

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

P: Yes.

I: And - maybe, you can come a bit closer to the ...

P: Yes, I am - I am a PhD student.

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

P: Methods and statistics.

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

P: No, I never did the Stroop task.

I: Ok. But could you describe a bit like your knowledge about it or the experience that you have with it?

P: I have learned about the Stroop task so I know what it means. And, well, once as a participant for a study, I performed the Stroop task. But I never administered the Stroop task to anybody.

I: Ok. And so you have learned about it during your studies did you also like read papers about it or so?

P: Well, for this experiment, for fabricating data, I looked it up on the internet to remind myself of what it was. But I didn't really look at studies.

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: Once a week would be Matlab and probably FSL which is a fMRI statistical language program.

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: Which kind of peers do you mean?

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

P: Ah in my field. 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, extremely insecure, to 10, extremely confident.

P: 3.

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, did you fabricate the data in one day or spread the data fabrication over several days?

P: I did it today in one day.

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

P: 1 hour and 15 minutes.

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: From 1 to 7, it would be 4.

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

P: Yes. I downloaded an empirical data set from the internet. And I thought beforehand on which characteristics I should keep in mind when fabricating the data.

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

P: 1 hour.

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: Ok. So you said that you thought about like different ways how to approach it or specific characteristics - could you elaborate a bit more on like what ideas you had?

P: I looked at - So, we had to fabricate four variables, two mean values for 25 participants and 2 standard deviations for 25 participants. For each of these 4 variables, I looked at characteristics such as the mean of that value, the variance of that value, and the correlations between the 4 variables. And I used this information for my simulation. And I also - I based the mean values, the variances, and the covariance upon one empirical data set that I downloaded from the internet. And, yes, so, and I also looked at outliers, because there was one outlier in the empirical data set that had a huge impact on for example the correlation between two variables. And when I looked at the scatterplot I saw that I should probably remove this case. So, I removed one case from the empirical data set and used the remaining of the empirical data set to - for the basis of my fabricated data set. So the means, the variances, and the covariances.

I: Ok. And how did you find this empirical dataset?

P: I just googled 'Stroop data set' and I found one on the open fMRI site. So, it's people in the fMRI scanner performed the Stroop task.

I: And how did this preparation influence your approach to fabricating the data?

P: Sorry.

I: How did your preparation influence your approach to fabricating the data?

P: A lot. So, I actually based everything on this preparation. So, the means, the variances, the covariances were all based on the preparation.

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: Well, I have written it down in a script. So, here I see: So, first, I downloaded the data set, then I inspected the empirical data set by looking at histograms and scatter plots and looking at means, variances, and covariances, and the ranges also. Then, I removed one outlier. And then I started fabricating the data set. ... It's actually the same as what we just discussed, but I just repeated it.

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

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

P: Well, first of all, out-of-range values. So when fabricating the data and (?) still using the outlier value as the basis for my fabricating data I had a negative standard deviation, for example. So, out-of-range values. I would look at distribution characteristics. So, you would expect a normal or maybe a skewed distribution. So, you could look at distribution characteristics and the size of the standard deviation, the size of the mean. But also higher order moments of the distribution such as the skewness or kurtosis. The correlation between certain variables, yeah.

I: And when you say that you could take a look at them, what do you mean like what would you look for?

P: So if I would have a data set and if I would have to test whether it is fabricated or not, I would look whether there will be out-of-range values, then it is clear that it cannot be real data. Then, I would compare the distribution with a distribution that I would expect. So if I would expect a normal distribution, I would test it against a normal distribution. And I would compare the means with the means that I would expect. I would compare the variances and covariances between the variables with those that I would expect based on the literature about these types of data.

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

P: Well, I think the answer would be similar to my previous answer. So, it would be genuine if the characteristics are similar to what other researchers found when performing this task. So, for the Stroop task, I would download several Stroop data sets, look at the statistical characteristics, and a specific data set would look genuine to me if the statistical characteristics would be similar to the ones that I have downloaded from the internet.

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

P: Yes, definitely.

I: And how did you take these into account?

P: I used a random number generator in R called `rnorm` - from the `mvtnorm` package and the function is called `rmvnorm` and you can simulate - randomly simulate data with a certain mean structure and covariance structure. So, I simulated four variables with the means that were similar to the means that I observed in the empirical data set that I download from the internet and the same applies to the covariance structure.

