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. So, yeah, some of the questions may be obvious, but I will just ask them. So, the first question is: Are you a PhD Student?

P: No.

I: No, ok. How many years has it been since you got your PhD?

P: 25, yeah.

I: 25 years?

P: Yes.

I: Ok. And what is your field within psychology? By field, we mean for instance social or cognitive psychology?

P: Cognitive.

I: Cognitive psychology, ok. And how many experiments including a Stroop task have you conducted in your career?

P: Uh, so many. And do you mean personally or on papers that I was on? Because most of the time, of course, I am on the paper, but not - I did not conduct the experiment myself, but it must be dozens and dozens, yeah.

I: Ok. So like if you could give a like rough guess of a number maybe of both - like how many you conducted yourself and how many you were on the paper.

P: Ok, yeah. Because if I, you know, supervise a graduate student I don't consider it conducting the experiment. I would say, ok, let's say that I roughly have 5 papers per year and 3 of those are experimental papers and they have roughly 3 experiments per paper, then I would say that I conducted myself maybe - well the first 5 years maybe - so that would be - well, I would say myself between 200 and 300 and then - but it is a rough estimate and then you know as a co-author many more than that.

I: And all of these experiments included the Stroop task?

P: Oh, the Stroop task itself, I have done - I was on a paper where it was used only once. And I think I have used it in the lab as a sort of, you know, pilot experiments for students who took cognitive psychology, you know, where they have do an experiment in the lab. But, yeah, there is one paper that I had where we used the Stroop [REDACTED], but I didn't run that experiment.

I: Ok, thank you. Which statistical analysis programs do you use at least once a week? Multiple answers are possible . . .

P: Ok. At least once a week, JASP. I used to use SPSS, but I don't use it anymore. And I use R for things like meta-analyses.

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

P: And with my peers you mean other cognitive psychologists?

I: Yeah. Relative to other researchers or scientists in your field of research.

P: Yeah, ok. Well, I would say better than average, but not much better. So, what would be the average? Let's say 7.5 or something.

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

P: 6.

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

I: 6, ok, thank you. This is the end of the first block about general information. Now, we will start with the second block and the goal . . .

P: Yeah, I have the - what I did . . .

I: Yeah, yeah, yeah. No, I was looking at the

P: Oh, sorry.

I: . . . whether the - yeah, no, no. So, the goal - now, we will start with the second block and 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 did - I have 8 steps and 1 through 7, I did in one morning and step 8 I did a few days ago, because I thought of something that you might be able to detect. And so I added an extra step.

I: Ok. So, the next question is then: On how many days did you work on fabricating the data and -?

P: I would say a total of 3 hours.

I: Oh, first days that were then 2, right? Like . . .

P: Yeah.

I: Ok. And the next question would then be how much time do you estimate that it took you?

P: Yeah, about 3 hours, I would say.

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

P: I would say 6. It was more work than I thought.

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

P: No. You mean like reading literature?

I: Yeah.

P: No, no.

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

P: Yes. Well, I didn't look at them for this, but [REDACTED]. I had a lot of interaction with Uri Simonsohn. So, I knew that - you, know, the things that he looks for like a lack of round numbers or something like that. So, in Smeesters' data file, the numbers 10, 15, and 20 occurred much less often than you would expect. And there (?) were also streaks of data if I remember correctly. There were very few streaks. If you order the data from low to high - of course, it works different with reaction times, these were ratings where you only have 7 options or something like that, but Simonsohn found that Smeesters data looked too regular, not random enough. So if you had - if you ordered all the ratings, you would have three 1s, three 2s, three 3s and so on, whereas in actuality you would have, you know, two 1s, six 2s, five 3s and so on. So, I tried to make it a little bit more random. I tried to avoid these traps.

I: Yeah, ok. So, you used your knowledge from the - ok. And like can you say again specifically how you used it by . . . ?

