Evaluation of statistical procedures used to evaluate the scientific veracity of the PhD thesis of Marleen Gillebaart

Herbert Hoijtink
Section Methodology and Statistics
Faculty of Social and Behavioural Sciences
Utrecht University
h.hoijtink@uu.nl

June 20, 2016

1 Summary

This report is based on the information provided in

- Koopman,L, Oort, F.J., and KLaassen, C.A.J. (2016). Evaluating the scientific veracity of PhD theses written under supervision of Prof. Dr. Jens Forster. Section 4. PhD thesis of Marleen Gillebaart.
- Peeters, C.F.W., Klaassen, C.A.J., and van de Wiel M.A. (2015). Evaluating the scientific veracity of publications by dr. Jens Forster. Introduction and Appendix.

subsequently referred to as KOK and PKW.

This report evaluates the appropriateness and quality of the statistical approaches used to evaluate the veracity of the PhD thesis of Marleen Gillebaart (subsequently abbreviated to MG). This report focusses on the evaluation of Chapters 2 and 3 of this dissertation by KOK published as Gillebaart, M., Forster, J., and Rotteveel, M. (2012). Mere exposure revisited: The influence of growth versus security cues on evaluations of novel and familiar stimuli. Journal of Experimental Psychology: General, 141 699-714.

- The dF approach assumes that the data are fake unless proven otherwise. Referring to "guilty unless proven otherwise" that is never used in the Dutch judicial system, it is clear that this is an invalid point of departure for evaluation of MG. Furthermore, Fisher's methods is used to combine the p-values resulting from the dF approach. This can only be done if the p-values are independent. Because sets of p-values are computed using data from *the same* respondents, they are not independent and Fisher's

method should not have been applied. Finally, the probability of a mean structure as linear as or more linear than in MG (if the data come from a population in which the mean structure is linear) is 1 in 1/.071, that is, 1 in 14.08. Lucia de Berk was convicted because of a probability of 1 in 342 million. It turned out that she was innocent. KOK should have firmly concluded that 1 in 14.08 does not even cast the shadow of a doubt on the data analyzed in MG.

- The EV approach is based on a quantity that has a lower bound of 1. The value 1 implies that it is unclear if the data are fake or real. The larger the value, the larger the evidence for faked data. However, EV cannot provide evidence that the data are real! This is an unacceptable bias. Nine of the twenty EV values reported for Chapters 2 and 3 in Table 6 in KOK are 1 or close to 1. Using a fair quantity it is very likely that these nine values provide evidence in favor of the data being real. Furthermore, the EV values are multiplied into an overall result. This can only be done if they are mutually independent, which they are not because sets of EV values are computed using data from the *same* respondents.

- Underlying dF and EV is a clear idea about data characteristics that could indicate that the data are fake. This is not the case for the final analyses reported by KOK. They observe high effect sizes, do a variance component analysis, and test equality of item means and compound symmetry. However, why these analyses might shed light on whether or not the data are faked, is not elaborated and also not obvious. Therefore these analyses do not provide information with respect to whether or not the data are fabricated.

2 A discussion of dF

Issue 1. For each of 20 plots related to Chapters 2 and 3 from MG, KOK evaluate the hypotheses:

$$H0 : \text{linearity of the three means} \tag{1}$$

and

$$Ha : \text{nonlinearity of the three means.} \tag{2}$$

Using dF a p-value is computed. If the p-value is small, e.g. smaller than .05, this is evidence against $H0$. However, as is clear from the statistical literature, p-values cannot be used to provide evidence in favor of $H0$. Nevertheless KOK use the p-value in the latter manner. It should also be noted that $H0$ implies that the data are faked and that $H1$ implies that the data are real. Stated otherwise, the point of departure is that the data are faked unless proven otherwise (guilty unless proven otherwise). This approach is never used in the Dutch judicial system and any approach investigating the veracity of MG should account for that.

Consider, for example, a two independent groups randomized design. In group a participants receive medication a and in group b participants receive medication b. If medication b is the innovative result of investments in terms of time and money, the researchers hope that it will outperform medication a. Therefore their point of departure is the null-hypothesis "the performance of a and b is equal" and only if they can reject this hypothesis by means of a p-value there is evidence that the performance is not the same. Given the research question this is a valid approach. However, it may also be that medication b is more expensive that medication a. Therefore insurance companies would like to show that the performance of a and b is equal. Their point of departure is the null-hypotheses "the performance of a and b is not equal" and only if they can reject this hypothesis there is evidence that the performance of a and b is the same. The latter can be done by means of equivalence testing. Only if the resulting p-value disqualifies the null hypothesis (which is the only thing a p-value can do) there is proof for equal performance.

