Reply to the Comments by Hoijtink on the First Version of the Report by Koopman, Oort, and Klaassen

Chris A.J. Klaassen and Frans J. Oort
University of Amsterdam

September 11, 2017

We would like to thank Prof. Hoijtink for his contribution to the discussion about the scientific value of Gillebaart, Förster and Rotteveel (2012) by his manuscript Hoijtink (2016).

Dr Hoijtink has put forward several objections to the first version of our report Koopman, Oort and Klaassen (2016), in particular to the applied statistical methods, namely the **V**- and **C**-method, as Dr Förster calls them. Many of these objections have been formulated before in discussions about Peeters, Klaassen and Van de Wiel (2015) and have been completely and convincingly refuted in our replies to be found at
http://www.uva.nl/en/content/news/news/2015/07/update-articles-jens-forster-investigated.html

Therefore, in the revision of Koopman, Oort and Klaassen (2016) we have maintained the application of these methods, but we have also applied an additional straightforward method based on the statistic $Z_{\mathbf{V}}$, which underlies the **V**-method. Let us call this straightforward method the $Z_{\mathbf{V}}$-method. In Appendix A of our revised report this statistic $Z_{\mathbf{V}}$ is defined and it is shown that the **V**-method can and may be viewed as a way to interpret $Z_{\mathbf{V}}$ within a Bayesian framework. Moreover, it is shown that the evidential value **V** in this Bayesian approach is a coarsening function of $Z_{\mathbf{V}}$ and that the threshold of 6 as used in the **V**-method introduces a very rough further coarsening to an indicator variable. This unnecessary coarsening explains why conclusions based on the $Z_{\mathbf{V}}$-method are much stronger than conclusions obtained via the **V**- and **C**-method. These conclusions are obtained by studying the behavior of $Z_{\mathbf{V}}$ under ANOVA model assumptions, which have also been used in the statistical analyses in all considered PhD-theses themselves. However, we avoid the assumption of constant variances, an assumption that does not always seem to be fully justified.

Let us return to the letter of Dr Hoijtink. In his **Summary** he mentions three main objections (at bullet points, which we enumerate for ease of reference):

1. *The ΔF approach assumes that the data are fake unless proven otherwise. ... Fisher's method is used to combine the p-values resulting from the ΔF 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 ...Lucia de Berk was convicted because of a probability of 1 in 342 million. It turned out that she was innocent. Koopman, Oort and Klaassen (2016) should have firmly concluded that 1 in 14.08 does not even cast the shadow of a doubt on the data analyzed in Gillebaart, Förster and Rotteveel (2012).*
   There seems to be a misunderstanding as to the hypotheses we test in our report, as we will explain below, where **Issue 1** from Section **2 A discussion of ΔF** is discussed. The independence issue is handled correctly in the revision of our report by grouping the subexperiments into sets of independent subexperiments. The data in the Lucia de Berk criminal court case turned out to be erroneous, whereas the results in Gillebaart, Förster and Rotteveel (2012) are published in a scientific journal. Finally, application of the $Z_{\mathbf{V}}$-method yields much stronger results.

2. *The $\mathbf{V}$ 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, $\mathbf{V}$ cannot provide evidence that the data are real! This is an unacceptable bias. Nine of the twenty $\mathbf{V}$ values reported for Chapters 2 and 3 in Table 6 in Koopman, Oort and Klaassen (2016) 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. ...*
   In a scientific publication every single claim should be scientifically reliable. So, a dubious claim cannot be made valid by adding some reliable, exonerating claims. In other words, there should not be exonerating $\mathbf{V}$s and we should always have $\mathbf{V} \geq 1$.

3. *Underlying ΔF and $\mathbf{V}$ 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 Koopman, Oort and Klaassen (2016). 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.*
   In our report it is nowhere claimed that data were fabricated, but it

is claimed that the veracity of the presented results is in doubt. The cause of this cannot be determined from just the publication itself.

In our revision we report more extensively on the exploratory analyses that we conducted when we found various peculiarities in the pre-processed data that Förster made available; see Appendix B. As explained in the appendix, errors in the tables and in the reported effect sizes caused us to repeat the analyses of Gillebaart et al. We then noticed the extreme effect sizes, and also calculated effects sizes d and the associated standard errors. That is when we found the negative correlations between repeated measures, which caused us to calculate all correlations. The highly unusual patterns of correlations, in turn, caused us to estimate the reliability of the outcome measure, by calculating generalizability coefficients. To our surprise, a large number of these estimates were zero, as if all variance is random error. When we conducted further analyses on the item responses, we found that the items were not interchangeable at all, which convinced us that the pre-processed data do not match with the design of the experiments, which might be due to honest mistakes in the pre-processing of the data. Still, we have to conclude that either the experiments have not been conducted in the way as described by Gillebaart, Frster & Rotteveel (2012), or that serious errors have been made in the pre-processing of the data, both of which invalidate the conclusions as reported by Gillebaart, Förster & Rotteveel (2012).

In his Section **2 A discussion of ΔF** Dr Hoijtink discusses three issues.

