-3 hours reading - also refreshing up some statistics and ...

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

P: No.

I: Ok. Or did you look into previous cases of data fabrication and how they had been detected?

P: Well, I am pretty familiar with the Stapel case. So, I did not really take that as an inspiration but I am quite familiar with that case anyway. So, I had that in the back of my mind.

I: Ok. And did your knowledge of the Stapel case influence your approach to fabricating the data?

P: No, not really. The nature of the Stroop task data is quite different than the stuff that Stapel fabricated. So, the main thing that I took from that is that the fabricated data should look as natural as possible and not too good to be true.

I: Ok. And could you describe in a bit more detail how you prepared?

P: I particularly read up about models of the Stroop effect. Because my idea was that you can get the best quality fabricated data if you can model the individual participant and then simply let your model participants do the task. So, that is in the end what I did.

I: Ok. And like and how did you read about it? Like did you search for articles online or ...?

P: Yeah, I - well, because of that I have worked with the Stroop task before I knew a couple of models. So, I reread those particular models. And went over my own experiments again.

I: Ok. And so your preparation influenced your approach to fabricating the data?

P: Yes. Yeah, because you want to have natural looking data. Yeah, that is definitely made me a bit more careful in fabricating this particular data set. I mean the manual (?) approach is that you can try but I decided to go for more the model-based approach. In the end, I went with a very simple model but I really explored some options there.

I: Ok. And which other approaches did you consider?

P: Well, if you look at this - the basics of a Stroop task are quite simple: You got two conditions of which one of the two is about 30, 40 milliseconds slower than the other one. So, what you can do is look at the average reaction time for congruent trials in a population, average incongruent reaction time in a population, look at the standard deviations, and then simply draw random numbers with those particular properties. But the problem is that your data

will then not really look convincing. So if you do just your t-test then, sure you will get a significant effect. But if you look a bit closer at for example correlations between the reaction times, in particular the standard deviations within trials of individual participants, it won't work. So you really have to look at the individual trial level and individual subject level. So, the simplest approach which, for demonstration purposes, is sometimes used with students is simply generate an array of numbers with a particular mean and standard deviation, generate two arrays of numbers with a mean and standard deviation and then introduce the difference. So, they can practice that t-tests, correlations, or whatever. But for this particular purpose I found that that would not work because that would be a bit too easy to spot.

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

P: Ehm, well, if you look at this particular task, looking at the amount of time to do this properly, it probably would have been faster to just run a Stroop task via MTurk and submit that data. So fabricate the data fabrication.

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

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

P: Well, if the data is too clean, that is one thing. So if you look at real experimental data, there are going to be extreme reaction times in there, there are going to be participants showing opposite effects. Real data looks messy. And if you look at fabricated data - in particular how Stapel did it - the data is just too clean, things not really feel right in a way, standard deviations being to small etc. Another thing that might be a give away is if, for example, fractions don't add up. So, in this case, you are asking for reaction times in milliseconds in whole numbers and you got only 30 trials. So, that means that you have very specific divisions of your numbers and if you are not careful, then that will quite easily stand out. [?] believe that you have this recent paper with Tim van der Zee, middle author - forgot his name - and Nick Brown looking at all this work by Brian Wansink's food lab you got similar instances: You got tables where the fractions simply don't work - they don't add up to whole numbers if you multiply them by the number of participant that are averaged. So, that is also a sign of data that might not necessarily be fabricated but where you can notice that something is off. So, that would be also a clear sign of, well, something is wrong. So, those are things that I would look at: data that is too good to be true, too clean, and data that doesn't add up.

I: Ok. And could you name specific characteristics that would make data look genuine or more genuine in your opinion?

P: Well, in this particular case, what I did for this Stroop task - you are looking at a within-subject design which means that if you look at the variability in reaction times for the two conditions within your participants, it should be a correlation there. So, that is one thing that I would be looking for. The other thing is, the average reaction time should be kind of normal for what you would expect from a student population. The magnitude of the effect should be kind of in the range that you expect from previously published literature. So that is generally what would look like - make data look more trustworthy. Another things, of course, is the sheer volume of amount of data. I mean, here you ask of 25 participants, 30 trials, well, that is not that much. But if I, for example, look at fMRI data or EEG data, the sheer quantity of that data makes it extremely unprobable that data would be fabricated. So, that is another thing that I would say, well, not going to assume that [?] potential data is going to be fabricated. Of course, it doesn't exclude the possibility but it is not sort of basic assumption.

I: Ok. And did you take these characteristics you just mentioned into account when fabricating the data?

P: Yes, I double checked myself on those.

I: Ok. And how did you do this?