If KOK evaluate hypotheses by means of a p-value, the hypotheses should have been:

$$H0 : \text{nonlinearity of the three means} \tag{3}$$

and

$$Ha : \text{linearity of the three means.} \tag{4}$$

The procedure to compute the p-value can be developed based on the principles of equivalence testing.

Issue 2. Using Fisher's approach KOK combine the 20 p-values resulting from the application of dF into one summarizing p-value. This can be done if the p-values are mutually independent. However, there are sets of p-values that are based on data from the same participants. For example, MG2.1a, MG2.1b, MG2.1c, and MG2.1d concern data from the same 66 participants. This implies that the four dependent variables are correlated and therefore that the four p-values are not independent. The summary of p-values using Fisher's approach is invalid and can not be trusted.

Issue 3. The combined p-value for Chapters 2 and 3 from MG reported by KOK is .071. Using the approach presented in PKW in Section A.1.3, this implies that "roughly speaking, under the assumption of perfect linearity in the population, the probability of finding results at least as consistent w.r.t. linearity amounts to 1 in 14.08 (1/.071)". According to KOK with 1-.071=.929 the evidence in favor of the null-hypothesis (for which the p-value can not be used!) "is high, but it does not meet the 0.999 criterion of strong evidence of low veracity". Odds of 1:14.08 do not justify the conclusion "This is high". In statistics smaller odds of 1:20 are deemed acceptable when an alpha level of .05 is used in null-hypothesis significance testing. Furthermore, what do these odds mean and imply? Sally Clark was convicted for double murder because the odds of having two children dy of cot death were 1:73miljon (https://en.wikipedia.org/wiki/Sally_Clark). Lucia de Berk was convicted for multiple murders because the odds of accidentally encountering so many dead patients were 1:342miljon (https://en.wikipedia.org/wiki/Lucia_de_Berk). As

has become clear, both ladies were innocent. See, for example, https://en.wikipedia.org/wiki/Prosecutor%27s_fallacy where the prosecutors fallacy is elaborated for the Sally Clark case: the probability of having two children dy of cot death is very small, but the (not considered in her court case) probability of sequentially murdering your two children is even smaller! The only conclusion that is justified based on odds of 1:14.08 is that there is not even the shadow of a doubt with respect to the veracity of the data analyzed in MG.

3 A discussion of EV

The EV approach as applied in KOK targets the hypotheses

$$H_{nonzero} : \text{one or more residual correlations are nonzero} \tag{5}$$

and

$$H_{zero} : \text{residual correlations are zero.} \tag{6}$$

As the names given to both hypotheses indicate, they are on equal footing, that is, neither is designated to be the null hypothesis. Note that, $H_{nonzero}$ implies that the data are faked and H_{zero} implies that the data are real. Using a likelihood ratio test (EV), KOK summarize the evidence in the data for both hypotheses for each of the 20 plots presented. EV has a lower bound of 1. The value 1 means that the data do not prefer $H_{nonzero}$ or H_{zero}, there is no reason to prefer one over the other. The larger the value of EV the larger the support for $H_{nonzero}$. If, for example, EV is 6, then the support in the data is 6 times higher for $H_{nonzero}$ than for H_{zero}.

Issue 1. Values of EV smaller than 1 would provide evidence in favor of H_{zero}. However, EV is designed such that it can never provide evidence in favor of H_{zero}. This is unacceptable, EV can be undecided, it can support $H_{nonzero}$ (faked data), but it cannot provide support for H_{zero} (real data). The reason for this is that EV only accounts for the fit of both hypotheses (how well are the data represented by each hypothesis) but not the complexity of both hypothesis (how parsimonious is each hypothesis). In all model selection criteria that exist (note that KOK use EV as a model selection criterion) both fit and complexity of hypotheses are accounted for, but not in EV. Note that, H_{zero} is a point hypothesis (each of the three correlation involved is zero) which makes it very parsimonious. Note also, that $H_{nonzero}$ implies a range of values the three correlations can attain. This hypothesis is therefore less parsimonious than H_{zero}.

