Contents lists available at SciVerse ScienceDirect

Journal of Mathematical Psychology

journal homepage: www.elsevier.com/locate/jmp

Commentary

The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses

Joachim Vandekerckhove*, Maime Guan, Steven A. Styrcula

University of California, Irvine, United States

HIGHLIGHTS

- The consistency test is designed to be conservative and has low detection power.
- Studies with greater bias are less, not more, likely to be considered inconsistent.
- Systematic application of the test would not improve the quality of the literature.

ARTICLE INFO

Article history:
Available online xxxx

Keywords:
Meta-analysis

ABSTRACT

If the consistency test were used to select papers for inclusion in meta-analysis, the resulting estimates of true effect sizes would be no less biased. Increasing its detection rate at the risk of a higher false alarm rate biases the pooled effect size estimates more—not less—because papers reporting large effect sizes are less likely to be judged inconsistent.

The consistency test discussed in the target article has several uses in what might be called "statistical forensics", in which data sleuths audit published work to determine whether reported data are credibly the result of an unbiased scientific process. On the one hand, it is hoped that the existence of such methods could serve to incentivize both good research practice and unbiased publishing of results. On the other hand, it has been suggested (e.g., by Francis, 2012) that methods like the consistency test could be used to assess the general credibility of a paper. Here, we explore the consequences of the application of a policy in which the consistency criterion is used as an inclusion criterion for literature reviews and meta-analyses. Like Francis (in press), we evaluate the effects of such a policy through Monte Carlo analysis.

1. A Monte Carlo experiment

To evaluate the usefulness of the consistency test, we performed a Monte Carlo experiment in which we simulated a corpus containing a large number of projects. Each project consisted of several studies on the same treatment effect that was tested with a paired-sample t-test. Then we performed a meta-analysis on the corpus to estimate the true underlying size of the effect. Throughout, we assume that a research project consists of some number of studies, and that those studies that yield a significant results are published in a single paper. We then evaluate the results of meta-analysis on the full corpus (including all unpublished studies), on a biased corpus (including only the published studies), and on a "corrected" corpus (excluding from the biased corpus all papers that report a set of studies that is judged statistically inconsistent).

1.1. Simulation constants

We simulated large fictitious corpora, with 50,000 projects in each. The number of studies involved in a project was a draw from a negative binomial distribution with parameters 0.5 and 5, so that it ranged from 2 to 24, with median 7. The sample size within each study was a draw from a negative binomial with parameters 0.05 and 1. It ranged from 10 to 160, with median 23. The conclusions of our experiment were not sensitive to reasonable changes in these settings.

1.2. Manipulations

As our main independent variable, we manipulated the selection criterion for studies to be included in a meta-analysis. In the *full access* condition, all studies were included in the meta-analysis. In the *publication bias* condition, file drawer bias was introduced by censoring those studies in which the null hypothesis was not rejected. In the *after correction* condition, we removed studies that failed to reject the null hypothesis and also removed

[*] Correspondence to: Department of Cognitive Sciences, University of California—Irvine, Irvine, United States.
E-mail address: joachim@uci.edu (J. Vandekerckhove).

0022-2496/$ – see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jmp.2013.03.007

