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, then we will start with the first block now. 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 am [REDACTED].

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

P: I got my PhD in 2011, so that is 6 years ago.

I: Ok. And what is your field within psychology? With field, we mean for instance social psychology, cognitive psychology, and so on?

P: Social psychology is my field of origin but I do [REDACTED] in, so also cognitive kind of things. So, social cognition, the field.

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

P: Any experiments that include a Stroop task? No, not a literal Stroop task. But a lot of similar experiments.

I: Ok. And could you describe your knowledge or experience with the Stroop task a bit?

P: So, I know the Stroop task just from being one of the strongest effects and replicable effects in the cognitive literature. So, I know how it works, how it is supposed to work, I have never actually - I have programmed Stroop tasks but also for fun but also for assignments that I give out to students. So, I teach programming of experiments and data analysis, and so I understand the Stroop task quite well.

I: Ok. And so you have also done and taught like the analysis of Stroop task data?

P: Yes, yes.

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: For me, that would be R and Python.

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: Relative to my peers in my own field?

I: Yes. To other researchers/scientist in your field.

P: Ok. Then I would say it is relatively good. So, I would say 8.

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 would say a 7.

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: No, one day.

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

P: I wrote the script and that took about 20 minutes. And then actually fabricating the data - because it is simulated data - took 1 second.

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: I would say - 1 to 7? I would say 2.

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

P: Well, as I am sure we will talk about in a few minutes when we talk about the way I fabricated I used a drift diffusion model and I knew about this model beforehand, but that is because I read a lot about that model for analyses and now I used it in a generative way. So, it was not preparation for this specific research, but there is a lot of knowledge in there that I had that you don't have immediately if you just - So you need to be involved in those modelling procedures before you can actually use them to simulate data.

I: Ok. And how much experience do you have with this?

P: With modelling data?

I: Yeah.

P: I teach it. And I - so, I teach R and I teach modelling with R and I would love to use it more for my own research.

I: Ok. And did you read any literature on detecting data fabrication?

P: I know of several – [REDACTED], and those refer to a lot of way of detecting things - so I know about like p-curving, but I am not sure you can - that is something you can detect - that is not something you will detect in a single data set, but I understand things like granularity testing which is something you won't be able to detect in here as well. I followed the whole, all the reports of Uri Simonsohn, all these other people who are trying to detect whether there is fake data in play or not. The stuff on Jens Foerster, on Diederik Stapel, I have read all about that stuff, yeah.

I: Ok. And can you describe or name the methods that you read about?

P: So, some of them is about things being - the most general method - most general description, I guess, would be to show that some attributes of the findings, some statistical attributes, is very unlikely under the assumption of random sampling. That is usually, I guess, what they are looking for. And that can be any attribute where it is - and another way is seeing whether the actual combination of data or the data points are actually plausible data points given the setup or design of the study. These would be the general approaches, I think, that people take. And I have seen people take different - look at different attributes of statistical - or of data for different kinds of papers or different kinds of data fabrication.

I: Ok. And would you like to – or like can you name more specific methods that have been applied or . . . ?

P: GRIM, for instance.

I: Ok.

P: That is the one that does the granularity testing seeing whether the - given that there is some mean and you only have whole integer numbers that can – that could have been answered, then that means that some means are plausible, but some means are implausible, you cannot get them with any combination of whole numbers. So, then there is a problem. So, that is the GRIM test. That is one. And you want – you want one, right? Yeah, ok, good.

I: Yeah, but you can also name more methods like if you think that they influenced your approach to fabricating this data set or so?

P: In this case, no, no.

I: Or did you – so you already said that you are familiar with previous cases of data fabrication and how they had been detected. First, can you name all the

cases that you are familiar with and then . . . ?

P: All of them?

I: Sorry?

P: All of them?

I: So like if you say that you are familiar with a lot of them, then of course you don't have to name all of them. Maybe you can just name a few.

P: Ok. So, I am familiar with the Jens Foerster case where the data seemed to be way too linear than you would expect to happen purely based on chance. Well, Diederik Stapel, I read the book about how he fabricated his data and [REDACTED]. What else? I know about – I have read about Marc Hauser's - made some "mistakes" – I am quoting – using quotation marks – mistakes in the encoding or coding of video clips. What else? I think Smeesters, I have read about Smeesters stuff. It is mostly psychology or social psychology kind of people. I am reading now a lot about – I am not sure yet whether people are convinced yet that it is fraud, but about the Brian Wansink stuff who seems to have some weird approaches to beating (?) data into submissions. What else? . . .