P: Well, by very carefully looking at what I am doing. So, just to give you an example: If you look at the random number generation I used in Matlab, by itself it doesn't give integers unless you specifically ask for that but I used a random normal distribution. That function does not give you integers so you have to round up that data and it is very to forget that because the average that you get are your three numbers and then a whole lot of decimals after the comma, but you have to double-check whether those decimals actually fit. So, that is stuff I did. What I also did is I double-checked whether the correlations between the standard deviations and reactions times for the congruent versus the incongruent trials were, well, in the range that I would expect. So, yeah, I double-checked for that whether that was actually in there.

I: So, you had sort of a checklist or . . . ?

P: Yes.

I: Ok. And did you take into consideration relations in the data other than the Stroop effect itself?

P: Yeah, the standard deviation of . . . yeah, that should be fairly constant within participants.

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

P: I looked primarily at the . . . of course, whether the - there was an effect and if

it was in the appropriate direction. And I looked at this standard deviation or (?) the consistency within my simulated participants. Those were my main things. So, I deliberately did not look at p-value, I mean I got a couple of simulated data sets when I was fiddling around with really very significant results that would - in theory could be true, but in the end, I am now happier with the data set that I just gave you with the errors in there. I mean p is .03-ish, but actually with an experiment with just 25 participants and only 30 trials, I would not be surprised if that is what actually would come out of it.

I: Ok. And did you also have specific and different criteria for the means and standard deviations?

P: What I did there - I am not sure whether you have a question with exact details, but what I did is I took an old data set with Stroop data - I only had 11 participants in there, but with a lot of trials. So, per participant, I computed the average congruent reaction time plus standard deviation, average incongruent reaction time plus standard deviation, and the Stroop effect per participant plus standard deviation. And I simply used the model that the Stroop effect is an interference effect that your sort of baseline reaction time would be the congruent reaction time, but that for the incongruent trials you would have this Stroop effect added up there. So, I had my 11 participants, I took the parameters from those 11 participants, and I sort of made random participants based on those parameters. So, I can also give you an excerpt of the Matlab code I used for that. But, basically, the entire process is for each fabricated participant, pick the parameters of a real one, draw random numbers based on those parameters for the congruent set and then for the incongruent set look at the error rate of that individual participant - or simulated participant - simulate if there is going to be an error - yes or no - if there is an error take congruent reaction time, if there is no mistake then take the congruent reaction time, add the Stroop effect, of course, with noise added to that and that is your incongruent reaction time. So, what this procedure does is it gives you a pretty natural looking data set but also consistency within participants. In particular, if you look at things like a speed-accuracy-tradeoff and things like that, so what I have gotten here, for example, that my fastest fake participant also makes the most mistakes. So, in terms of decreasing Stroop effect, because there are actual mistakes in the data set. So, by doing that you actually have quite a bit of internal consistency within your simulated participants. The only downside is I actually had only 11 participants in my - in the data set that I could find so fast, so quickly. It sort of means that if you are going to look at this particular data set, the variability in reaction times would be a bit less than you would expect for 25 participants. So, that is the reason why I am not saying this is 10 out of 10, you are not going to pick this up. I think you will pick it up, because the variability in the reaction times over participants is not big enough.

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

P: One thing I have been struggling with - it is not going to make a huge difference,

but I would have used Poison distributions in order to get the congruent reaction times rather than standard normal distributions. I don't think it makes a huge difference but it would be nicer. Of course, it is possible but it is more difficult to scale the Poisson distribution - I have, of course, tried it, but the built-in Matlab function - the standard deviations of the Poisson distributions were not large enough, so my data was too clean. So, I got back to the standard normal distribution and, I think, that as a result of that a couple of reaction times are faster than you would expect on the single trial level. Probably, it will average out on the group level, but that would have been nicer to actually have Poison distributions and use that as a basis. And what would have been really ideal, but I find - I don't think it would have made a huge difference - but if I would really have had 2 or 3 days time I would have implemented this model by Cohen where you actually have a pretty good [?] network model of the Stroop task. That would have been ideal, but, I think, this comes pretty close.

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. The main thing here is that I used real data in order to extract parameters for my fake participants and - yeah, that is the main thing. I - given that the Stroop task is such an easy task and something that everybody has some data lying around it is something that would be quite widely available for people to fake their own data set.

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

I: Yes. 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. So, could you indicate what steps you took to fabricate the means for the participants?