P. Well, by trying to avoid certain regularities. So, I used - well, I will show you in a minute, but some randomizer here and there, because I thought, well, if I am gonna do that by hand, it is probably not going to be random enough and that's what I learned from these earlier experiences, that people tend to create patterns that are actually not random enough, basically, and like avoiding round numbers and those sort of things, because they think that makes it looks random, but in actuality those numbers do occur, of course. So, I tried to use some of that by using a - by basically changing values using a random number generator.

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

P: No, not really. It was - I have to say, though, that it was much more effort than I thought. It is almost easier to run the experiment to be honest. If I had run the experiment on Mechanical Turk, I would have been done more quickly. Yeah, that is - that surprised me. Also because you are second guessing, so I am thinking the whole time if I were in their position, I would be looking for this and that. So, I should make sure I, you know, don't do these things. That makes a little, yeah, more work, more intense than I had expected. And also it feels strange to do it, it feels very odd, almost like it is a criminal act which normally it is, but . . .

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 more or in general fabricated in your opinion?

P: Well, one of them would be, you know, the lack of round numbers relative to what you would expect in the data. Something else that would be - at least that I tried to look for is having a reasonable effect size, sort of the effect size you might in the literature. Trying to have enough variability in your - here we have response time, of course, and standard deviations and one of the things - even though by a large there is a relationship, lower reactions times have lower standard deviations, it is certainly not a perfect correlation, so your data should reflect this. You could have a, you know, a low reaction time and still much higher standard deviation than a person with a longer reaction time and you could have a smaller standard deviation. So, I tried to look for that and then I realized that - but maybe then I am getting ahead of myself - that it would be difficult for me to produce that by hand and so I used an existing data set and tweaked it. That is basically what I did. Because I thought there may be all kinds of hidden regularities that I am unable to see but that you guys will be able to pick up on. And so, I thought if I use a naturally occurring data set and I tweak it in a certain way, then it probably still has these regularities that are sort of hidden to me right now but that you guys will be able to pick up. So, that was basically my thinking.

I: Ok. And related to the last question: Could you name specific characteristics that would make data look genuine or more genuine in your opinion?

P: Well, I think data are usually a lot noisier than people think. And so, having - and this was the problem, I guess, with the Smeesters for example is that there was much too little variability across subjects and the effect sizes were unreasonably big given the nature of the manipulation. So, you know, with some social priming experiments or he had like a three-way interaction that was highly significant and I just knew from my cognitive experiments that even though he used between-subjects designs, with within-subjects designs those are already extremely hard to get. And his data just looked completely unrealistic from that point of view - the strength of the manipulation and the effect size. And then, of course, you can have one lucky shot where you have a small n and still a big effect and, of course, it has to be big for it to be significant given the n, but if you have a string of those, then something odd must be going on. So, what I tried to do with this set is, you know, make sure I maintain sort of a normal variability in response times which I think I am quite familiar with, because I do a lot of work on response times, so I have an idea of how much subjects can vary along that line. So, for example, if you have a within-subject manipulation - even if you have a Stroop task where the effect is typically big, it is never the case - or in my experience that all the subjects are showing the same effect, going the same way that the match in this case the color-match is faster than the mismatch is true on average, of course, but certainly it is not true in every data set that all the mismatch conditions are slower than the match. So, that is for example another pattern that I tried to avoid, but I have seen other cases of data fabrication that I can (can't ?) mention where people had every subject in the experiment showed one condition being faster than the other one in a response time experiment and I have never seen that in real data. So, that's why I, you know, tried to avoid it here as well.

I: Ok. And, yeah, did you take these characteristics you just mentioned into account when you fabricated the data? And if so, how?

