**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: 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: Yes. I would say I am a PhD candidate, not student, but yes.

I: Ok. And what is your field within psychology? For instance social or cognitive psychology.

P: I'd say officially social, but more cognitive.

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

P: Yes. I have done one experiment with the Stroop task.

I: Could you describe that a bit more like in what context did you do it?

P: Yeah, we wanted to - we actually used it more as a filler task but it also showed us nicely that the Stroop task actually worked and we found sort of the predicted results, but we used in a larger study - it was a 45 minutes study with a manipulation and we used the Stroop task as a filler and also to sort of be sure that people's minds were sort of empty that they didn't have previous thoughts or instructions in their mind, but that they really had something to do during the filler task. So, that's what we used and there was a - we used an automatized Stroop task on the computer so people had - I think 4 colors and I am not 100 percent sure, we might have used a touch screen to make them touch the screen for the right color or we had 4 buttoms taped off with different colors that they needed to press when they saw the color on the screen - or the word with the color on the screen. So, I am not sure but it was automatized, so it was not - it was all on the computer and without the - without me present. So, I just programmed it in [?].

I: Ok. And did you also analyze the data from the Stroop task?

P: Yeah, I also analyzed the data from the Stroop task. Yeah, it is actually very easy to analyze - and it was beautifully - I think always, I think the Stroop task is the most robust effect. So, it worked, so we found a Stroop effect. So, it was

nice to see that we could sort of replicate this basic effect in our lab. Even if it was only for us a filler task, it was nice to see that we could sort of replicate this effect.

I: 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: I normally use SPSS, but I am trying to learn R. So, I am not doing analyses in R, but every week I try to read a chapter about R and then do a little bit, but I am not good enough to use it for analyses yet, but - so normally, SPSS.

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: So, my peers are PhD students?

I: Your peers would be other researchers or scientists in your field.

P: I think compared to other people in my field - other researchers in my field, I would say maybe a 6 or a 7. But compared to my peers as in PhD students I would rate my knowledge an 8 or 9.

I: 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: I don't know. It depends a bit how you are gonna check it. Are you comparing it with real data or are you doing it sort of based on an algorithm? So, I think it really depends on how you are gonna do it.

I: Ok. Like I don't want to provide you with additional information for this question. So, like you could either like give a global evaluation or you could also like specify like you just did for the last question like depending on what the test would look like how you would rate your confidence.

P: I think - so, I did a little bit of a research on what the data should look like. So, I based on a Many Labs Stroop task data. So, if that is used to compare it with, then I think the chance of detection is very low. So, that would be an 8, I think. But in general, I don't know. I mean I hope you could detect it, so I would say 5, maybe 4.

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

I: Ok. Then this is 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, did you fabricate the data in one day or spread the data fabrication over several days?

P: I spread the data fabrication - simulation over a couple of days.

I: And on how many days did you work on fabricating the data?

P: I think, maybe on 3 days. So, on one day I looked a bit about what the assignment was and how to do it, how - what I want to get out of it. So, one of the reasons I participated is also because I want to get better at simulating data. So, I thought this was a nice opportunity to sort of test if I could simulate data and make it look as real as possible for my own power analyses or preregistrations, for example. So, I wanted to do - like really put some time in it to see if I could actually do it. So, the first day, I made a plan of how to do it and read a bit about it. And then another day, after I had some thought about it in my free time, I think I spent like half an hour trying to do what my plan was. And then I noticed that it didn't work. And then on the third day, I tried a different approach that also didn't work and then I tried another approach on the same day just to have something to submit. So, that's a bit of the timeline.

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

P: I think in time that I spent on my computer I think I spent like 3 hours on it, but I also thought a lot about it at home, but I think that is difficult to estimate - yeah, I don't know.

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: Can I say 5.5?

I: Sure.

P: Ok.

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

P: Yes. Like I told you, I did it in different stages. So, one day I sort of thought about it, prepared it, and then over time before I actually started making data, so I had some thought about it, yeah.

I: Ok. And could you roughly estimate the time that you spent on preparing?