P: Yes. Like I said, I have a data set which is by now over 10 years old. It is a data set in which participants – it is (?) the [REDACTED] data set of which I am not 100 percent sure if it is actually completely genuine, but have all the reasons to believe it is. So, this is a very simple experiment where participants have 3 possible colors and congruent and incongruent conditions. 11 participants participated in this experiment and I only looked at the baseline data. So, this is a Stroop task which they did before any kind of [?] stimulation whatsoever. It is an average of from the top (?) - we had 100 trials, 50 congruent, 50 incongruent, and like I said 11 participants. So for each participant, I computed mean, standard deviation of the - mean and standard deviation of the congruent trials, same for the incongruent trials. I computed the mean Stroop effect, standard deviation of the Stroop effect, and the percentage of errors per participant. So, for 11 participants, I had these parameters. Subsequently, I simulated my 25 fake participants by for each participant randomly selecting one of my 11

models. And then based on these model parameters, first create 30 congruent trials and - let me see if there is anything interesting in my notes - yes, the 30 congruent trials were simply created by drawing a number from a standard normal distribution with standard deviation - or a normal distribution with standard deviation of that particular model participant and the mean of that particular model participant. Now, for the incongruent trials, I assumed that people would make mistakes. So, I assumed that people would not make mistakes for the congruent trials, I assumed they would make mistakes for the incongruent trials. So, for each incongruent trial, I first randomly decided if this would be a mistake trial or not by comparing a random number to the model's error rate. So, you draw a random number between 0 and 1. If it is smaller than .07 or .10, then it is a mistake. Otherwise, it is a correct trial. In the case of a mistake, I simulated the reaction time of that trial similarly to a congruent trial. So, that is, a random number drawn from a normal distribution with standard deviation of the incongruent trial added with the congruent trials. Oh, I actually see it could be one step better, one tiny mistake here in the model. And ... there is indeed a tiny mistake there. I don't think it would matter too much, but ... Ok, then if there was a - if the trial was correct, so that means, this is an incongruent trial, so there is Stroop interference here. I used a random number with standard deviation of the congruent reaction times [?] the congruent reaction times, but added a Stroop interference and again that is a random number drawn from a normal distribution with a standard deviation of the standard deviation of the Stroop effect and mean of the Stroop effect for that individual participant. Then, I rounded these numbers in order to get to these rounded numbers of milliseconds. And yeah, that's it. That is the way I computed each of this individual trials. Then per participant, you average that and compute the standard deviation. And that is what goes into the excel sheet.

I: Ok. And how did you decide about which data set you would use as a baseline? What were the reasons for selecting this data set?

P: Convenience. I had that data set available. There are many other possible data sets that you can use. For example, you got these open data sets which include Stroop tasks. I believe you got another data set which contains the Stroop task, but it was purely convenience. These were 1st year undergraduate students from [REDACTED] university who would be paid quite well in order to perform at a particular level. So I thought, ok, this would be representative participants. Moreover, the reaction times and magnitude of the Stroop effect were quite in line with what I observed here as well. So, I believed that these would be good model participants. Moreover, I had access to the raw data of this particular data set which allowed me to create rather detailed - or parameters for my models of my fake participants. So, this was the data set I had available with all the data I needed at this particular moment.

I: Ok. So, the next question would be: Could you indicate what steps you took to fabricate the standard deviations for the participants? But I ...

P: Well, no, that is quite detailed in there.

I: Yeah. Did you repeatedly fabricate data until you were satisfied with the results?

P: Yes. I tried out a couple of things. In particular, the instructions were not clear about whether the 30 trials would include mistakes or that is (?) what we clean data. In the end, reading the instructions, I took them very literally. And that is, each participant sees each trial or sees 30 trials of each condition. So, that is in the end what I simulated and I included the error trials and I did not get rid of the error trials, but that is something that [?] struck me after working on it for a bit. And I thought, no, I should not be using the cleaned data that is simulated but actually also include the errors. So, I went through a couple of iterations of the data until I was, well, satisfied with it. So, the data actually has a good internal structure, good correlation between standard deviations, and sufficient number of errors and not a too big effect which you wouldn't expect with a relatively small data set like this one.

I: Ok. And how did you determine whether you were satisfied with the fabricated data or that they needed to be adjusted?

P: I looked at the properties of my original data set, in particular at the consistency - the averages - that is something that you get because that is the parameters that you put in there. So, I was not worried about that one. The main thing I was worried about is the correlation between standard deviations within participants. So, that is pretty high, .85 for my real data set and I aimed for at least .8 for my fabricated data set. [?] after a couple of runs, it turned out that I had actually .95 or close to 1 even - well, that was too high. Indeed, that turned out to be a programming error. So, I repeatedly checked that. That was my main cutoff (?) to see whether things were right or not.

I: Ok and did you try to inspect whether the fabricated data looked weird?