P: Yes. Well, I did that by using a irrelevant (?) data set, I even have the link to the set. It is on the Open Science Framework. And I did that because I said - first, I thought I am going to simulate it, but then I thought, well, I am not so good at that and that will take me more time, but then - and I thought, you guys will probably be able to detect that. So, it is much better if I use an existing data set and then manipulate it a little bit. So, I used a data set, you know, from a Stroop experiment and it had 25 - aeh it had 24 subjects, so I created an additional one basically taking a mean and a standard deviation that were somewhat comparable to the rest. And I - for the existing data set, I computed all the condition means and standard deviations and then they had 21 observations per condition, but you wanted 30, right? So, I thought, well, if you have more observations, then the standard deviation is going to be lower per subject, so I took the 21 response times from a condition and then copied the last 9 ones, then I recomputed the standard deviation so that it became a little bit smaller overall, not always by the way. And then, I did that for the first 3 or 4 subjects and then after that I sort of got a sense of ok it is going to be reduced by 20 or 30 milliseconds to that - or 10, yeah, 10 to 30, so I basically said, ok, this one had 225, I will make it 220. And this one had 234, I will turn it into 214 or something like that. And I did that for all the standard deviations. And so then I had a pattern that looked realistic, I thought, because now I have 25 subjects and I have standard deviations that look reasonable for if you have 30 observations per condition. But then I thought, well g (?), maybe they have some sort of a program that just looks for these kind of reaction times on the internet, you know, some search program, so I shouldn't have an exact match. And then I basically took the condition means and in Excel I randomly added or - plus or 5 - plus or minus 5 milliseconds, so, you know, so the difference between the means would still be fairly realistic for a Stroop task, but if you did a search, you wouldn't find these exact numbers online. And then I also re-ordered the subjects to make it a little bit harder to match it up with the original file. And that is what I did that one morning. And then a few days ago, I thought, well, but maybe the condition means are still going to be very close to the original, because all I did was randomly add or subtract, you know, minus 5 to 5 milliseconds, but on the average you should get pretty much the same means. So the only difference was the 25th subject that I added, but that didn't change the overall mean so much. So, I basically - after having re-ordered the subjects, there was a different number 25 and I gave that subject longer reaction times, so 100 milliseconds in each condition, and I also made the standard deviations a little bit higher. So that that influenced the overall mean so that now if you guys compute the overall means for my conditions, they are not the same or not, yeah, highly similar to the ones you can find online. Because I figured, well, one thing they are going to do maybe is for sure look for regularities in the data or lack thereof. But something else they might do is just, you know, figure out, oh, some people may use existing data sets, so we are going to have a program that just goes through the entire Open Science Framework or something like that. So, I shouldn't have perfect matches to the data in there. So, that is why I did that.

I: Ok. Yeah, the next question is: Did you take into consideration relations in the data other than the Stroop effect itself?

P: Yeah, like the relation between the mean and the standard deviation. And that is why I, you know, because you were asking for 30 observations per condition and not 21, I lowered the standard deviation based on what I saw on the first few subjects that you get a reduction of 10 to 30 milliseconds, you know, per standard deviation. And so I made them all a little bit lower, some I may have made a little bit higher because that could happen, too, of course, but that's it. So, because I knew that the mean and the standard deviation are roughly correlated in the sense that lower means give lower standard deviation, but it is certainly not perfect like I said earlier, so ... But that was already taken care of because in the original data set you saw that very clearly. Let's see I can even show you, that is maybe easiest if I - [Participant searches for file on his device]. Ok, here this is it. Yeah see, so here for example you see - oh, well, that is a rather high mean, 530, but then the standard deviation is only 93. And so that is very low given how high the mean is. Here, you have a lower mean or a mean that is about 540, but now the standard deviation is 221, so it is still a lot of variability. So, I thought the easiest way to create that is to use an existing data file and so that is what I did.

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

P: So, well, one of the criteria was what I knew about, you know, Simonsohn the things that he looks for. And then I thought, well, you need to have a reasonable effect size, comparable to what you find in the literature. The relation between the mean and the standard deviation needs to be reasonable, but certainly not too highly correlated. And the - but the standard deviation on average needs to be proportional to the mean, so that is why I went from, you know, the 21 observations per condition I had to 30 to lower it a little bit, because otherwise I think the standard deviations would have looked too high for the means on the average and then of course the effect size would also have been smaller, because now (?) we are dividing a difference by a larger number.

