Do p-value misreporters cover their tracks? How readability of the text can help in detecting statistical errors.

Aurelio Sabini
Tilburg University
a.sabini@uvt.nl

Noah Smeets
Tilburg University
n.smeets@uvt.nl

Sanam Vaswani
Tilburg University
p.p.vaswani@uvt.nl

Marco Wedemeyer
Tilburg University
m.a.wedemeyer@uvt.nl

Sen Yang
Tilburg University
s.yang@uvt.nl

Abstract

A growing stream of literature argues that the current academic climate seems to push scholars to engage in questionable research practises, such as p-hacking. The primary aim of this work is to determine whether a feature such as language complexity can be regarded as a sound indicator in detecting misreporting of results, particularly p-values, in scientific literature. The R package Statcheck, in combination with a set of readability indexes, were used to check the existence of misreported p-values for 129 Academy of Management Journal (AMJ) articles. We apply mean difference tests to represent the relationship between p-value misreporting and readability scores. We find that, contrary to what was speculated, readability is insignificantly correlated to p-value misreporting. The better part of the readability indexes unveils this pattern, yet only in one case the relationship is statistically significant.

Introduction

The scientific community works by sharing results and building on each other's' work. The compensation schemes are built around this aim. Namely, for scientists to get tenure they have to publish a certain amount of articles in top tier journals. This system is heavily criticised as it pressures researchers to make an unpleasant choice: publish or perish. This does no good to the quality of scientific work, as unpolished papers are issued, work that should be bundled together is split up into separate papers, and some scholars resort to short cuts or questionable research practices to make their work more interesting to top tier journals. In this report we focus on one of these questionable practises: P-hacking, in the form of statistical misreporting (i.e. the reported t-statistic does not match the reported p-value.)

Statistically significant effects are more likely to get published, and getting published can play a substantial role in the upward progression of a scholar's career. According to a team of researchers (David Chavalaria et al., 2016), the number of studies published containing p-values in their abstracts has doubled from 1990 to 2014 and of those studies that included a p-value 96% were below .05. This bias of journals for significant results ultimately may lead scholars to be drawn to the lure of p-hacking. P-hacking is defined as artificially altering non-significant results so that they become significant (e.g. collecting or selecting data for statistical analyses until significant effects are found (Head et al., 2015)). Recent work by Baum and Bromiley (2018) and Head et al. (2015) provides evidence of the existence of the phenomenon in top-tier journals, however, singling out papers which have engaged in p-hacking remains difficult.

In other domains it has been found that when people engage in questionable behavior they often also engage in obfuscation. Obfuscation is defined as "the production of misleading, ambiguous and plausible but confusing information as an

act of concealment or evasion" (Brunton & Neissenbaum, 2013, p.164). One way in which obfuscation in text has been detected is by examining the readability of the text. There are several ways to measure readability but all have to do with the complexity of the language used, such that more readable texts use less complex language. Readability has for example been found to be related to the sentiment of texts of annual financial statements of companies (e.g. Courtis, 1998; Moffitt & Burns, 2009; Laksmana Tietz, and Yang, 2012). When the financial news conveyed was more negative, the language used to convey it was more complex. Researchers who misreport and researchers who report accurately might differ in the extent to which they engage in obfuscation. If we assume that researchers who use p-hacking attempt to obfuscate their work, we might be able to find a relation between the level of readability of the paper and whether a statistical error was made.

Goal

The goal of this project is to examine whether there is a significant correlation between the misreporting of p-values (purposefully or not) and the complexity of the language used in the papers. The present study speculates that authors who have used questionable research practices are (sub)consciously aware of this and may use vague or convoluted language to report their findings. This obfuscation serves to hide the fact that the results might not be as significant as they first appear.

Relevance

This research makes a theoretical contribution to the literature on readability and obfuscation, as well as to the stream of literature around p-hacking. To the best of our knowledge, no prior research exists which examines the effect of readability on statistical errors. When people engage in obfuscation, readability has been shown to be a good predictor (Courtis, 1998). The present research contributes by applying this rational to a new context. Moreover, investigating whether readability scores can detect the obfuscation of questionable research practises opens the door for further analysis of linguistic aspects to detect (the lack of) good conduct.

Even though the p-value is set arbitrarily and is no substitute for sound scientific reasoning, its strength lies in the use as a cut-off by all scholars. The combination of journal's bias towards significant effects, and researchers pressure to publish, ultimately weakens the quality of published scientific work. The practical relevance of this piece lies in calling further attention to this issue, and additionally functions as a first attempt to detect p-hacking.

Approach

The data on which the following analyses are based are drawn from the text of 1104 AMJ articles in HTML file format. All articles are taken from the years 2002 to 2019. The Academy of Management Journal publishes empirical studies contributing to management practice based on quantitative, field, laboratory, meta-analytic, and mixed methods. Importantly, the final analysis was conducted on 129 articles, as the program used to determine statistical errors was limited in the sense that it was unable to read tables.

