The prevalence of statistical reporting inconsistencies in management research: A replication of Nuijten et al. (2015)

Johannes M. van Zelst[1], Peter Kruyen[2], & Chris H. J. Hartgerink[3]

[1] Department of Organization Studies, Tilburg University

[2] Department of Public Administration, Radboud University Nijmegen

[3] Department of Methodology and Statistics, Tilburg University

Author Note

Correspondence concerning this article should be addressed to Johannes M. van Zelst, Warandelaan 2, 5037 AB Tilburg, The Netherlands. E-mail: j.m.vanzelst@uvt.nl

Abstract

This study documents reporting inconsistencies in a sample of over X p-values reported in forty management journals from 1995 until 2015, using the new R package statcheck. We find that X% of the articles in management research contains at least one reporting inconsistency and X% contains at least one gross inconsistency, which might alter the conclusions. This corroborates/disagrees with findings by M. B. Nuijten, Hartgerink, Assen, Epskamp, & Wicherts (2015) who found similar/dissimilar results.

The prevalence of statistical reporting inconsistencies in management research: A replication of Nuijten et al. (2015)

Introduction

Incorrect reporting of findings in science has a trickle-down effect; not only are the conclusions potentially affected, but all reuse of those findings are tainted (e.g., they bias meta-analyses). The use of Null Hypothesis Significance Testing (NHST) is widespread in management research (Lockett, McWilliams, & Van Fleet, 2014; Orlitzky, 2012; Schwab, Abrahamson, Starbuck, & Fidler, 2011) and the reporting of statistical test results (e.g., $t(40) = 2.19, p = 0.03$) has proven to be subject to mistakes. Mistakes, or what we will call reporting inconsistencies, can be found across different fields in science (e.g., medicine and psychology; Bakker & Wicherts, 2011; García-Berthou & Alcaraz, 2004; M. B. Nuijten et al., 2015) and occur in approximately 1 out of 10 reported results. Given that any empirical researcher reports numerous statistical results over their careers, we are all bound to be affected by such reporting inconsistencies. Some inconsistencies can even affect the statistical significance, which has been indicated to happen in approximately 1 out of 8 papers (M. B. Nuijten et al., 2015).

There is no reason to assume that management and organization research would not be afflicted by such inaccuracies. An assessment of the prevalence of statistical inconsistencies in the field of management and public administration is required to uphold the trustworthiness in our results and theories. Goldfarb & King (2016) already found that effect sizes in management research are inflated by around 24-40%; others noted that "honest mistakes and possible scientific misconduct pose a worrisome threat to the trustworthiness of accumulated knowledge" (Bergh, Sharp, & Li, 2016, p. 2).

In this paper, we present the results of a direct replication (M. B. Nuijten et al., 2015) investigating reporting inconsistencies in management and organization research. We investigate the prevalence of statistical reporting inconsistencies in 33 leading management and public administration journals across a timespan of twenty years. A reporting inconsistency can occur when either the test statistic, the degrees of freedom, or the resulting p-value is misreported. Given its substantive importance, we focus on whether the p-value matches the reported test statistic and degrees of freedom. An inconsistent p-value can arise from misreporting any of these three reported results. Nonetheless, the misreporting of any of these values has consequences for the drawn conclusions, be them with respect to the underlying theory or effect of interest.

Reporting inconsistencies are primarily the result of honest mistakes, but can also be the result of purposeful misreporting. Honest mistakes can be manifold, of which the following two are non-exclusive illustrations. First, authors can mistakenly round a p-value. For example, a researcher incorrectly rounds a p-value after their child has kept them up all night, resulting in a p-value of 0.056 being rounded as $p = 0.05$. Second, a researcher might make a minor typographic error. For example, $F(2, 56) = 1.203, p < .001$ instead of $F(2, 56) = 12.03, p < .001$. The latter example produces a reporting inconsistency, without the p-value being incorrect (M. B. Nuijten et al., 2015, p. 10).