I: Like could you say how you think that this knowledge influenced your approach to fabricating the data for this study?

P: Well, yeah, I think that these questions are very much related to people who actually tried to fake the data by hand. But given that I have just – I have used a model – generative model that we know that describes reaction times and accuracy of responses pretty well. So, for me, I did not take anything of this into account. I just made the assumption that if you can analyze reaction times and responses well – and that has been proven and validly with drift-diffusion modeling, then using it in a generative manner should also work given some parameters that it uses. And actually I think that this is a very interesting test because if this data set is detected as fake data, it means something for the validity of using drift-diffusion modelling of (?) modelling responses- latency data. So, it would actually be interesting to understand what it means for the method in itself if the data are detectable as fake.

I: Ok. And so like did you consider any other approaches to fabricating the data?

P: No, no. I wouldn't have said yes to this if I would have needed to really by hand – I don't like doing data analyses by hand, I want to use scripts for that. So, I also want to fake data with scripts, I guess. I am too lazy.

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.

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: In general or in this specific?

I: Yeah, like here it is more about in general or like if you say that you have too much then you can also like ...

P: Ok. So, I guess, we have already talked about the inclusion of impossible errors – or, sorry, of impossible mean values, given you know the granularity of the data that goes into the means. That is one thing. Another thing is, I guess, that – specifically with reaction times – you would expect it not to be normally distributed but to have a tail because people are – can go really slower, but they can't really go a lot faster than some specific speed. I would think that looking at the data themselves I guess you can estimate how likely it is that you get the data distributed as they are assuming that these should be distributed a certain way. Also, one thing actually by the way that I thought in specifically here was not clear in the instruction is whether the way the data are analyzed is including or not including the trials on which people make mistakes. So, I assumed that you just want all the trials but if you would also look at the raw data – now this is not really the raw data because you only get the response means of the participants – but I could have provided you with the raw data because this is what drift-diffusion models give you. So, for each trial instead of just the mean for each block, I could give the data for each trial that has been simulated and then you know, there you would even have more options to see whether the data are fake or not, because it will allow you to test all kinds of like impossible combinations of speed-accuracy-tradeoff, I guess, whether participants too similar in how they respond. But basically more in general, I think, it is all about whether the values are plausible given some assumptions about the world (?) – about randomness. And I think that the problem more generally about faking data is that humans are not good at generating random variables cognitively, so I guess that when people try to fake by hand, they copy-and-paste or they change some values but don't take into account how this changes the distribution or the likelihood of the data. And I guess most of the people who do that kind of fraud fake that data in that way. They are just not knowledgeable enough to actually know what are the features of fake data or what changes that make their data unlikely where a data detective might look at those things.

I: Ok.

P: Is this – am I still answering the question?

I: Yeah, yeah, absolutely.

P: I didn't have much sleep tonight, so ...

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

P: So, more genuine would be in general when the combinations of data or features of the data are not unlikely given that the assumption that these are normally distributed variables, for instance. So, in the extreme example, if something has – like for instance, let's say we are assuming that people make speed-accuracy tradeoffs. Then, showing a Stroop data set where people are both slow and make a lot of mistakes would be very implausible, not from a statistical point of view, but from a theoretical point of view. So that is one reason why something might be more or less genuine. So when it adheres to all this theoretical knowledge about how the Stroop task is performed, that would be one way to make data more plausible. The other way would be that - so this may be a point where it becomes repetitive, but would be where each value is possible given the granularity of the input and where the distributions follow the distributions that you would expect given a random process or in this case a process that would give you like a long tail in the distribution for the reaction times. These kind of things.

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

P: I took them into account by choosing to use a drift-diffusion model in order to simulate the data. But I didn't then also check, I just did one run of the model and used that whatever it was instead of , you know, running the model until I thought it looked plausible, which would have been another approach. But if I would have done that, then I guess that one way to - now that I think of it - would be like if by accident now I end up with a data set where the effect size is implausibly big, for instance. That would be a potential problem.

