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: Now we will start with the first block. The goal of this block is to get some general information about you. Are you a PhD Student?

P: Yes, I am.

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

P: I work at the department of clinical psychology, but I would characterize my research more in the field of between cognitive and forensic psychology.

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

P: Not a Stroop task, but I am quite familiar with reaction time tasks.

I: Ok. Could you describe a bit more like what (a) is your like knowledge or experience with the Stroop task and then in addition to elaborate a bit more on your other familiarity with the reaction time tasks?

P: Yes. So, my experience or familiarity with the Stroop task is, I would say, the general familiarity that every person who studies psychology should have. So, I am aware of what the effect is, also where I hope how this effect comes about. Other than that, I have not conducted - I have participated in Stroop task for sure when I used to - when I was a Bachelor student and I had to participate in research, then there were some Stroop tasks that I participated in, but I have never collected Stroop task myself. So, this would be my answer to the first part. And the second part is, my experience with reaction time based paradigms is predominantly in the domain of [REDACTED]. And especially - maybe you will come back to this later, but the analytical procedure when it comes to aggregating the raw reaction times to one participant and then get the means for example for the congruent and incongruent trials is very similar to the Stroop task.

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 think I use both Python and R daily. I can use SPSS, but I prefer not to.

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: With my peers being fellow PhD students?

I: Yours peers being researchers/scientists in your field.

P: I would say maybe an 8, because I specialized - when I did my Bachelor I specialized in psychological methods and statistics. So, I would say it is more than the average, but I would not compare myself to a professor in psychological methods, because they obviously know much more.

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 hope you won't detect them, but that is just - ok, I would say I - [?] I would say 8, so I am quite confident that you would (?) not detect it.

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

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

P: I spread it over several days. I begin thinking how I could do it, begin writing the code for this, and then let it rest for a couple of days, then look back. So, it is - it took in total three weeks - I spread it over two or three week, the whole process of fabricating the data.

I: Ok. And on how many days during this time did you work on fabricating the data?

P: I would say 4 or 5 days - I spent - yeah on 4 or 5 days I worked on it, yeah.

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

P: In total in hours or?

I: Yes.

P: I think maybe every day 2 or - 2 hours or. I think in total maybe between 4 and 5 hours.

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: 7, a lot of effort.

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

P: Well, the preparation was mainly on finding similarities to analyses that I have done before with real data that we collected with participants. And then I - so, the preparation was mainly coming up with a concept of how I can fabricate it and thinking about how I could, for example, add noise to the data and stuff like that. So, this was more the thinking part and preparation part and then I looked - although I did this at a later stage. I also looked at some studies that used the Stroop task to see what actually is a - because as I said, I have not used the Stroop task myself for my own research. This is why I thought I should know at least a bit of what range realistic reaction times in the Stroop task fall into.

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

P: Maybe 1 hour. 1 to 2 hours maybe in total.

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

P: No.

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

P: No, I did not look into them explicitly, but I am aware of a couple of the well-known cases, particular the Jens Foerster case, where I think one of the indicators was the - I should (?) say the linearity in the effects that were reported that were a bit suspicious. So based on this, I was convinced that right from the start I should use a simulation approach to simulate my data rather than just - or to simulate my data with a program like R for example rather than coming up with a data on my own. Because I thought I myself would be too - how shall I say - too non-random in coming up with things. So, then I myself would have maybe the linearity in my data that is then based on the notion that they found in the Jens Foerster case, for example. So, this is what - how I used previous knowledge on data fabrication for this, but I didn't really read in depth about the work of the guys who detect it or who came with the idea that the data was fabricated in the Foerster case.

I: Ok. And so you said that like your preparation process basically consists of two different parts so that you first looked up previous papers on the Stroop task and then that you thought about your approach how you would like to do it. Could you first - or could you describe both in a bit more detail what exactly you looked up or what you thought about and then how this preparation influenced your approach to fabricating the data?