What is presented in Table 6 of KOK are 20 EV values based on an evaluation of *only* the fit of both hypotheses. Of these 20 values 9 are (close to) 1 which implies that the fit of both hypotheses is the same. If the hypotheses would also be penalized for their complexity, then $H_{nonzero}$ would be penalized much more than H_{zero} because it is less parsimonious. This implies that an EV-measure that would account for complexity (a Bayes factor would automatically do this) would probably render 9 EV-values that are (substantially?) smaller than 1,

that is, there would be 9 EV-values that would render evidence in favor of H_{zero}, that is, in favor of the data being real!

The bottom line is that EV is biased in favor of $H_{nonzero}$. Therefore it can not be used for a fair evaluation of MG. Furthermore, imagining a correction of EV that is fair (that not only accounts for the fit of an hypothesis, but also for its complexity), it is likely that the 9 EV values that are currently close to one, would become (substantially?) smaller than 1, thereby providing (substantial) support for H_{zero}. Multiplying 20 EV values corrected for complexity would therefore not render an overall value of 60075, but a smaller number, possibly even a number smaller than 1.

The overall conclusion of KOK based on dF and EV was "inconclusive evidence for low veracity". This is, as argued in this and the previous section, not supported by the analyses executed. The only conclusion that can at this point be obtained is that there is not the shadow of a doubt with respect to the veracity of MG.

Issue 2. Analogous to Issue 2 from the previous section, the EV-values are not mutually independent and can therefore not be multiplied to obtain an overall value.

4 A discussion of effect sizes, variance components, and equality of means/compound symmetry structure

After computing and evaluating dF and EV, KOK execute additional analyses. They compute effect sizes that turn out the be larger than usually observed. Furthermore, they execute a variance component analysis. Finally, they evaluate equality of three items means and compound symmetry of the covariance structure. The results presented by KOK show that equality and compound symmetry do not hold.

Issue 1. Based on dF and EV, KOK conclude "inconclusive evidence for low veracity" (note that the previous sections strongly suggest "no evidence for low veracity"). Why then continue with further analyses? It is clear that prior to looking at the data and doing analyses it was decided to use dF and EV. However, the final set of analyses seem to be post hoc. How and why was it decided to do precisely these analyses? Were other analyses considered and not executed?

Issue 2. KOK observe large effect sizes. However, they do not elaborate why and whether or not these are indications of faked data. One might argue that large effect sizes are rare and therefore "suspicious". One might also argue that "fakers" would therefore create their data such that effect sizes are medium, small, and occasionally about zero. Still other lines of argument can without doubt be construed. As it is, the presence of large effect sizes do not provide any information with respect to whether or not the data are faked.

Issue 3. KOK execute a variance component analysis and note which com-

ponents have larger and smaller contributions. It is, however, not elaborated why and how the information resulting from the variance component analysis can be used to determine whether or not the data are faked. Therefore, this analysis does not provide any information with respect to whether or not the data are faked.

Issue 4. KOK expect approximate equality of item means and a compound symmetry covariance structure. First of all, they do not define what they mean by approximate and they do not test for approximate equality but for exact equality. Furthermore, KOK do not elaborate how the information resulting from their tests can be used to determine whether or not the data are faked. Therefore, this analysis does not provide any information with respect to whether or not the data are faked.

5 Conclusion

The methods used to evaluate the veracity of MG are inadequate. In summary and as elaborated in the previous sections, both dF and EV are inadequate and can not and should not be used for the evaluation of the veracity of the PhD thesis of Marleen Gillebaart. Furthermore, it is not clear why the additional analyses were executed, nor how these analyses can be used to determine whether the data are faked or real.

Finally, the readers of this report should take notice of Buchanan, M. (2007). Statistics: conviction by numbers. Nature, 445 (7125), 254-255. The main message is that "numbers" are not enough. Without a "confession" it is irresponsible to base a "conviction" only on "numbers". In the faked data context there is a way out. It is not possible to *proof* by means of statistical analyses that data are faked and thereby "convict" the faker. However, it is possible to replicate all the suspicious studies. If none of the replications corroborate the suspicious studies, it is still not possible to "convict" the researcher in question, however it does allow claims like "the research projects of this researcher can not be replicated". Note that this paragraph *does not* express an opinion with respect to the (non)replicability of the results of MG. However, it *does* express an opinion with respect to the approach that should be used to evaluate suspicious studies: replication research.