For the analyses, this study makes use of the *Beautiful Soup* Python package in order to parse the HTML files. The *RE* (regular expressions) package was used to clean the articles of graphs and tables, and generally preparing them for data analysis. The R package *Statcheck* is used to check whether an article includes misreported p-values. The readability scores of each article are evaluated by means of the Python package *Textstat*, which determines readability scores of a particular corpus. In the initial stage of the analysis, a selection of scores that are widely used in practice to allow for generalisability is chosen. The main scores and indices included in the package are the following:The *Flesch-Kincaid Grade Level*, *SMOG* index, *FOG*scale, and *Automated Readability Index* are measure of readability that estimates the years of education needed to grasp the meaning of a piece of writing, for instance, a score of 9 on these indexes means that a ninth grader would be able to read the document. Dif-

ferent from other tests, Dale-Chall Readability Score uses a lookup table of the most commonly used 3000 English words and returns the grade level using the NewDell-Chall Formula, whereas, the Linsear Write was developed for the United States Air Force to help them calculate the readability of their technical manuals. The hypothesis was tested by performing mean difference tests to detect whether articles with misreported p-values had higher readability scores. The distributions of the readability scores where also visualised to provide a visual check (figure 1, appendix).

Results

Due to the limited amount of data, it was difficult to come to a sound conclusion. Most of the readability scores are insignificant (table 1, appendix). Interestingly, the Coleman Liau index was significant, but the direction is opposite of what was hypothesized; the papers in which the p-value was misreported had lower scores on readability, making them easier to read as opposed to harder. Three out of the six remaining readability indexes showed a similar pattern, although the difference was not significant. A characteristic of the Coleman Liau index, which sets it apart from the other indices, is that it considers the character length of the words rather than the syllables. This feature was a functional adaptation rather than a theoretical one. Not using syllable boundaries made this score easier to code (Coleman, Meri & Liau, 1975). The validity of this scoring method has also been called into question. Due to the small sample size and its in-congruence with the other measures we do not wish to draw hasty conclusions from these results. Further study using more data is necessary to more accurately determine whether there is an effect. Notably, out of the 129 papers which reported the p-value and t-statistic in text, 41 (31.8%) misreported a p-value, which shows again that more care should be given to the accuracy of statistical reporting during the reviewing process.

Figure 2 and figure 3 (appendix) show that there are deviations in p-value misreporting and readability scores over time. The first graph shows that the rate of misreporting per year is slightly decreasing, however, with the small sample size per year, this trend is not generalisable. The second graph shows the normalised readability scores per year. An interesting trend is the decrease in readability scores (easier texts) in 2018 and 2019.

Evaluation & Future Directions

The goal of this research was to determine whether there is a relationship between readability and statistical misreporting. Our hypothesis was that researchers who misreport obfuscate their work, and therefore have higher readability scores. The results indicate that there likely is no significant effect of readability on statistical errors, however, more data is necessary to solidify this finding. The findings suggest that researchers who misreport, and those who report accurately, write with a similar linguistic complexity. Our hypothesis is not supported by the data, and actually contradicted by the one significant effect of the Coleman Liau index. It could be interesting to explore the possibility of the Coleman Liau reflecting the true relationship.

That the Coleman Liau index was significant in the opposite direction of our hypothesis is surprising in light of previous research on readability and obfuscation. For example, Laksmana et al. (2012) showed that organizations which attempt to hide negative financial results use more complex language (i.e. statements about finances which convey negative news are less readable than their positive counterparts). Perhaps, researchers who misreport p-values do not attempt to obfuscate their misreporting by shrouding it in complex language, however; what could explain that the papers with statistical errors were *more* readable? Our hunch to the answer is related to a controversial thought, that perhaps readability correlates with the chance of statistical error as well as the level of competence of the researcher. Further analysis of the text on a more fine-grained level could perhaps give some further insights, but this is beyond the scope of this project.

Future research could delve into the different linguistic aspects of readability scores to more precisely define how papers with and without statistical errors differ in terms of use of language. Interesting textual features to examine for example are: the length of sentences, the complexity of words used (length, amount of syllables, or meaning), the length of the paper itself or even to what extent the author makes strong claims (are they more or less cautious when interpreting their findings). Moreover, there might be specific words which signal misreporting, which could be identified by determining what words occur often in papers with statistical errors, and few times in papers without such errors. However, a larger database of papers would be necessary for such tasks.

A substantial limitation of this project is the amount of data used. We were limited mainly by the StatCheck program which was only able to detect misreporting of p-values when they were explicit in text. Most papers report their p-values in a table only. To our knowledge, there is no algorithm available which is able to very accurately get the necessary data out of tables in papers (not to mention the large variation in presenting tables across journals). Finding, refining, or developing a tool which can identify p-values in tables would instantly greatly increase the amount of data available for this task.