P: Yeah, so the order was a bit different. So, I first came up with an approach of how can I actually do this technically, so how can I - if I begin with this, how - I thought, ok, how can I do it? I know that the idea was, there were 25 participants in this case it is within-subjects, so then you have the congruent

and the incongruent trials and the Stroop task and each subject got 30 trials. So, I thought, ok, I must rather than just coming up with the means and standard deviations that I need to report for the incongruent and congruent ones, I should simulate the data in a higher level data meaning simulating all the trials for each individual in the both conditions. So, that was the first consideration. The second consideration was how - when - after - while simulating the data, how can I actually blur my simulation as much as possible and make my own influence in the simulation procedure as small as possible. So, this is why I added the second layer of randomness to this where I said, ok, I need - I want to give a range of possible values that I want to use for my other simulation then, so this is why I had a first random draw on saying, ok, what - if you have (?) a range, from this range take for each individual one value that falls into the range and simulate data based on this value. And then as a third step in the simulation, I said, ok, how can I add noise to this and again there are this step-wise procedure I had. First, a range of noise parameters and then I drew from this range of noise parameters and put them into another random generator that added the noise and then I added them up. So, these were the three main considerations in simulating the data and coming back to your first question with the looking into previous literature I would say I should have - or not I should have, I could have done this more thoroughly. First - my first thought was I just look up meta-analyses on the Stroop, but what I was mainly interested in is just what are the realistic reaction time ranges they fall in, because for the paradigms where I work in reaction time differences are much smaller than for example for the Stroop task. So, I thought ok if I now come just with some difference that shows ok incongruent ones take longer than congruent ones the most obvious - or I thought one of the most obvious things that you could detect my fabrication with is if I just have unrealistic reaction times for the Stroop task. So, basically, I don't know my domain and just fabricate data and get a significant difference, but just completely off from what Stroop task shows. And if I remember correctly, I looked into, I think, just one article that I found - it is very bad - via the Wikipedia page of the Stroop effect and looked into - it was one of the very first publications on the Stroop task where they measured reaction times and then I looked this paper up and looked what reaction times do they actually report. Just to get roughly an idea of what the means of the reactions times for congruent and incongruent ones could be.

I: Ok. And did you also consider different approaches to fabricating the data?

P: I was aware that there was one approach would just be to come up with the means for the - so the aggregated means for each individual, but I thought in the beginning (?) and I knew that this was an - could have been another way to do it, but I never considered doing this, because I thought if I want to make it as difficult as possible for you to detect it, I should go to the length and really fabricate data on the stimulus level or the trial level and then aggregate it from there. And other than that I - no, this was the other only alternative I thought of.

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

P: Let me think, 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: So, the - I would say data that has some strange things in it is more - data that have some random - how should I say this - some strange observations in them are more likely to be genuine, this is what I would say. So, if you got data that were super - let's say you have data that is perfectly normally distributed with a - for a relatively small number of observations, this is what I would say, yeah, this is a bit strange, because you would never get this really in real-life-data. So, one criterion for how should I say good fabricated data in my opinion should be that there are some strange things going on. For example that you don't observe that for each participant there is a big difference between congruent and incongruent trials. Because you have some participants for which there maybe is a small difference or it goes the opposite direction, so this [?] some apparently counter-intuitive stuff going on in your data, this is for me what makes good fabricated data.

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

P: Yeah, similar to what I said before. Genuine I would say if there are some odd things going on, maybe even you have missing data, which in this case I have not included this, because you asked for the 25 participants. But something like this, you have missing data, you have strange observations, you have some data is not recorded properly, some - you have some outliers in your data. Stuff like this, which really from a researcher perspective is not really nice if you have it, but which you always have. This is what I would say is a characteristic of genuine data.

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

P: Yeah, one of - this was one of the reasons why I added noise to my data to make - to blur the underlying random generation process as much as possible and to get some strange observations into my data.