P: I think like 1.5 hours.

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

P: No. That is something I will do that later. Yeah, that is interesting.

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

P: Also no, no. I am in [REDACTED], so they sometimes talk about like simulating data and data fraud and that sort of stuff. So, sometimes I see [REDACTED] posts coming by about that topic, but I have never really - no, I have never read a paper about it.

I: Ok. And could you describe in a bit more detail how you prepared? So, you said that you like put some thought about like how you would do it. Could you describe that in a bit more detail? Like what you thought about, for instance?

P: Yeah. So first I thought, so you wanna - so let's see if you can detect data and my assignment is to make data look as real as possible so that you cannot detect it. So maybe, I hoped to come up with the best algorithm to test it in the future. So, I thought about how data should look like and to make it as real as possible. So, first, I looked at my own data that I had about the Stroop task to look - to see what it looked like. And then I looked at a Many Labs project about the Stroop task, so it is from multiple labs and I thought this would be more sort of reliable than my own data. So, I looked at what it looked like, I looked a bit at the means and the standard deviations and the correlations between the means to really see like what does it look like. And then I noticed that in the assignment it says there were three colors and two conditions, so in total 30 trials per condition. So, then I thought, ok, so right now in [REDACTED], we have this policy now that we have to upload all our data that we collect. So, not just the means, but all data. So, I thought if I were Diederik Stapel, for example, what would I do? I would not just fabricate means, but I would fabricate the entire data set. So, I thought, so for every color, I need to have a data point, because that is what you will need to upload at least here at my university to sort of show that you did not fabricate the data, but that it is real. So, I thought, I would create that for that reason. And also I thought, maybe if I create the real data point - or all the data points instead of just the means, maybe it will look more real instead of if I just come up with a random number for the means and standard deviations. So, I thought that would be the best approach - so based on real data and the mean difference there - so, we (?) chose a 15 milliseconds mean difference with on average 600 to 700 milliseconds latency. I thought I would fabricate a data point for each point, for each color, for each condition. So, then, I also thought maybe you would ask for it to sort of show how I did it. So, that is also why I did it. So, I really simulated the entire data set and used that to generate the averages. So, that is my thought process of how to do it. And I also wanted to take into account the correlations between means, so I thought, ok if people are faster on average, they will also be faster in the incongruent trials. And it makes sense that if people are super fast on congruent trials, their standard deviation will also be a bit smaller. Whereas if they are very slow, I thought maybe the standard deviation would also be higher. That is also what I see back in my own data. So, this is what I sort of wanted to simulate. So, I thought this was also a nice project to see, can I simulate real data and can I sort of use that algorithm that I create now for my own research if I wanna do preregistration, if I wanna test like ok, this is an analysis I wanna do, will there be (?) outliers, how will they look like. So, that was sort of my thought process for the simulation

I: Ok. Then this is the end of the ... Or no ... And how did this preparation influence your approach to fabricating the data?

P: How it influenced my ... Yeah, so that was really the sort of the outline or structure of how I was gonna do the data fabrication.

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: No, not really, no.

**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: That would make data more fabricated?

I: Look fabricated or more fabricated?

P: I think maybe specific repetition of numbers. So if people use the number 7 a lot of the time, but never the number 1, for example. I can imagine that when you are making [?] and it looks to even maybe. I can imagine that. I recently read a report about Brian Wansink's lab that he used really weird averages, really weird decimals that couldn't be existing in two populations. So then maybe if the averages are weird or couldn't be real. I think, maybe when you are creating data you might forget about that, but when it happens for real that would never occur. So maybe - I wouldn't know how to detect this, but an algorithm maybe could. And I think, maybe - yeah, for the Stroop task, it is difficult because there are so many ways to do it, but if there would be a standardized task that could only be done in one way, if the means were really off, that would be a sign. Or if the standard deviations would be really off, that would be a sign. Or if the correlations would be off. So, for example, I think the congruent and incongruent should be correlated highly because if you are faster on average then you will also still be relatively fast in the incongruent trials. So if something is odd, like if the correlations between data points are odd (off?), I think that would be a sign that something is wrong.