Authors sometimes engage in intentionally misreporting of p-values to make the result come across as statistically significant while it actually was nonsignificant. Banks et al. (2016) find in a large-scale survey of management scholars that more than 10% of their respondents engaged in this form of questionable research practice in at least one study, confirmed by C. H. Hartgerink, Aert, Nuijten, Wicherts, & Assen (2016) where 14% of p-values reported as .05 (statistically significant) were in fact larger than .05 (statistically nonsignificant). Given the serious consequences of reporting inconsistencies, whether accidental or purposeful, we attempt to systematically document the prevalence of statistical reporting inconsistencies in the fields of management and organization research.

This paper is a direct replication of M. B. Nuijten et al. (2015) and our results can be directly compared. As such, it provides a first estimate whether reporting inconsistencies are just as prevalent, more prevalent, or less prevalent in the management and organization research fields, when compared to psychology. Additionally, we offer several solutions that might help to partly solve the problem of reporting inconsistencies in the future.

Methods

Sample

The first- and second author compiled a list of 35 journals in management and organization research that are analyzed for reporting inconsistencies. These journals are (primarily) empirical journals and are widely read throughout these fields. For each journal, we collected the articles published from 1995 through 2015 from CrossRef with the command line utility `getpapers` (`v0.4.9`; ContentMine, 2016a), and subsequently downloaded all articles in `HTML` and/or `PDF` format available within the University of Cambridge subscription with `quickscrape` (`v0.4.7`; ContentMine, 2016b). Table X depicts the list of journals and the downloaded articles per file format.

In order to scan the collected articles for reporting inconsistencies, we applied the `R` package `statcheck` (`v1.2.2`; M. B. Nuijten et al., 2015). `statcheck` extracts statistical test results and recalculates p-values based on the reported test statistics and their degrees of freedom. `statcheck` executes the procedure in four steps. First, statcheck processes an `HTML` or `PDF` file into a readable format. `PDF` files are more problematic given the document structure, and `HTML` is to be preferred (M. B. Nuijten et al., 2015). For example, text is frequently placed in multiple columns, where a test result might span multiple columns and will not be properly extracted in the conversion of the `PDF` file due to the way this document is structured. `HTML` files have fewer processing problems and are hence preferred.

`statcheck` extracts t, F, r, χ^2, and Z test results from the text and checks whether there might be a reporting inconsistency. Considering that `statcheck` is an automated procedure, it should be regarded as identifying potential reporting inconsistencies and should not be considered definitive. The algorithm is currently capable to read results that are reported in the format prescribed by the American Psychological Association (APA). This format dates back to 1983 (American Psychological Association, 1983, 2001, 2010), encompassing the timespan we investigate (i.e., 1995-2015). For example, an APA formatted F-test is reported as $F(1, 238) = 2.94, p = 0.09$.

Based on the reported t, F, r, χ^2 or Z test results (and degrees of freedom), `statcheck` recalculates the p-value and compares this to the reported p-value. This recalculation assumes that the test result is correctly reported and that the p-value is two-tailed. However, to catch potential one-tailed tests, `statcheck` searches the article for any mentions of a one-tailed test and does not consider it a reporting inconsistency if the recalculated p-value divided by two is equal to the reported p-value.

If the recalculated p-value differs from the reported p-value, `statcheck` considers this a reporting inconsistency; if the statistical significance of the recalculated p-value is different from the reported p-value, this is considered a decision inconsistency. As such, a decision inconsistency are those inconsistencies that warrant the most attention, given that they might alter the substantive conclusions (depending how important the result is to the main findings). In order to prevent unnecessary overdetection of reporting inconsistencies, the algorithm takes into account potential rounding errors in the test-value (e.g., when rounded to two decimal places, a value of 1.22 can be the result of anything from 1.215 through 1.224). We investigate decision inconsistencies under $\alpha = .05$ (the default of `statcheck`) and $\alpha = .10$.

The advantage of the automated procedure is that it allows us to assess the prevalence of reporting inconsistencies on a large scale. Furthermore, the automated procedure eliminates human errors which are bound to be made when results are recalculated by hand. The disadvantage of an automated procedure is that it will miss statistical tests that are not reported according to APA standards and can introduce machine error when the algorithm is misspecified or unable to handle specific cases (e.g., corrected p-values; Schmidt, 2016). An extensive validity check, where manually extracted results from a set of research papers was compared to the results after applying `statcheck` to the same research papers, indicated that the inter-rater reliability between manual and automated was 0.76 for reporting inconsistencies and 0.89 for decision inconsistencies (M. B. Nuijten et al., 2015). Nonetheless, the algorithm might incorrectly find reporting inconsistencies when corrected p-values are reported (Schmidt, 2016).