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

P: So, one thing that I did not include but which I should have included is the - I thought a lot about this and then in the end I think I did not have the time to do it is the - I would expect to be there some correlation between the two - between the congruent and incongruent trials, since this is a within-subjects design. So, ideally I should have accounted for that as well, which I did not do explicitly - or not do at all. This is something that I should have done, I think. The other thing is that I could also have done (?) about a realistic effect size that I would want to yield with my fabricated data that should ideally also be realistic (?) of what is commonly found in the Stroop task, which I didn't do either. I did it on a more superficial level, I just looked at what are the realistic ranges of what congruent and incongruent reaction times fall into and then took them as some indicators on where to start my random number generation process. The - ideally you could have said, ok let's say meta-analyses showed that the just say the mean Stroop effect is for Cohen's f effect size of I don't know .15, then ideally I could have said, ok, generate data that yields an effect size in the difference between congruent and incongruent trials approximately in that range of the meta-analytical findings. But I didn't do that either.

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

P: What - why I thought they would go undetected or?

I: Yeah, so like criteria that you used to see at the end whether like your fabricated data would be like hard to detect or?

P: Yeah, so mainly the layered random generation process and the noise that I added to the data.

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

P: For congruent and incongruent trials or?

I: Well, like separated for . . .

P: Oh, yeah.

I: . . . first the fabrication of the means and then the fabrication of the standard deviations.

P: Let me check . . . let me just check here my script . . . no from what I - I can also give you the code that I used, but from what I remember here I came up with - I simulated the means only - not the means, the reaction times per trial, so the reaction times for 30 trials for each individual, 30 trials congruent and 30 trials incongruent, and then in order to obtain the mean - the aggregated mean that I reported in the Excel file is just the aggregated mean and the standard deviation is the aggregated standard deviation based on the raw reaction times, so I did not simulate the standard deviations separately, they would just come from the reaction time simulation.

I: Ok. And did you have criteria that you sort of used as a checklist to see whether this would make sense as a like sort of genuine fabricated data?

P: I did not use an explicit checklist, but just that I - mainly what I used is as I said before I used some indication of what the Stroop reaction time difference could look like in order to not come up with data that maybe shows a significant difference but just reports reaction times that are so off the Stroop task that someone looking at this would just say he obviously doesn't know what the Stroop task is. So, this was the main criterion that guided how - what kind of values I used to generate the data.

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

P: Yeah, as I said, I should have accounted for the within-subjects nature more by adding some correlation between the two types of trials and I should - could maybe have paid more attention to what actually an effect size of the Stroop task is or what magnitude the effect of the Stroop task is and then generate the data based on this.

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: No, other than what you would probably be able to see in the code that I used to fabricate it that as I said I did not simulate the standard deviations but where I added some randomness for the standard deviations is in the random generation functions, so for example when I generated data from a normal distribution where I then feed in the number of observations that I want, the mean which in this case came from a sequence where I drew the random mean from, and the standard deviation of the data from the normal distribution that I simulated. The standard deviation came as well from a sequence of what I thought were possible standard deviations for the congruent or incongruent trials and they then generated the reaction time data, so there was some way of how I fed some standard deviation into the random number generation process, but I did not generate the standard deviation of each individual.

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. So, let me just check, I can maybe walk through this script quickly. I came up with some random - with some indications of what the mean could be for the random generation. And how I did it is I came up with a plausible maximum for congruent trials, a plausible minimum for congruent trials, same