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

P: I think maybe if it is in line with the general results. So, the Stroop task almost always replicates. So, it should replicate now. The p-value should be very low. I think, if you have a p-value of .048 or something, it is so unlikely that you would get that p-value that that would look weird maybe. And I think also not every participant should show the effect. I think normally in social psychology or at least that is the field that I work – well, that my department works in, is using (?) the effect of averages and not an effect of every person. So even in

the Stroop task which is super robust, there would be 3 or 4 people that do not show the effect. So, I think, maybe people could forget that when they create data because want to make it look really nice. But I think not everyone should show the effect.

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

P: Yes. And that is why it was so difficult.

I: Ok. And could you describe how you tried to take these into account?

P: Yes. So, I tried to simulate the data based on what I mentioned before. I found a mean difference that was by 50 milliseconds as I wanted between the congruent and incongruent means, but there was no effect. And then I noticed that because there was no correlation between the congruent and incongruent means. So then I thought: Ok, so it is a within-design. So, how do I simulate random data that is also correlated with other data? So, then I googled a bit about how to do that. So, then it worked. So, I could create data – or simulate data without typing it in by hand that is correlated .8 – as I wanted or as I expected. The problem was that the standard deviations didn't correlate with the latencies of the congruent trials. So, then I tried to make that work as well. So, I used a specific formula so that it would also correlate. So then it worked but it messed up the rest of my data. So, it stopped looking nice. So, in the end, I did have to make a decision about trying better and better to do it, but that would take a lot of time. So, in the end, I decided to go with the data set I had and sort of manually adjust some numbers, because I couldn't make the correlations that I wanted with what I got. But so I did take it into account.

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

P: Can you give an example what you mean?

I: For instance, the distribution of the scores or other aspects that could be inspected with the data set.

P: No. I did think about it, but I thought it would be too difficult for me. So, I assumed that it would be a normal distribution. So, I used a normal distribution for the Stroop data. I also looked at the standard deviation of the standard deviation of every individual. So, I took that into account. But no, not really. I could have looked at it, but I don't think my skills are proficient enough to do it.

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

P: Yeah, so I think what really helped is that I really created 60 data points per participant, so, and I created means from that. So, I think that these data points would be likely to be found when someone actually do it. And because I created so many data points instead of just creating the means, I think that really helps to make it look real. And then afterwards, I checked in SPSS what the mean

difference was and there was an effect, there was a strong correlation between congruent and incongruent trials. But the correlations were not in order, so then I manually adjusted some of the variables in a certain way. And I checked with each change I made if it went in the right direction. And then, the effect was maintained and also I don't think I messed up the standard deviations and means so much. So, I think we still have the beneficial effects of all the data points that I generated, but also the benefit that the correlations were - between standard deviation and means were in the way I wanted them to be, which I think would make it more difficult to notice that it is fabricated instead of real data. But then also I thought, yeah, it really depends on what kind of Stroop task you use, because maybe it is different in other Stroop tasks than the one that I based my hypotheses on. So, that is, I think, the reasons that I think it would be difficult to detect.

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

P: Yeah, so I looked at my own data and I looked at the many labs data to look at what is the mean on average and what is the mean difference on average. So, I used that mean difference as sort of a starting point from which I generated 60 data points per participant. So, that is what I used. I also looked at the standard deviation and the spread within the standard deviations. So, I took all that into account.

I: Ok. And had you different criteria for the means and standard deviations to check whether your fabricated data might go undetected?

P: Can you give an example?

I: So, like ... I don't want to like direct you in a specific direction, but so like you mentioned that you like checked a couple of things for your data points where it might be detected with or so. And now the question is did you have like separate or different, specific criteria for the means and the standard deviations? Because on the one hand you had to fabricate the means and on the other hand the standard deviations.

