Carel F.W. Peeters, Chris A.J. Klaassen, Mark A. van de Wiel 02/07/2015

The report 'Evaluating the Scientific Veracity of Publications by dr. Förster' has been the focus of some public discussion on platforms such as Twitter, PubPeer[1], and certain blogs[2]. We welcome and appreciate the considerations and deliberations contained therein. We noticed, however, that at times the discussion seems to lack some nuance and that certain points and misunderstandings are reiterated. We would like to respond to these. To avoid the response being scattered we will give a 'meta-response' here.

On the Type I error of the evidential value

We understand that an evaluation of the Type I error when the assumptions underlying the evidential value are violated, is of interest. The Type I error will depend strongly on the extent of the violations. However, the data are available to us only in the form of summary statistics obtained from the publications under investigation. This does not allow for a verification of (abidance of) the assumptions.

While the type I error of a single experiment/study is of interest, all of the discussed publications are composite studies, i.e., they are composed of multiple (sets of) independent (sub-)experiments. The difficulty, w.r.t. the Type I error, lies with dependencies among the (sub-)experiments. Explicit modelling of these dependencies (if they can be known at all), without having access to the raw data, is highly problematic. Therefore, we resorted to the use of approximate bounds as mentioned in the Tables on Pages 2 and 3 of our report. The approximate bounds are obtained, also from the perspective of the central limit theorem, on the basis of reasonable assumptions.

Please also note that the probability that the evidential value exceeds the threshold, has been computed under the assumption of true super-linearity (for lack of a better word) in the population. If the truth would not be super-linear the probability of observing large values of V would decrease. Therefore, the probabilities reported in the table on Page 3 of the report are likely conservative upper bounds. Moreover, the threshold V > 6 is reasonable when considering the control/reference studies (see Table 1.2 of the report).

On the purported leniency in concluding 'low veracity of the data'

Online discussions often focus on the approximate Type I error (for a single study). In these discussions a Type I error of about 8% is not deemed stringent enough to conclude 'strong evidence for low veracity of the data' (more on our classification below). However, as was explained above, this probability is a conservative upper bound. Also note that we often observed V's that extend far beyond '6', which further diminishes the False Positive probability. Finally, other evidence, including the delta F-test and visualizations of the data were used to support our classifications. Once more, we regret that an exact computation or approximation of the Type I error is impossible due to absence of the raw data. Likewise, a more formal Bayesian probabilistic statement would require a good prior for each study. What should this prior account for? The fact that the raw data is absent? The empirical evidence that super-linear patterns were observed for many of the studies by Dr.

[1] https://pubpeer.com/publications/5439C6BFF5744F6F47A2E0E9456703

[2] http://rejectedpostsofdmayo.com/2015/06/09/fraudulent-until-proved-innocent-is-this-really-the-new-bayesian-forensics-rejected-post/

Förster? The current scientific reputation of the field of social psychology? Inclusion of such prior information would have led to a subjective analysis.

On the absence of exculpatory evidence and the cumulative evidential value

The evidential values of independent studies/samples can be combined (by multiplication) into an overall evidential value. Some commenters object to this, because of the absence of exculpatory evidence in V, arguing that individual studies with low evidential values may combine into a high overall V. This argument convolutes a statement in the paper that develops V, and our use of V in the report regarding the publications of dr. Förster. Let us explain.

First, we believe the absence of exculpatory evidence to be quite natural, as it implies that (sub-) experiments for which the data have high veracity are not able to exonerate (sub-)experiments for which data veracity is low.

Second, the conclusions reached in the report are never based on overall evidential values. They are, among other things (see also *On cumulative evidence* below), based on the number of evidential values of individual samples/sub-experiments that are considered substantial, in relation to the total number of sub-experiments that constitute an individual publication. That is, we approximately control the Type I error for a publication by considering the number of substantial evidential values in a total batch of experiments; see also the table from page 3 of our report. As explained above, the probabilities reported in the table on Page 3 are conservative upper bounds.

On the relation between the delta F-test paired with Fisher's method and V

There have been some questions on why the evidential value V was used in the report and if the delta F-test paired with Fisher's method used in the original whistleblower report[3] would not have been sufficient to analyze the additional publications.