for incongruent trials. Then I said, these - I did this for the means and standard deviations of the normal distributions that I would later simulate my data with. And with those maximum and minimum values I created a sequence with certain steps in it and from the sequence I drew randomly - for each participant I drew one number for 30 congruent trials, one for the 30 incongruent trials, based on this number, the mean and the standard deviation I generated by two random processes reaction times for the congruent and incongruent trials for each participant. After I had done that, I added noise to the data, so I had - for each participant I had a participant ID, I had the trial type, and a reaction time. And then I added noise to this and I again generated - came up with two different ranges of how to generate the noise for the data, so I had - I came up with the variables I called noise-congruent- and noise-incongruent-trials which consisted of - here of a sequence from 5 - from specific values you can look up (?) from which I drew random one number. That number I then put into a random Poisson generator, generated 1000 observations from a Poisson distribution with a lambda parameter of the noise value, then I multiplied this randomly with +1 or -1 to add or subtract from the reaction time as a noise parameter and then added this to the raw reaction time. So, I did this for congruent and incongruent trials and then the only other steps were that I then had for each participant 30 observations, 30 trials congruent, 30 trials incongruent, and the according reaction times and then I aggregated this for the means and the standard deviations.

I: Ok. So, you mentioned that you had different types of distributions in the different steps. What was your reasoning behind using those specific distributions?

P: I used the - the main generation was based on a - generated random data from a normal distribution where I fed in again randomized parameters, but so the idea for this was that I looked at some graphs for the - I think it was in the paper that I mentioned previously from the Stroop task or maybe it was some - a read-off (?) slides for a lecture somewhere about the Stroop, where they looked approximately random, aeh normally distributed. So, I thought ok if they are normally - I used a normal distribution as a sort of initial point where I start and then add some noise to this. And in order to not add the noise in a random fashion as well I just said, ok, then I just use the Poisson distribution with a highly skewed characteristic and draw randomly from a Poisson distribution which again I fed into with a random value for the lambda parameter. And then in order to blur this even further I said ok, it would be nice to do both to add some noise really in terms of adding something up to the reaction times and to include noise by subtracting something from the generated reaction time. Because otherwise - so, in this case the noise was really more - really something just to blur it rather than something that comes on top of the reaction time, because it could go either way, either increase the reaction time for each trial slightly or decrease it slightly.

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

P: Oh yeah, as I said before the standard deviations were not really fabricated - oh yeah, they were - everything was fabricated - but they were not as fabricated as the means in the way that the standard deviations were really just the aggregations or - so, the standard deviations were based on the reaction times that I generated.

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

P: I initialized my random - the whole random generation process by seeding the randomization, so that you could hopefully replicate exactly the way I fabricated my data. Well, I did this a couple of times until I got the - because the random nature and all the random layers or the layers of randomness that I included implied that sometimes you could find larger or smaller differences between them, so I looked at this and I said, ok, I don't want them - a massive difference but I don't want a very, very small or just marginally significant. So, I did this a couple of - not too often really, I think maybe 3 or 4 times and then - because it just goes a couple of seconds and then I thought, ok, this is a fine random initialization (?) parameter and then I just sticked to this and this is what I reported in the end.

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

P: I based it mainly on the t-statistic and on the p-value and again I should have - actually should have used effect sizes to get standardized indication of the magnitude of the difference, but the - I looked at the t-statistic and p-value and there were cases, I think, in of the incidences where I generated all of it, it happened that I only got a p-value that was marginally significant, it said .039 something like this and I found this too unusual for the Stroop, but just again this is more the intuition that I had, it is not that I based this on empirical observations really what the p-value or again effect size should look like. Because effect size I didn't include this at all. And there were other cases where the t-statistic was so high and accordingly the p-value was very, very small and I thought, no, this is maybe too obvious. So, I tried to find a middle ground between not just significant but also not being super significant - if - yeah in a way this is what I would say.

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

P: I looked at the data after I fabricated it, but I did not inspect it with a step-wise procedure. I looked at it and I was quite satisfied because I found exactly what I intended to do. That for some participants, the difference was very small or maybe even not existing or even went the other way so that the reaction time in the incongruent trials was faster than for congruent trials, which is a bit odd but ok, I thought this could happen. So, I was satisfied with it after I looked at it and saw, ok, there are some odd things going on. So then again this was more for me an indication ok the fabrication seems to have gone ok rather than just saying oh no there seem to be strange things going on so I must get data that looks cleaner, so the main idea behind this was if it looks strange

this is for me an - or if it looks at first sight a bit strange when (?) you think if you generate the data you should come with this perfect nice effect throughout everything is consistent. And then I thought, no, you must add some strange things into the data in order for it to resemble more real-life data.