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

P: Well, I guess that the drift-diffusion model takes into account just more generally all kinds of assumptions about reaction times and response and how they relate to each other. And that is a fairly complex story about the model itself, but it is basically the model using something called Brownian motion which is a concept from physics where - Brownian motion is basically you have some kind of particle, and then you have even smaller particles that you can't see, they are really order of magnitude smaller, they all hit from random directions, because we can't explain where they are coming from, but they hit the bigger particle from random directions. So what you see is the particle will move around what looks like random ways - and this is called the drunken man's walk, right - and now, this concept has been used in drift-diffusion modelling by simulating lots of data based on several parameters that have to do with how fast a response accumulates given some input and given some signal in the data. It models how fast it accumulates to surpass some criterion to say one or the other response.

And this takes into account the way that the cognitive process might work in order to get to the answer of yes or no over time or in this case of a specific color, of naming a color. And I have used a very basic model here, the number of parameters that determine what this path looks like can differ from task to task and several tasks are modeled more - better with complex drift-diffusion models where there are lots of parameters and I chose a basic one based on the assumption that the Stroop is a pretty basic effect where there is not much - you know it is a very straightforward cognitive process. For me, that was the idea. So, that is why I used a basic model, which I am not sure whether that is correct thing to do or not. But that, I guess, is the kind of things that I took into account while simulating or faking the data.

I: Ok and what kind of parameters did you specify in the model?

P: It is easiest if I show you the script, right. I can give you the script if you want.

I: Yeah, that would be great, yeah.

P: Let me see. It is actually pretty straightforward. Ok, so, here, if you give me your USB stick later on you can see it. But here, this is the whole script. So, here we have the 25 participants, here we have the number of trials, here is a function to simulate a single trial where you have as parameters the strength of the signal, a constant that is added to the drift, it is like a bias whether some people are, you know, biased to say one color or the other in this case. This is another bias parameter, the starting point, then there is the noise is (?) how much of variance accumalates per unit of time, and there are some other parameters, here bound, small time steps, it's the granularity of, you know, how much if you make the time more granular, so if there is a higher resolution of time, there is more randomness added to it. And then there are some other like fuzz (?), response time fuzz (?), I don't know (?) the number of steps, it is not even - there are not even that many parameters actually. And then this is actually doing the work here, this is the full simulation of one single trial. Here, you see some extra - oh, yeah, this is from an old script - this is not informational. So, I used this drift-diffusion model to create fake data to simulate data for students to then practice their analysis skills on. And then because of using this model technique I can, you know, choose the way I want, what type of data do I want to produce, do I want a small or large effect, do I want an interaction or not or anything like it is very easy to do that when you - and then you can look for special cases where you find specific outliers or whatever and you want people to train on that. So, potentially, being versatile in modelling and in R also makes you a very dangerous data faker, I guess. But people just have to trust me. But here you can have the full script if you want to. And I can actually run it for you if you want. So here we have simulated a data where we have a t-value - so this is a very strong effect now. And this is the congruent, these are the incongruent reaction times and, well, you can do so. Like here, I can just go on simulating data sets. I have simulated ten now.

I: And does this also allow you to like take into account specific relationships for instance, between the means and the standard deviations or correlations within subjects etc.?

P: It would possibly, but in this script I didn't do anything with that. So, actually, every subject is now treated - actually, these data are treated as just one single giant subject doing 60 times 25 trials and then I just, you know, took subsets from that to create single subject data. So in this case, there are correlations between subjects to the extent and within subjects to the extent that there are - let me see, am I correct in the way that I am saying this, I did this a month ago when you guys asked me to do this - yeah, it is correct what I am saying. So, basically, there is - I didn't take into account any correlations within subjects or between subjects that there should be.

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

P: No criteria. So, my feel - so, my assumption was given that it is produced by this drift-diffusion model which - that it should be better than I can myself fabricate. But I realized now that we are talking that perhaps like the difference signal strengths or the different parameters that I chose might show up as being implausible for a typical Stroop data set. I don't know that. And given that maybe I have not simulated it in terms of enough complexity including you know correlations within and between subjects, maybe that is something that will show up. I don't know. We will see. Do we get to hear that actually as participants how high you ranked in . . . ?

I: Yeah, we can talk about this after the interview.

P: Ok, good.

I: So like or did you have any sort of checklist for the means and standard deviations or so?

P: No.