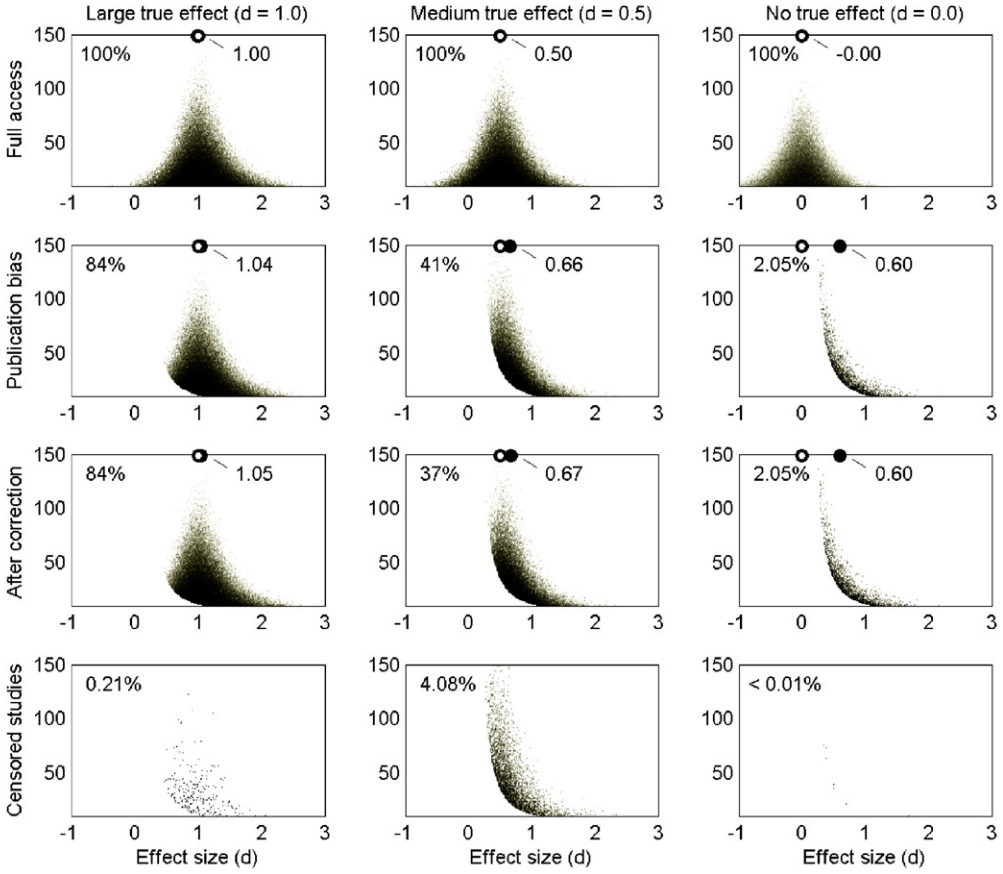

Fig. 1. Funnel plots of simulated literatures showing various types of bias. Funnel plots have sample size on the vertical axis and recovered effect size on the horizontal axis. These funnel plots are rendered as densities, with darker colors indicating a higher density of studies in that region of the plot. (Note: the different panels use the same vertical and horizontal scales, but the color scale is different for each panel.) The columns differ in the underlying (true) effect size, with values 1.0, 0.5, and 0.0, respectively. *Top row*. The *full access* condition shows symmetric funnel plots. *Second row*. *Publication bias* causes the disappearance of studies finding effect sizes close to 0. For smaller true effect sizes, the amount of censoring increases. *Third row*. Using the proposed conservative consistency test to correct the literature has little effect: *after correction*, the funnel plots look virtually identical. *Bottom row*. The distribution of studies that were censored due to inconsistency shows that more studies are censored for medium true effect sizes.

studies if they were part of a multistudy paper that did not pass the consistency test. Independently, we manipulated the true size of the effect, with Cohen's d ranging from 0 (no effect) to 1 (a large effect).

Our dependent variable was the pooled effect size estimate, which was defined as Cohen's d, pooled using its precision as weights (using the relevant formulas found in Francis, in press).

1.3. Results

1.3.1. The conservative consistency test has little effect

Fig. 1 shows funnel plots of the simulated corpus. In a funnel plot, each study's sample size is plotted against its reported effect size, yielding a characteristic funnel shape where the tip of the funnel contains large-sample studies with low estimation error. Three corpora are shown, with true Cohen d-values 1.0, 0.5, and 0.0 from left to right. The true effect size is also displayed as a circle near the top of each plot, and the estimated (pooled) effect size is written out and shown as a black dot. The percentage of the total number of studies that was used in the construction of each panel is indicated on the top right of the panel.

In the top row of the figure, three unbiased corpora are shown (*full access* condition). There, 100% of the studies are used, and the effect size is recovered well (and the dot near the top of the plots is occluded by the circle).

In the second row of the figure, three corpora with file-drawer bias are shown (*publication bias* condition). The funnel plots look distinctly different, with a smaller funnel shape missing from the plot in the region around $d^{\text{est}} = 0$, where null hypotheses were not rejected.[1] When the true effect size is large, the effect on the estimated effect size is small (and only 16% of studies are hidden), but, as the true effect size diminishes, and then vanishes, the influence of file-drawer bias on the pooled effect size becomes larger as the proportion of censored studies increases (as expected; Greenwald, 1975) to 59% and then 98%.