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

P: Yeah. Well, for the means, I basically - I stuck with what I had from this study and I just did the, you know, set the random adding or subtracting minus 5 to plus 5 milliseconds. And so that means that the times are still very close to what was found in this original study. And for the standard deviations, I had to do more, because I basically didn't have 30 observations.

I: Ok, so you have named the research done by Simons a couple of times. Could you name the specific papers that - or the specific methods that [?] . . . ?

P: [REDACTED] it should be in the report for the Smeesters committee. And I also read the Stapel committee report. I guess those analyses were maybe done by Jelte, I am not sure, but so I also read that. And I know they also look at other things like p-values and so on, but that wasn't applicable to this case, of course, because we only have one p-value here.

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

P: No. I think, I mean I am sure you guys will find things that I did not think of, but right now I can't think of anything else that I should have done.

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 that you think could be interesting for us to know?

P: I think, well, one thing that occurred to me is even though I am an advocate of open science, you know people publishing their data or posting their data, there is a risk of course, because they could be doing what I am doing here. You know you could basically create some sort of Frankensteinian data set out of other people's data sets. That is why I think it is useful that I did it this way, because it would be very good if you could detect this, because I think this might be what people will start doing, because they will start thinking, well, it is way too difficult for me to create my own data set, I would have to, you know, learn how to simulate in R or whatever and even then I am not completely sure if I capture everything, why not take an existing data set and tweak it. And pretty soon, there will be data sets for pretty much every experiment that you want. So, you can just look at those and that is why I think your detection program should have something where it would automatically scan all those files on the

Open Science Framework for example and then not look for direct matches, I guess the way you could find this is if you ordered all the response times and then you would do a correlation with the ones that I have and the original ones, then that correlation would be extremely high, maybe higher than you would normally expect. That would be how I would probably try to approach cases like this.

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

I: Ok, thank you. Ok, now we will start with the fourth block. The goal of this block is to get some information about the specific steps of the data fabrication process. I know that you have already told me quite a bit, but I will just ask the questions and you can decide to what extent you like . . .

P: Ok, yeah, I have it open here.

I: Ok, so first question here in this block is: Could you indicate what steps you took to fabricate the means for the participants?

P: Yeah, so I took those from an existing data file, it only had 24, so I created the 25th and I randomly added minus 5 to plus 5 milliseconds to those condition means except for the one that I created, yeah.

I: And how did you create the one that you created yourself?

P: I basically looked at - I eyeballed the data and I thought well ok if I have let's see this one I think is the - yeah, here I think the 25th, no, this is - ok, here I already reordered the data, but say if I had to create another one I would say ok well I see the fastest one I see is 498 but then the slowest is 668, so I will say if I have 595 or something then by enlarge (?) the other condition is I don't know 50 or 30 milliseconds slower, so I will say 595 and 620 or something like that. So, I pick somewhere in the middle, I wouldn't do that if I had to generate more, but with one I thought I could take the risk.

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

P: Yeah, I - well like I said they also came from the data file that I used but they only had 21 observations, I needed 30, so I took for the first three subjects I computed the condition standard deviation and then I took the last nine items and copied and then I computed the standard deviation again and then I saw well it was about 10 to 20 percent reduction and because I did that for the first three subjects and then I basically lowered the other standard deviations by eyeballing the data by 10 to 20 milliseconds so that the relation between the mean and the standard deviation - the correlation would not be very much affected, but overall the standard deviation would be lower because you have more items, so otherwise it might not have looked realistic. People - you guys might have said, oh, those data are too noisy or something like that and the effect size would have also been smaller of course, because the standard deviations would have been higher, whereas the means would have been pretty much the same.

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

P: No, I think the steps that I took were the ones that I had in mind except for you know at the very last moment deciding that I needed to make the means for the last subject higher because I thought maybe you are just going to compute the means and see if you can find those means anywhere in the literature or something like that or on the internet. And so I tried to make sure the means were a little bit more different from the original - this thing keeps coming up - from the original than it would have been otherwise.