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

P: Well, I wouldn't have done a lot of more than I did now given the time that I wanted to put in it, but perhaps the - one thing would be the correlations we just talked about, within-subjects correlations probably. I can actually check what happens here. So here with the correlation between congruent and incongruent in that case, let's see, yeah, but we want to . . . yeah, we can check that at least at the mean level for this specific - this is not the simulation that I gave you, so this is just another iteration. Yeah, that would not be very plausible, I guess, if you have a negative correlation between congruent and incongruent reaction times. So, that would be something that I would have had looked at. Although you know who knows maybe if I run this again I don't know what actually happens in my simulated data set, because I didn't look at it. So, that is one

thing that I would have needed to check. So, I am becoming less confident about not being found out. And then another thing that I think now that I might have not looked at is how actually - I didn't check at all whether the average and standard deviations are kind of representative of typical Stroop effects. I didn't check that either. But I guess that could be sampling error, right? Yeah.

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

P: No.

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

I: Ok. Then, we will now start with the fourth block. The goal of this block is to get some information about the specific steps of the data fabrication process. So, could you indicate what steps you took to fabricate the means for the participants?

P: Yeah, actually I just explained it by saying we used a drift-diffusion model with - I used a drift-diffusion model with several assumptions on the parameters which are all actually pretty basic. In the end, the only thing that deviated from what you would do by default is I made the decision about how strong the signal would be given a congruent trial or an incongruent trial, these are these values actually here.

I: Can you name them for the . . . ?

P: Yeah, so 2.6 and 2.0. 2.6 is the amount of signal that people see in the stimulus if (?) it is an incongruent trail. And 2.0 is the amount of signal they see in terms of congruent trial. That is actually the only thing that then - these are usually parameters that you are trying to model, but now I gave them as input and then based on that the model - the drift-diffusion can generate the data.

I: And how did you come up with these values?

P: Because I have simulated more drift - more data with this drift-diffusion model. I knew already intuitively that this would give me a strong effect as I assumed the Stroop effect would be. Now, what I could have done is I could have played around with not necessarily with the difference between these numbers but with how high these numbers would be. And that would give me faster or slower reaction times. And I didn't do that.

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

P: Yeah, they follow from the model, yeah.

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

P: No, so I just hit run once and I used that. But I could have just keep on reiterating the whole simulation until there was one that I thought looked plausible. But I didn't want to put time in that.

I: Ok. So you didn't determine whether you were satisfied with the fabricated data or that they needed to be adjusted or so?

P: No.

I: Or you inspected whether the fabricated data looked weird or genuine or so?

P: Well, I actually also didn't do that. I just made the - for me, it was really interesting to see what would happen if you have just one run with drift-diffusion models which should be generating plausible data - whether that indeed would follow to be plausible given your procedure to detect it. But I see something, when I look at this now, at this template file, is that because you asked about standard deviation, I hadn't pay much attention to it, but that for instance in the incongruent condition, the standard deviation are larger than in the congruent condition, and I think you would expect this. Also given how reaction times behave, because you cannot - because when you go more towards fast reaction times also the spread will truncate, because there is - on the left there is kind of a minimum speed that you need to have. And also I guess when a block gets difficult sometime you are fast, because it is still easy for you, easy trial, but sometimes it is a very difficult trial, so you get larger dispersion while with the congruent stuff you might get - you might - you get faster on every trial, because it is all easy for you. So, I - it is not something that I took into account while I was fabricating, but the model gave me that. So now, I think that actually looks kind of nice.

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

P: Just one.

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

P: Yeah, so R.

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

P: Yeah. This is doing that, yeah.

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

P: No.

I: Ok, so but like were some of the parameters inspired by like papers you looked at or that you know from your experience with the Stroop task or so?

P: No. Well, yeah, I mean, yeah, I got knowledge about the Stroop task, that it is a strong effect, so the parameters that I chose were inspired by that it is supposed to show a strong effect.

I: Ok. And like how you selected like the average mean or the . . . ?

P: No, yeah, I didn't pay attention to that. I should have actually to have a better chance at winning this game, but I didn't.

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, except for the fact maybe that I should also give attribution that parts of this stuff, like this drift-diffusion-model [?] is just available on the internet and I used a lots of default stuff from there.

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: No.

I: Ok. Do you think that it would difficult if you would take more complexity into account?