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

P: Again, I did not use a checklist. I inspected it, yeah, I inspected whether it was genuine just by looking what I know or observed from my own research with (?) comparable reaction time data where you have some individuals where you have very strong differences somewhere you don't have strong differences, others where you have the complete opposite of the differences. And once I found that there are some observations in the fabricated data that I made that are not all going as nicely as someone would expect they go. So, well, I thought, ok, they are more in line with real data that I have observed in my own research with the strange things and oddities in it then I thought, ok, this looks genuine.

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

P: So, for each individual I fabricated all the trials, but if I just look at the aggregated stuff, so basically it comes down to how many runs I did for my simulation script until I was satisfied and then as I said before it was probably 4 or 5 runs of it and then I was satisfied.

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

P: Yeah, I used R to generate all the random observations.

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

P: Yes, for all steps I used several layers of randomization in it.

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

P: No, no, there was, no, I used the indications of what the means could look like, but there was no real data involved, no.

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

P: No, I don't.

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

I: Then, we will now start with the fifth and final block. The goal of this block is to get some information about the underlying rationale of the data fabrication

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

P: Well, yes, so I think there is some irony in it. I would say at least the - what I would expect the quite extensive length that I went through for fabricating my data set I thought why would someone fabricate the data if it takes so long, I could just collect the data myself, right. So, if I would have just programmed the Stroop task, put it out there on Amazon Mechanical Turk, some online platform, and collected the data, this would probably have been quicker than coming up with all the fabrication steps. So yes I would say it is difficult compared to just collecting the data, but this, I would say, depends on the way how you fabricate the data. In my case, I tried to put (?) a lot of randomness in it for all the trials. Maybe if I just sat down, came up with 30 - with 100 values for the 25 individuals in total and just spent 1 hour coming up with some random numbers in my head, this would probably be much easier than what I did. And then it would be easier compared to collecting the data. But in my case I would say you might just as well have collected the data.

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

P: Well, it depends on how you will do it, but I would say it is definitely more difficult in the fine details compared to someone just sitting down and just coming up with the values in their head.

I: And could you think of ways how it would be possible to detect your fabrication?

P: Yeah, I thought about this. I thought about how would you try to detect my data fabrication when I generated the data. I thought one way would be - and this might actually be the K.O. criterion for my data from the start - to look at does this person actually know their domain. So, have I actually come up with some values that are realistic for Stroop task? I hope to counter this by looking up what values could be reasonable for a Stroop task, but this would be the first thing that I thought, ok, this is how they could detect my fabrication and I hope to have countered this. The second thing is what I have mentioned before is the within-subjects correlation between the congruent and incongruent trials, I thought if you assume this within-subjects design implies some correlation between the observations of congruent and incongruent trials, you could detect that I did not include - not only that I did not include this, but that this is not in my data, so that this is not something happening in my data. And if this happens in real Stroop data, this could be an indication that the data is not generated by the real, let's say cognitive mechanism behind it.

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

P: I think it is really interesting and I think it is necessary to come up with solutions that try to counter this issue of data fabrication in a hopefully automated way so that you maybe in the future have machine learning classifiers that you can run on data and say ok, maybe this data - or at least this data

looks suspicious to make it more difficult - or to increase the chances of someone doing it of being detected. I thought, yeah that would be nice to help with this to do this.

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

P: I talked about it with a colleague of mine, but I did not discuss any specific procedures on how we - I don't know whether [REDACTED] participated in this study actually - but on how I did it. I did not discuss specifics.

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

P: No, I did this all on my own.

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: Let me think, no.

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

P: Let me think. No, no, I can't.