# Reply to the Comments by Förster 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. Förster for his contribution to the discussion about the scientific value of Gillebaart, Förster and Rotteveel (2012) by his letter to Prof. Kamphuis Chair of the Department of Psychology, UvA.

Dr Förster 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 Förster. He claims: *Specifically, I think that the estimates of low veracity are inflated because*

*(1) Some of the* **V** *and* **C** *values in that paper are dependent on each other, as they come from the same experiment, but the authors treat them as if they are independent.*

*(2) There is only one out of 23 computed* **V** *values that exceeds the critical level of 6. The left-tail probability of that is 86%. The left tail probability of getting one such result in 13 papers is close to unity.*

*(3) The authors count-in upper-bounds of* **V** *values, although the PKW-report clearly says that only lower bound values should be counted.*

*(4) The authors of the KOK-report say that the effects are too large, but do not provide an assessment of how distinctly large they are, nor do they take into account the possibility of benign file drawering of weaker results (which I admitted doing many times and which was common practice when I did the studies).*

*(5) The authors of the KOK-report say that the existence of stimulus-related variance is evidence of irregularity, whereas in fact it is a normal finding that might occur in studies on evaluation.*

The criticism as formulated in issue (1) is justified. Therefore, we have grouped all subexperiments into sets of independent subexperiments in the revised report. The computation in issue (2) is valid only under independence, but it illustrates that the extreme coarsening of $Z_{\mathbf{V}}$ in the **V**-method does not yield a realistic picture of the situation. This is the reason why we have applied the $Z_{\mathbf{V}}$-method. Issue (3) also illustrates that the coarsening in the **V**-method should be avoided, as is done in the $Z_{\mathbf{V}}$-method.

So, the objections raised in issues (2) and (3) become irrelevant given the application of the $Z_{\mathbf{V}}$-method as described in Appendix A and applied in the new Section 4.3 of our revised report. The conclusion of this Section 4.3 clearly necessitates retraction of the paper Gillebaart, Förster and Rotteveel (2012).

In addition, we now 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. We maintain that the effect sizes are extremely large, especially considering the subtle experimental differences between conditions, but we do not have a criterion for judging an effect size as too large (issue 4). Still, our follow-up analyses do show that the reliability of the outcome measure is not sufficient to find any effects of the independent variables, and that the items are not interchangeable (issue 5). We therefore conclude 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, as the validity of the experiments hinges on the interchangeability of the items, we cannot attribute the effect sizes to the experimental manipulations.

The objections raised by Dr Förster in his letter in section I. Specific shortcomings in the analyses used for Gillebaart, Förster & Rotteveel (2012) elaborate on issues (1) through (5) and on points raised by Dr Hoijtink. As mentioned before, in the revision of our report we have taken care of the justified independence issue (1) and we have addressed issues (4) and (5) by more extensively reporting about the additional exploratory analyses. We have also countered objections (2) and (3) by applying the fundamental, straightforward $Z_{\mathbf{V}}$-method. The points raised by Dr Hoijtink will be discussed in a separate letter.

The points elaborated on by Dr Förster in his letter in section II. The statistical analyses used in both reports are invalid have been addressed in earlier discussions to be found at
http://www.uva.nl/en/content/news/news/2015/07/update-articles-jens-forster-investigated.html

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.

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*.