P: Yeah, for sure, for sure, yeah. Although not as difficult as it would be when I do it manually, right. I would not know whether to start actually. Because if I would do it manually I would probably already have in terms of being able to detect it I would already probably make so many mistakes that it would be detectable compared to - I would - you know, the more knowledge you have about statistics, I guess, the more you know what should go right in order to be plausibe. And if you - I think manually faking is only - you can only do that when you don't know a lot about statistics, otherwise you go crazy. So if you know a lot about statistics you can use this model, but then you know you ask questions about, yeah, well, but how plausible - how valid is this model? And that would become a PhD project in its own, right, because that means you are actually investigating the validity of the model, which would be a really interesting methodological project, but you know for this specific task I wouldn't put that much effort in it.

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

P: I thought, it would. I became a little bit less confident given that we found out that maybe I should have paid attention to a lot - some of the complexities of the Stroop task more. But I think generally that the approach of using a

drift-diffusion model if you use the right parameters should be way less detectable than more straightforward techniques, for instance manual fabrication, yeah.

I: And why do you think so?

P: Well, yeah, that would repeat the answer to the earlier question. If you do things manually, there is just too much to keep into account and basically you will probably make mistakes on some dimensions that you were not aware of. And these models are supposed to generate plausible data. That is why they are used to explain reaction time data.

I: Ok. And like considering your data set like can you think of ways how you could detect it as fabricated?

P: Well, one thing that I thought about beforehand is that maybe if the - because this is based on random number generators and true random number generators don't exist in computers, that maybe in some way it could be detected. And I think the way to detect it would be then if that is true, given that this is generated with a drift-diffusion model, if you can then use a drift-diffusion model to analyze it and explain 100 % of the variance, which should be possible given that it is first generated with it that it also comes up with it, if that happens then that is too good to be true, right. So then maybe you can draw conclusion that there is a high likelihood that it has been generated by a drift-diffusion model.

I: Ok.

P: Maybe. I don't know, I have never tried this. I should actually do it.

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

P: I think it is important work to find out. I have - in the past couple of years, I got a very dark view of our field and I think it is a very good idea to be able to find out ways to quantify - well quantify is not a good - to well to detect fraud actually. And although I think that statistical methods will never be the only - it will never be the nail in the coffin, right, the final nail in the coffin - it is a good way to start looking at papers that are already published and see which one we should or should not trust or who we should further investigate. And I have seen too many people whom I have trusted already - or were big names in the field, I am nut sure whether you can see I trusted them, but who were big, big names in the field that turned out to have fabricated data. And one is already too many actually, but there are more than one. And this shows to me that we cannot rely on just that people say that they are trustworthy, right. So, I think it is a good way to start looking, but I also think it is very dangerous because it has to be - all these methods, these automatized methods of figuring out data fraud need to be validated well if you ever want to really use them to detect fraud. Because otherwise you could make, you know, type I errors and call someone guilty who is not guilty. And so that's why I thought it is important to contribute to this study, because the better it gets validated now, the better these tools are going to be when they are actually needed. That being said, I also thought there is an

ethical dilemma here, because I also thought that maybe it could lower the bar for people who might have less of a [?] to actually fabricate data once they have done it once. And so for people who use drift-diffusion modeling or anything like I do, it is already so easy to fabricate data that the only thing holding you back from doing it is just because you think, well, I am a scientist, I care about the truth, right, but I also know how easy it could be. So, I can - and that for me trumps the ethics part that I think it needs to be found out. And I actually hope that this can be detected, but it is also problematic - and I hope that that is because I used implausible parameters, but it is also a problem for the - if I didn't use implausible parameters, then this is a problem for this specific method actually. Not necessarily for this method of faking data, but for this method of analyzing reaction time data that this model doesn't fit. So, I thought that was an additional reason for me to participate in this study to see what happens with this specific method when you try to detect it as faking data. So yeah, it is a little bit of an incoherent story maybe but ...

I: No.

P: No, no? Ok.

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

P: Well, yeah, I told them that I would do that. Not participants in this study. They are not participants in this study.

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

P: Some of them offered thoughts on what a plausible Stroop data set would look like. But I said that I don't really care about this because I have this model that is going to do everything for me.

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

P: No.

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

P: No. I am going to give you the script and then you can see all the details in the script, right. There is a - it is all documented, so it is really open science. And no, no.