I: Ok. And at the end how did you then determine whether you were satisfied with the fabricated data?

P: Well, what I tried to do is second-guess what you would be doing. So, I thought you will be doing the sorts of things that Simonsohn did, so I need to make sure that I pass that test. And then I thought, but you may also just, you know, look on the internet to see if you can find these data, so I need to make sure that they are not easily found. And those were the two steps, I think.

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

P: Well, I didn't think they would look weird because of them being relatively close to the original data.

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

P: No, I thought they would. I didn't think I need to do anymore inspecting because you know these are pretty much the original data plus or 5 milliseconds and so that is not weird. I mean if I had done plus or minus 100 milliseconds then they would have looked weird, because then I would have really fast reaction times for a Stroop task, here the fastest is - well, what is the fastest - 400 and something, then I would have had 300 and something on the average and I would have found that unrealistic myself. If I would have looked at a dataset saying for a response time 350 milliseconds that is unrealistic for the Stroop task. But with the 5 milliseconds up and down change I don't think it is gonna look weird.

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

P: Aha, ok. So, well, depends on what you call fabricate. I would say that I fabricated all the means because I basically changed the original ones and the standard deviations I definitely fabricated because I, well, took the existing ones and then reduced them by a certain amount.

I: Ok, and was this process that you changed them one time or did you first change it one time and then changed it again?

P: No, just once every time.

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

P: Well, I used - because another thing I thought is maybe you are going to look at different versions, so I had a different Excel file and I pasted everything into this one and in that Excel file I created a new column that basically said, you know, take the original mean and then add or subtract, you know, -5 to 5 milliseconds and that was the new column and then I copy-pasted that column into this file.

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

P: No, not to simulate data, but, well, just to change data, I guess, yeah.

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

P: Yes.

I: Ok. And so, the next two questions are how much real data did you use and how much - aeh how did you use these real data and I think you have elaborated . . .

P: Yeah, I answered that.

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 now start with the fifth block and this is also the last block. The goal of this block is to get some information about the underlying rationale of the data fabrication process. First question is: Did you consider fabricating these data a difficult task to complete?

P: Yes, more difficult than I thought, because it is a - in fact, it is actually, I think, a useful exercise, because it gives people a much better sense of what data should look like. So, when you do your own experiment or, you know, you are a checking let's say a data set that a collaborator has produced, you can more easily detect if there is something weird about it, which could be fraud but could also be a mistake that somebody made, you know. I have seen it where students use cut-offs for response times if two standard deviations from the mean, but then they did it in the wrong column and so they got significant effects where I would not have expected effects and I looked back and I could find the error in the data file basically because so they had the cut-offs for the means but I said also compute the medians and the medians showed no effect, but the means did. And then I thought, oh, there is something wrong with how they did the means and I think those are the sorts of things you become more sensitive to when you do an exercise like this. So, I actually thought it was useful to do it.

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

P: I am not sure because I think you guys obviously know that people will be able to - or will probably use existing data sets. So, the data sets that I could find you can find as well. So, then if the transformation from my data set to that other data set is easier than I think, then you will detect it fairly easily. I am hoping that my - the few steps that I took will help mask you know this, but I am not completely confident.

I: Ok. And then the next question is: Why did you decide to participate in this study?

P: Ok, well, that sounded like a very interesting project that is kind of a counter-intuitive thing to do, but I think in the long run it will be very useful, because even though, you know, there are some fraud cases that have become public, I know of at least 2 or 3 other cases that have not become public and that probably will never become public, so if I already know 5 or 6, then you know there must be dozens and dozens. And so, that makes me worried, so I think if we have good software that allows us to detect fraudulent data then that will be a benefit to the field. And like I said, when I was doing this, I realized that the disadvantage of having open data is that there will be many more data sets for people to work with, to harvest so to speak, to create their own data set so the better the techniques we have to detect these things, you know, the more the field will benefit. So, that was my - I was intrigued and I thought it might be very useful

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

P: No, I think I have pretty much said everything I have done, yeah.

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.