In the third row of the figure, the three corpora have been "corrected" by systematic application of the consistency test (*after correction* condition), censoring studies that were part of a multistudy project that did not pass the test. As noted by Francis (in press), the proportion of papers censored is small. As a result, the funnel plots are virtually identical, and the bias in the pooled effect size estimate remains.

To illustrate how few papers are censored by the consistency test, the bottom row of panels shows the distribution of studies that were visible in the literature but were then censored due to lack of consistency. An interesting effect now appears: more

[1] Analytical functions to describe the shape of the major funnel plot and the minor negative funnel plot are readily derived, and are driven by the inverse of the sample size N.

Please cite this article in press as: Vandekerckhove, J., et al. The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. Journal of Mathematical Psychology (2013), http://dx.doi.org/10.1016/j.jmp.2013.03.007

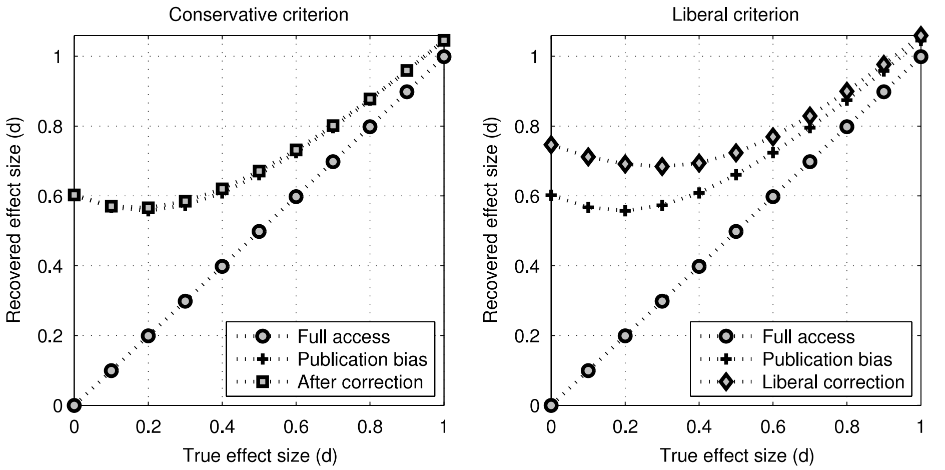

Fig. 2. Effects of censoring strategies on the estimation bias of a population effect size. Effect sizes are given in units of Cohen's *d*. Both panels show the recovery function under *full access* (circles on the diagonal) and under *publication bias* (squares). *Left panel.* Recovery function *after correction* using a conventional credibility criterion of 0.1. *Right panel.* Recovery function after *liberal correction* using a credibility criterion of 0.8.