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

P: Yes, the covariances between the four variables that I had to fabricate.

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

P: So why I would think that my data would go undetected? Well, the fact that I took into account these statistics - so, for example, the correlation between those variables - it should be very high because they are from the same participants - so over participants it should be correlated highly. And I took that into account so not everyone would think about that. So, relative to others, my data set would then be difficult to detect with regard to, for example, correlations between the different variables.

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

P: In the end, not really. So, the means are of course different. The means of the mean value are not the same as the means of the standard deviation values, but I would say that the distribution should be different as well. The mean values can be distributed normally, that is not really strange, but the standard deviations are, I think, not distributed normally. So, I wanted to change that to a different kind of distribution, but I didn't have time for that.

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

P: Yes. I think that I should have looked more elaborately at the form of the distribution of both the mean values of the standard deviations - I only downloaded a small data set with 28 participants and I think it is not reliable enough to really infer the type of distribution that you would normally see. So, I think that the standard deviations are not normally distributed but there is a floor effect, but first of all I was not entirely sure of that and second of all I didn't have time to put that in my simulation.

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.

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: So, I looked at an empirical dataset and I looked at the means of the congruent and incongruent conditions and I used that as the bases of my simulation. So, I used the [?] means for my random generator and I looked at the standard deviation in the empirical dataset and used that as my standard deviation for

my simulation. And also the means of the congruent and incongruent conditions are correlated over participants and this correlation was also taken into account.

I: Ok. And how did you simulate the data?

P: How did I simulate the data? With using the `rmvnorm` function from the `mvtnorm` package in R.

I: And then you just took the values that this simulation gave you?

P: Yes, yes. I used that function with the input that is required. So, the mean values of all the variables and the covariances between the variables and the variances of the variables and the output of that function was my fabricated data.

I: Ok, so the next question is: Could you indicate what steps you took to fabricate the standard deviations for the participants?

P: I took the exact same procedure. So, I looked at the mean value of the standard deviation of the congruent and incongruent condition within my empirical data set - and also the relation between the standard deviation of one data set and the standard deviation of the other data set. And I used that to - as the bases of my data fabrication.

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

P: Yes. I did it two times. The first time I was not satisfied. The second time I was still not satisfied but I didn't have any time left to further improve it.

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

P: I inspected the fabricated data similar to how I inspected the empirical data set that I used to base my fabrication on. And I looked at whether they looked similar in terms of means, standard deviations, correlations, ranges, yeah, that's it.

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

P: Yes.

I: And how did you do that?

P: For example looking at out-of-range values.

I: Ok. And how did you determine what out-of-range values would be?

P: Out-of-range values are negative values. Yeah, that's the only ...

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

P: Yes.

I: And did this differ from like the criteria that you had for looking weird or was it the ...?

P: No, the same criteria.

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

P: 2, because I based the means and standard deviations first on the whole empirical data set and later on the whole data set with one participant removed.

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

P: R.

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

P: Yes, in R the random number generator.

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

P: Yes, I based it on an empirical data set.

I: Ok but you did not like copy real data into your ...?

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.

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

I: Ok. 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: Whether I found it difficult?

I: Yeah.

P: Well, no, I didn't think it was difficult. But I think I should have done it better and to do it better would be a difficult job.

I: Ok. And what do you imagine to be difficult about it?

P: To get the distributions of your fabricated data the way you wanted it to be.

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

P: No, I think that it is easily detected that my data is fabricated.

I: And how you think could it be detected?

P: Because I have very small standard deviations which, I think, are unusual.

I: Ok. Are there any other criteria you can think of like how your data set could be detected as fabricated?

P: Related to this, the distribution of my standard deviations is simulated as a normal distribution but, I think, in reality it is not normal. But I am not sure.

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

P: I think it is a really nice idea. So, I like the topic of your study and I also like the fact that I can contribute to that. And because I get money.

I: Ok. And did you discuss this study ...?

P: [REDACTED]

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

P: Yeah, I spoke about it with one of the other participants of your study. [REDACTED]. We know from each other that we both participate. And we previously - a while ago - we talked about what could be important for fabricating the data, but in the end we did it independently.

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

P: Well, we spoke about it and while speaking about it, you get ideas. But he doesn't really help me other from that.

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 process that you think is worth mentioning?

P: No.