P: Yeah, so that is a good question. So, now that we are talking, I didn't even check like with a histogram like how they are distributed. There are so many things I didn't check and now I think I could have done a better job. That is also why I said 5.5, because I think I did a really good job and really put a lot of effort into it, but maybe I did forget a lot of stuff. So yeah, I think – but I did have different criteria. I thought that the means should be correlated for the incongruent and congruent trials. I thought the incongruent mean should be correlated with the incongruent st... – the congruent mean should be correlated with the congruent standard deviation, but the incongruent standard deviation not with the congruent mean. And the incongruent standard deviation also maybe not with the incongruent mean. So, in terms of correlations, I thought there would be differences between standard deviations and means and how they relate to each other. But I didn't really use any sophisticated methods to check it, I didn't check any assumptions or anything, no,

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

P: Yes, so at least the distribution. Maybe in my own data set about the Stroop task, is it normally distributed or not? And maybe if not, maybe look a bit better into how can I simulate non normally distributed data. Because now I used a normal distribution. Yeah, I should have checked that.

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, I don't think so, no.

**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. So based on an earlier data set which I talked about before, I noticed that it was on average between 600 and 700 milliseconds. So, I created an Excel file where I created absolute random numbers with a normal distribution with a mean of 610 and a standard deviation of 115 for the congruent trials. And for the incongruent trials, I used the – so, I created 60 – eh 30 data points per participants for the congruent trials and then for the incongruent trials I matched each data point and added 50 - a random number with a mean of 50 and a standard deviation of 100 to it. So, that it would be correlated but also higher. So that is what I did, so that is how I fabricated the data.

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

P: Yeah, so, I found that the standard deviation ... - yeah, it was basically in the same way, because I created data points for each color for each participant in each condition. So, I used the average of all those 30 data points and it also gave me a standard deviation.

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

P: Yes, so I tried a few different methods to generate random data that was correlated but also on average 50 milliseconds higher in the incongruent condition than in the congruent condition. Because I used different methods to try it – also to find out what works for me and what brought me the nicest results. So, that already gave some different options and different tries and in the end I finally settled with one method and sometimes I didn't like the results and then I pressed again, again, again, until the results looked like what I wanted. So, for

instance, one time I had one data point with a standard deviation of 10, which I thought would be way too low for someone, so then I pressed again. I am not skilled enough to automatize it so that it would be always above 50, for example, the standard deviation. So, I pressed again a lot.

I: Ok. So, you said that you tried out different methods.

P: Yeah.

I: Could you describe the different methods that you used?

P: I don't think the different methods are very relevant, because it is just because I am new at simulating data, so it is basically the same method, but different formulas, so different ways of creating correlated random numbers. So, I don't think it is relevant for this purpose.

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

P: Yeah, so I looked at the means. So, I created – yeah, sorry, I created many data points and then I created an average based on those and then based on those I had a formula or a box that created the average of the average of all participants. And I just looked at if there was a mean difference and if the standard deviations looked ok, so for example when it was 10, it was way too low, of course, for [?] participant. So, just visually, and then in the end, I still adjusted some numbers and then I checked in SPSS if the main effect and correlations changed in a way that I didn't want.

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

P: Yeah, so just what I said. Like if the means were really off or if the standard deviation was really off, then that looks suspicious and not correct, yeah.

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

P: Yeah, but that is difficult. Because what makes a data set look genuine? If you have paper for that, I would be interested to read it, but no, I have no idea how to check that, no.

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

P: I think maybe like 100.

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

P: Yes. So, I tried to fabricate the data in Excel to get all the data points. And then afterwards I transported them to SPSS to do some analyses. And then when they were correct, I transported them to the template.

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

P: Yeah. In Excel, I used the random number generator - or I don't if it is a random number generator - to fabricate the data points, yes.

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

P: I am not sure. I did check real data and that is what I based my random data points on, but I didn't use real data points. So, yes and no.

I: Ok. So, you used it as like a general inspiration for your simulation, but you did not like copy-paste some of the data points or so?

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, not really, no.

**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, the first question is: Did you consider fabricating these data a difficult task to complete?

P: Sorry, can you ask the question again?

I: Sure. Did you consider fabricating these data a difficult task to complete?