First, note that the Fisher method needs many independent samples to be reliably employed. The evidential value V enables assessment of data veracity for individual (sub-)experiments. In this respect, the V could be termed indispensable to evaluate those publications who have only few constituent (sub-)experiments.

Second, we belief that V has added value in that it quantifies anomalous linearity in a different way than the delta F-test paired with Fisher's method. To quote from Section B.ii.3 of Response_to_LDletter_JFcomments.pdf:

> "The *p*-values for the delta-F test must be uniformly distributed when the *null hypothesis of perfect linearity in the cell means holds in the population*. Seeing consistently high *p*-values should then raise suspicion. The impact of these *p*-values is formalized with Fisher's method, which allows probabilistic statements regarding the extremity of the observed *p*-values *under the assumption that the null hypothesis is true*. The problematic nature of the data then lies in the consistently small deviations from perfect linearity. The evidential value also assesses lack of variation but does so by allowing for dependence between the measurement errors of the respective factor-levels in the ANOVA model. It then directly assesses, for individual (sub-)experiments, the hypothesis of a dependence structure in the underlying data (which is indicative of low data veracity) against the standard ANOVA model assumption of independence (which is indicative of high data veracity). The evidential value in a sense quantifies excessive closeness of the cell averages to perfect linearity given the reported cell

[3] https://retractionwatch.files.wordpress.com/2014/04/report_foerster.pdf

> variances, under the assumption that *perfect linearity in the cell means holds in the population*."

While related, the methods employ different angles. Note that the results between V and the delta F-test are highly concordant: In all cases a large value of V coincides with a large p-value for delta F. This brings us to the next point.

On cumulative evidence

Some suggest that we base our classification of investigated publications solely on the evidential value V. Our classifications (per investigated publication) are, however, based on an accumulation of evidence resulting from careful analysis. This accumulation consists of the evidential values, the p-values for the delta F-test and (when the number of independent experiments allow) the employment of Fisher's method, the visualizations of the (sub-)experiments, and additional data-peculiarities (such as non-linear between-effects canceling out into a very linear effect for pooled data). The criteria stated on page 2 of our report then serve as a guideline, but they are not applied blindly. This also means that the weight of evidence depends on the publication under scrutiny. For example, as stated in the report, when a publication has few independent samples/experiments, the weight of evidence necessarily shifts to the evidential value.

On the classification of investigated publications

Some public claims seem to imply that we are concluding fraud or misconduct for certain publications investigated in our report. Again, these claims seem to stem from a convolution of statements in the paper that develops V and our use of V in the report regarding the publications of dr. Förster. As we state clearly in our report, we are assessing the *veracity of data* as reported in the publications under investigation.

On procedural decisions

There is some dismay with certain procedural decisions, such as the intention of the Executive Board of the UvA to ask editors to consider retraction of publications that are classified in the report as bearing 'inconclusive evidence for low data veracity'. The authors of the report, however, bear no responsibility for procedural decisions made by third parties.

On purported slander

Some ask if the public availability of our report counts as slander with respect to the person of dr. Förster. We understand that the report may be regarded as damaging. However, we view the report as a contribution to the debate regarding the scientific value of some publications by dr. Förster; a debate, which includes a previous retraction[4] and which has had previous public exposure.

It must also be noted that we scrutinize the empirical trustworthiness of publications by dr. Förster, not the integrity of his co-authors. As we state in Section 1.6 of the report:

> "Importantly, it must be emphasized that the empirical trustworthiness of publications by JF is under scrutiny, not the integrity of his co-authors. The report does not imply, nor does it intend to imply, that the collaborators of JF were involved in problematic or dubious practices."

[4] http://retractionwatch.com/2014/11/27/retraction-appears-for-social-psychologist-jens-forster/

On the field of social psychology

We have seen some comments online that seem to suggest that social psychology must not be taken seriously as a scientific endeavor. We state explicitly that the report in no way possible intends to condemn the field of social psychology. We distance ourselves from all such statements and regret if our report is used to support such claims.