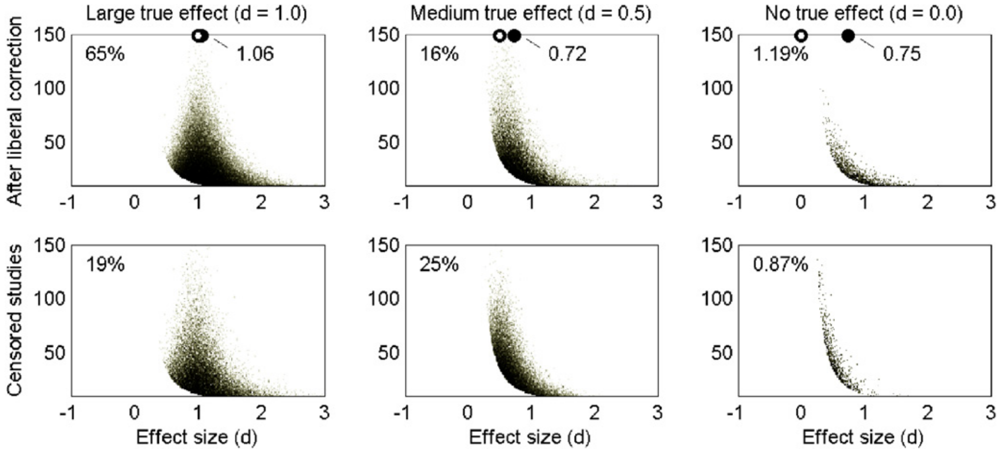

Fig. 3. *Top row.* Using a liberal consistency test (with critical value 0.8) to correct the literature has a perverse effect: *after liberal correction*, the estimated effect size is more biased than before the correction. *Bottom row.* The cause of this effect can be understood by inspecting the distribution of studies that were censored due to inconsistency. The studies that were censored tended to be those with relatively small bias, as is particularly visible in the bottom right panel.

studies are censored in the medium effect size condition (4%, compared to much less than 1%); when the true effect size is zero, almost no inconsistency is found.

1.3.2. A more liberal consistency test has undesirable effects

While the consistency test in the target article is conservative by design (one wishes to avoid unjustly accusing an author of bias or incompetence), this may not be a concern for one who wishes to do meta-analysis. Indeed, the test is conservative in rarely making false indictments, but it is liberal in that truly biased publications will frequently be left scatheless. A meta-analyst perusing a large literature may wish to excise biased papers with great certainty at the cost of incurring false alarms.

Here, we compare the behavior of a corrective policy with critical consistency value 0.1 (the current convention) to one with a liberal critical value of 0.8, hopefully eliminating all but the most consistent papers.

The unfortunate results are in Fig. 2. Both panels show the magnitude of the estimation bias over a range of true effect sizes. In both panels, the horizontal axis shows the true simulated effect size, ranging from $d = 0$ to $d = 1$. The dashed line shows the pooled effect size estimated from the entire set of studies. As expected, it follows the first bisector exactly. The full line with the downward-pointing markers indicates estimates based on the biased literature, while the dotted line with the upward-pointing markers indicates estimates based on the corrected literature. In the left panel, where a critical consistency value of 0.1 was used, the latter two essentially overlap due to the small number of censored papers. In the right panel, the consistency test was made much more liberal, now using a criterion value of 0.8. Perversely, censoring all but the most consistent-seeming papers from the literature causes *greater* bias in the effect size estimate.

The same effect is illustrated in Fig. 3, where the same three corpora as before have been treated with the more liberal consistency test. The proportion of censored papers is now considerable. The bias in the effect size estimate is now larger, and the lower right panel illustrates well why this is the case: the consistency criterion has preferentially eliminated studies with relatively small reported effect sizes. The most biased studies, with large effect sizes, are left intact, even though, in the condition where the true effect size was 0, fully 3/4 of the published studies were removed from the meta-analysis due to statistical inconsistency.

2. Discussion

The relatively greater likelihood of detecting bias in slightly biased studies is an unavoidable property of the consistency test. Highly biased studies will report large effect sizes and thus exaggerate the post hoc power. High power and large effect sizes are, in the logic of the consistency test, no cause for concern: after all, that is what unbiased well-designed studies reporting true large effects *should* look like. Without prior knowledge of the true effect size, the consistency test is unable to detect highly biased reports.

As noted in the target article, the consistency test is extremely conservative, and a ''negative side effect of such conservatism is that it also misses many situations with a strong file-drawer bias''. While useful as a test for individual audits, this lack of statistical power, combined with a propensity for flagging slightly biased rather than heavily biased reports, renders the consistency test all but useless as a tool in meta-analysis. And that is a pity.

Acknowledgments

This work was partly funded by a Summer Undergraduate Research Project grant from UC Irvine to MG and JV.

References

Francis, G. (2012). Evidence that publication bias contaminated studies relating social class and unethical behavior. *Proceedings of the National Academy of Sciences*, *109*(25), E1587.

Francis, G. (2013). Replication, statistical consistency, and publication bias. Journal of Mathematical Psychology, in press (http://dx.doi.org/10.1016/j.jmp.2013.02.003).

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*(1), 1–20.

Please cite this article in press as: Vandekerckhove, J., et al. The consistency test may be too weak to be useful: Its systematic application would not improve effect size estimation in meta-analyses. Journal of Mathematical Psychology (2013), http://dx.doi.org/10.1016/j.jmp.2013.03.007