Issue 1 Under this heading it is claimed that our methods test the null hypothesis

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

against the alternative hypothesis

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

Indeed, this would be wrong. However, in our report the **C**- and **V**-method test the null hypothesis

$$\begin{aligned} H_0 : &\text{ the data underlying the results as presented in the paper,} \\ &\text{ have been generated according to the ANOVA model assumptions} \end{aligned} \tag{1}$$

against the alternative hypothesis

$$\begin{aligned} H_1 : &\text{ the data underlying the results as presented in the paper,} \\ &\text{ have been generated in some other way.} \end{aligned} \tag{2}$$

In the revision of our report these hypotheses are evaluated via a study of the behavior of the $Z_{\mathbf{V}}$-statistic as well. This evaluation is done under the assumption of perfect linearity of the *population* means, which is the most favorable assumption for the authors of Gillebaart, Förster and Rotteveel (2012).

Issue 2 The independence issue is taken care of in the revision of our report.

Issue 3 Under this heading the independence issue is discussed again. We have taken good care of this in the revision. By the way, the Sally Clark criminal court case rested on the false assumption that cot deaths are independent among siblings.

In his Section **3 A discussion of** *EV* Dr Hoijtink mentions two issues, of which the second one is the preceding **Issue 2**. His first issue here is the fact that the evidential value EV or $\mathbf{V}$ satisfies $\mathbf{V} \geq 1$. As stated before, this is completely acceptable and even necessary. It cannot be the case that unreliable experiments in a scientific paper are compensated by reliable ones. This is in contrast to criminal court cases, where the adage is *In dubio pro reo.* However, when judging the scientific value of a publication it should be *In dubio pro scientia.*

In his Section **4 A discussion of effect sizes, variance components, and equality of means/compound symmetry structure** four issues are presented.

Issue 1 *Based on ΔF and $\mathbf{V}$, Koopman, Oort and Klaassen (2016) conclude "inconclusive evidence for low veracity" (note that the previous sections strongly suggest "no evidence for low veracity"). Why then continue with further analyses?*
When calculating ΔF and V we found errors in one of the tables with descriptive statistics. As Förster had made the (pre-processed) data that Gillebaart used available, we were able to correct the numbers and repeat the analyses. We then found errors in the reporting of the effect sizes (partial eta-squared). When we reported the correct partial eta-squared they seemed very large to us. To aid interpretation we also calculated effect sizes d and the associated standard errors. When calculating standard errors we found that many of the repeated measures were negatively correlated, etc. (see above).

Issue 2 *Koopman, Oort and Klaassen (2016) observe large effect sizes. However, they do not elaborate why and whether or not these are indications of faked data.*
We nowhere stated that the data are fake, we only stated that the pre-processed data do not show characteristics that fit with the description of the experiments. However, we do think that it is highly unlikely to find effects when the experimental manipulation is subtle and the outcome measure is unreliable.

Issue 3 *KOK execute a variance component analysis and note which components 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.*

The previous version of our report had a large number of separate appendices. By way of summary, we included a variance component analysis in the report. We agree that the variance components analysis is difficult to interpret. Instead, we have now reported the additional exploratory analyses more extensively in Appendix B of the new version of our report. The results of these analyses are not evidence of faked data, but rather show that the pre-processed data do not fit with the description of the experiments by Gillebaart et al. (2012).

Issue 4 *KOK expect approximate equality of item means and a com- pound 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.*

We expected equality except for sampling error. As hypotheses of invariant means, variances, and covariances must be rejected, we conclude that the items are not valence neutral, nor counterbalanced, nor randomized. This is contrary to the description of Gillebaart et al. (2012), and this is important as the conclusions of Gillebaart et al. (2012) require interchangeable items. Nowhere do we say that the data are faked, we only say that the data do not fit with the description. This may be caused by honest mistakes, but that would still render the conclusions of Gillebaart et al. (2012) invalid.

In his Section **5 Conclusion** Dr Hoijtink notes that *Without a "confession" it is irresponsible to base a "conviction" only on "numbers"* and he expresses *an opinion with respect to the approach that should be used to evaluate suspicious studies: replication research.*

Recently, some studies in the field of social psychology have been replicated and in quite a few instances the original findings were not confirmed. Finally, our point of departure has been and still is, that any scientific publication should be scientifically reliable. Otherwise its contents have no scientific value. The $Z_{\mathbf{V}}$-analyses show strong evidence of low veracity and the additional, exploratory analyses show that either the experiments have not been conducted in the way as described by Gillebaart, Förster & Rotteveel (2012), or that serious errors have been made in the pre-processing of the data, both of which invalidate the conclusions as reported by Gillebaart, Förster & Rotteveel (2012).

References

Gillebaart, M., Förster, 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.

Hoijtink, H. (2016). Evaluation of statistical procedures used to evaluate the scientific veracity of the PhD thesis of Marleen Gillebaart. *Report.*

Klaassen, C.A.J. (2015). Evidential Value in ANOVA-Regression Results in Scientific Integrity Studies. *arXiv:1405.4540v2.*

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 Förster. *Report.*

Peeters, C.F.W., Klaassen, C.A.J. and Van de Wiel, M.A. (2015). Evaluating the Scientific Veracity of Publications by dr. Jens Förster. *Report.*