P: I don't know. I think, yes, but also because I made it more complicated for myself, because I also wanted to get something out of it by learning how to best simulate data. So, yes, yeah.

I: Ok. But you think like one could also do it in a much easier way?

P: Ja, because I really thought – so, normally you have 500 participants. So then if you do a Stroop task, you cannot do it by hand. And you have more than 30 trials per condition, so you cannot do it by hand. So, simulation is really the way to go. Also if I want to use it for myself – not to fabricate data, but to simulate data. So, I really wanted to learn how to do that, but then for this data set you only needed 50 data points per participant. I could have done it all by hand, then it would be ... Yeah, I could have just written down numbers that seemed plausible to me. So, then, it would be less effort, I could be one in 10 minutes. So, I don't think – I think it could be easy, but I made it a bit difficult for myself.

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

P: That is an empirical question. I hope so, but also I hope not. I don't know. It would be nice for myself if you couldn't detect it. But I think, I hope that you will be able to detect it. Yeah, so I don't know.

I: Can you think of ways how like the team of researchers might be able to detect it?

P: I think maybe look at the distribution. What does the Stroop distribution normally look like, because I didn't check it. And all the things that you may have in mind of how to check it, that I didn't mention would be good ways, because I didn't check for those things.

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

P: I think for multiple reasons. I think it is interesting that you are working on a way to detect fabricated data. I think that is nice. I don't know. I wanna say it is sort of important to sort of make sure that people do not commit fraud and if they do to sort of detect it to really build a better science. I also notice with my students that we always tell them, yeah you have to collect participants and if you fabricate your own data, we can detect it. So really recruit your participants. We use it as a way to scare them to be sure that it is real data and not fabricated data. And, it would be nice if we could actually sort of check in a way if data is fabricated or not. Not only for our students, but for everyone. At the same time, you would hope it is not necessary. So, you wouldn't want it to be important. So that is one of the ways – one of the reasons I wanted to participate because I do think that it is important. So maybe, getting better at detecting fabricated data. At the same time, I thought it was a nice challenge for myself, because I took some course on preregistration and I think it is a really cool way to set up your own research. But then in order to do a pre-registration, you should also register what analysis you are going to do and how you are going to do it. So, for me it was a nice challenge to sort of simulate the data that I wanted to see like where are the outliers, does the data look weird, how would I detect outliers, to really use it for my own research this data simulation approach which I have never done before. To really come up with the best sort of analysis plan, knowing whom I should include and exclude for the reasons what might be in the data even if my hypotheses would be confirmed, so I thought it was also nice sort of exercise for me sort of to train myself in how to do this. So, two reasons: One selfish and one for science.

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

P: Yes, with my friends, but not with people in the department.

I: Ok, and did these people help you in fabricating the data?

P: No, because they are not in science. They have no idea, they think it is weird

– as it is that people can spend more than 10 seconds analyzing data, so 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: Yeah, one thing is that maybe I really tried to do it by simulating data using sort of random numbers, but it was so difficult that I really – I changed numbers by hand and I – now that I think of it maybe that's the way that most people if they fabricate data that is how they would do it. They wouldn't do it with some fancy computer program, but just do it by hand maybe. So, yeah, so that is my last thought about it.

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: Yeah, there was one thing. Maybe not worth mentioning, but what struck me as weird is that when I do a paired t-test in SPSS, I get different results for my own data than if I do the – or if your template does the t-test. So, I was wondering why. And does – so, is that template looking at my data differently than I looked at it myself. So, I thought that was kind of weird.

**Follow-up on the last comment of the participant**

Chris Hartgerink checked this and it seems that the approach to the t-test is different. The online calculator takes just the means response latencies per participant per condition. The template spreadsheet, on the other hand, creates the pooled standard deviation per participant across conditions, and computes the Stroop effect for that individual. Subsequently, those Stroop effect scores are supplied to a t-test with H0: mu = 0 (difference is zero). This way of calculating the t-test was based on the script from Many Labs (around line 134), which computes the Stroop effect in this way.