`statcheck` is currently not designed to read results that are reported in tables. Therefore, we are unable to assess the prevalence of errors in tables that report regression results, for example. Given that regression tables are also frequent in the fields of management and organization research, the results from this paper should not be generalized to all statistical test results, but only to the APA reported test results.

Analyses

Given that the algorithm is extracts (almost) all APA reported test results, the collected dataset is the population of APA-reported test results for the included journals. Hence, we refrain from using NHST in our analyses and only descriptively model the data. Considering that the extraction quality differs between `HTML` and `PDF` files, we will analyze the results from both separately.

We report the prevalence of (gross) inconsistencies per journal and explore trends of the extracted results over time. We also compare whether gross inconsistencies are more likely for results that are reported as statistically significant as compared to insignificant results. Last, we analyze a number of journal-level covariates to observe whether they influence the number of inconsistencies. We explore whether there are differences in the percentage of inconsistencies across different publishers: we included journals from general publishers such as Wiley, Elsevier, and Sage as well as dedicated publishers INFORMS, the Academy of Management, and one APA journal. We also report regression results for the relationship between journal impact factor and the amount of inconsistencies.

Since many journals only publish HTML articles for more recent years, we also downloaded all articles in PDF and used statcheck on these articles. As explained above, the conversion to text files is less reliable for PDFs than for HTML files. We therefore use this robustness check as a sensitivity analysis and the results ought to be interpreted with caution.

References

American Psychological Association. (1983). *Publication manual of the American Psychological Association* (3rd ed.). Washington, DC: American Psychological Association.

American Psychological Association. (2001). *Publication manual of the American psychological association*. Washington, DC: American Psychological Association.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: American Psychological Association.

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. doi:10.3758/s13428-011-0089-5

Banks, G. C., O'Boyle, E. H., Pollack, J. M., White, C. D., Batchelor, J. H., Whelpley, C. E., … Adkins, C. L. (2016). Questions about questionable research practices in the field of management: A guest commentary. *Journal of Management*, *42*(1), 5–20. doi:10.1177/0149206315619011

Bergh, D., Sharp, B., & Li, M. (2016). Tests for identifying "red flags" in empirical findings: Demonstration and recommendations for authors, reviewers and editors. *Academy of Management Learning & Education*. doi:10.5465/amle.2015.0406

ContentMine. (2016a). getpapers. Retrieved from https://github.com/contentmine/getpapers

ContentMine. (2016b). quickscrape. Retrieved from https://github.com/contentmine/quickscrape

García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and P values in medical papers. *BMC Medical Research Methodology*, *4*(1), 13.

doi:10.1186/1471-2288-4-13

Goldfarb, B., & King, A. A. (2016). Scientific apophenia in strategic management research: Significance tests & mistaken inference. *Strategic Management Journal*, *37*(1), 167–176. doi:10.1002/smj.2459

Hartgerink, C. H., Aert, R. C. van, Nuijten, M. B., Wicherts, J. M., & Assen, M. A. van. (2016). Distributions of *p*-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. doi:10.7717/peerj.1935

Lockett, A., McWilliams, A., & Van Fleet, D. D. (2014). Reordering Our Priorities by Putting Phenomena before Design: Escaping the Straitjacket of Null Hypothesis Significance Testing. *British Journal of Management*, *25*(4), 863–873. doi:10.1111/1467-8551.12063

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2015). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*. doi:10.3758/s13428-015-0664-2

Orlitzky, M. (2012). How can significance tests be deinstitutionalized? *Organizational Research Methods*, *15*(2), 199–228. doi:10.1177/1094428111428356

Schmidt, T. (2016). Sources of false positives and false negatives in the STATCHECK algorithm: Reply to Nuijten et al. (2015). *ArXiv E-Prints*.

Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers Should Make Thoughtful Assessments Instead of Null-Hypothesis Significance Tests. *Organization Science*, *22*(4), 1105–1120. doi:10.1287/orsc.1100.0557