P: Yes. In particular these overly high correlations, effects that are too large, standard deviations that are surprisingly large or small or - those kind of things.

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

P: Yeah. I did not look at each individually simulated trial, but I definitely double-checked whether the numbers are within the range that I would expect.

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

P: About 6 or 7. With different distributions and time (?) and programming errors and stuff like that. So, yeah, something like that.

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

P: I did the entire model in Matlab.

I: Ok. And did you use a random number generator to simulate data during this study?

P: The pseudo number generator. I would have liked to use my [?] number generator, but I didn't have access to that from home, so.

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

P: Yes, indeed, as a baseline for - and generating the models for fabricating data.

I: Ok, but you didn't like copy-paste like cases and . . . ?

P: No.

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, I think I went through everything. No, the only thing is that, now, I am looking at my pasted code here that there is still one variable in here which - I used the wrong standard deviations for the incongruent trials, in which participants made a mistake. So, it is not going to have a huge effect, because - [?] probably going to be about 20 trials at most over all participants, but still. It might have been a small effect there.

I: Ok. And how would this influence the fabricated data?

P: The standard deviations between the - the correlations between the standard deviations for the congruent versus incongruent trials is probably a bit smaller than it would be if I did not make that mistake.

I: Ok.

P: But like I said it is not going to be a massive thing.

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

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

P: I would not say difficult, but it was an interesting challenge. I mean there are different ways you can approach this and, in particular, I was interested in something that would look real. So, like I said, the easy way is to just generate two - well, arrays (?) of random numbers based on population characteristics, but I wanted to go for something a bit more sophisticated. And it was not particularly difficult, but thinking of every single detail like the exact number of trials and stuff like that, rounding up your random numbers, that is - it took some attention.

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

P: With a data set of this size, I think it will be difficult to spot that this is fake data, but I don't know. I think, it might be difficult, yeah.

I: And can you think of ways how it could be detected as fabricated?

P: I think that the averages - Because I used a normal distribution, it might be that the distribution is just a bit off. That is one thing. The other thing is that the standard deviations might be a bit larger than you would expect. Those are the main things I can think of right now.

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

P: Oh, for several reasons. First of all, I think it is a very interesting and important project. I am not sure if it is going to work - that is for sure. But I think it is a very important project and there should be a lot of - it is something I really would like to contribute to and I hope that many other people contribute to it as well. So, that is one thing. And the other thing is that it is kind of interesting to think about, ok, how would I go about fabricating a data set that looks as real as possible? So, in that sense, it was also a fun challenge.

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

P: No.

I: Ok. So, you had no help in fabricating the data or so?

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 think that the main - I think it depends a bit on how people would go for this. I am not sure what your sample is but for me as a cognitive psychologist it was very natural to go for this as a - there (?) should really simulate this at the lowest possible level and that is go for the individual trials. Don't try to simulate the entire population all at once. So, that was a very important consideration for me as for (?) the background and rationale for the fabrication process.

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

P: Like I said, the instructions were a little bit ambivalent and a bit unnatural (?) for an experimental psychologist. First of all, the number of trials that has been asked for in the sheet is a bit low. I mean, 30 trials for a reaction times task is not really I would typically recommend people to use. In particular not, if you want to know something about a proper fit of parameters. Second, it was not directly clear from the data sheet what kind of cleaning of the data would we apply. If you look at the data I (?) submitted, that was completely uncleaned data. If it was for an actual experiment, I would first remove outliers. And if

you are interested in Stroop interference I would also remove the incorrect trials. Outlier correction [?] is a big debate of course, but typically is done by removing trials more or less than 3 standard deviations from the mean. And excluding incorrect trials, if (?) you would do that to this particular data set, the effect that I report in the Excel sheet would be probably larger. And therefore, you would get a smaller p-value. This right now is just - well, p is .03, it is nothing to be - to [?]. It is statistically significant, but not brilliant. But yeah, that is a more general thing. So, I had to really check myself: Am I not making any kind of hidden assumptions here? So, in the end, I really stuck to the letter of the fabrication instructions, just generate 30 reaction times, 30 trials, average those, even though I would probably not do that in reality when I am analyzing data of a RT task.

I: Ok. And the other data set that you submitted ...

[Phone ringing, participant decides that it can wait]

I: And so the other sheet that you submitted, you did all the things you would usually do?

P: I only excluded the mistakes. So, I simulated a participant there who does not make any mistakes. And that is, of course, not a realistic situation. And that is also why I think that the data set that I actually submitted is the better data set, because it also includes the speed-accuracy tradeoff. People who make more mistakes will generally be a bit faster.

I: Ok. But also in the other data set, you did not exclude outliers or so?

P: No.