It is important to keep in mind that we used data only of one journal in the field of Management (AMJ). The social sciences have been shown to suffer from p-value misreporting multiple times, but in the medical field as well this is becoming a widely acknowledged issue (Bruns & Ioannidis, 2016). Nonetheless, we can not assume that our findings will generalize to other domains, which is why similar studies (albeit with more data) should be conducted in other fields as well. This will definitely give more insights to this topic. However, when one is doing cross-discipline validity, keep in mind that language used in different fields might be in nature more complex than the one used in another study.

Future work could extend this project by taking into account other potential predictors of statistical misreporting such as; journal specific criteria (e.g. journal prestige, impact factors), author specific criteria (e.g. the number of authors, publication rate), institution specific criteria (e.g. ethical codes, number of scandals), and other paper specific criteria (e.g. topic modelling, publication year). A set of these predictors might be able to detect misreporting with some degree of accuracy, and subsequently be developed into a usable application. However, whether a witch hunt on p-hackers is desirable is highly debatable. P-hacking is a symptom of a broken system. We would suggest a different approach which tackles the root of the problem. This would include a rigorous examination of the current publication system and a serious consideration of alternatives, such as the As-You-Go approach (Hartgerink, Van Zelst, 2018).

References

Baum, J. A., & Bromiley, P. P-hacking in Top-tier Management Journals.

Brunton, F., & Nissenbaum, H. (2013). Political and ethical perspectives on data obfuscation. Privacy, due process and the computational turn: The philosophy of law meets the philosophy of technology, 164-188.

Bruns, S. B., & Ioannidis, J. P. (2016). P-curve and p-hacking in observational research. PloS one, 11(2), e0149144.

Chavalarias, D., Wallach, J. D., Li, A. H. T., &

Ioannidis, J. P. (2016). Evolution of reporting P values in the biomedical literature, 1990-2015. Jama, 315(11), 1141-1148.

Coleman, Meri; and Liau, T. L. (1975); A computer readability formula designed for machine scoring, Journal of Applied Psychology, Vol. 60, pp. 283â“284

Courtis, J. K. (1998). Annual report readability variability: tests of the obfuscation hypothesis. Accounting, Auditing & Accountability Journal, 11(4), 459-472.

Hartgerink, C., & Van Zelst, M. (2018). “As-You-Go” instead of “After-the-Fact” : A network approach to scholarly communication and evaluation. Publications, 6(2), 21.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. PLoS biology, 13(3), e1002106.

Laksmana, I., Tietz, W., & Yang, Y. W. (2012). Compensation discussion and analysis (CD& A): Readability and management obfuscation. Journal of Accounting and Public Policy, 31(2), 185-203.

Moffitt, K., & Burns, M. B. (2009). What does that mean? Investigating obfuscation and readability cues as indicators of deception in fraudulent financial reports. AMCIS 2009 Proceedings, 399.

Sullivan, Gail M., and Richard Feinn. "Using effect sizeâ”or why the P value is not enough." Journal of graduate medical education 4.3 (2012): 279-282.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASAâ^{TM}s statement on p-values: context, process, and purpose. The American Statistician, 70(2), 129-133.

Gray, W. S. and B. Leary. 1935. *What makes a book readable*. Chicago: Chicago University Press.

Appendix

Figure 1.

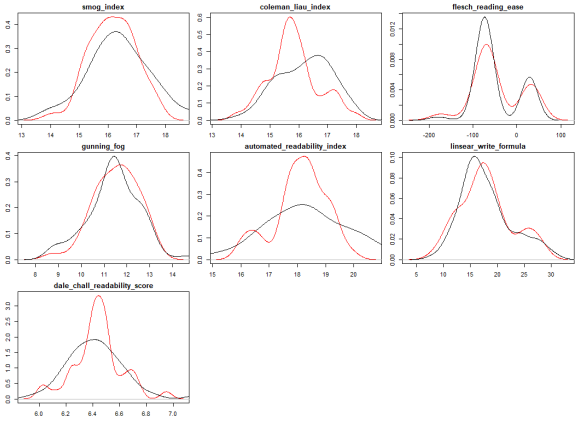

Figure 2.

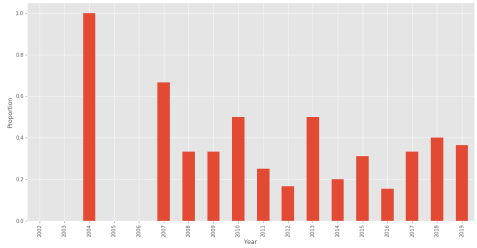

Figure 3.

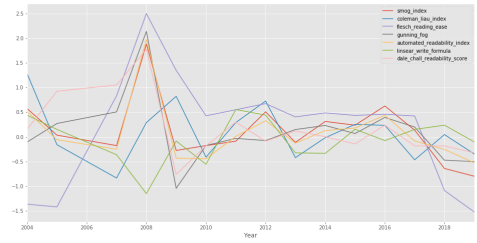

Table 1.

	Smog Index	Coleman Liau Index	Flesch Reading Ease	Gunning Fog	Automated Readability Index	Linsear Write Formula	Dale Chall Score
Estimate	-0.2182	-0.4374	2.8342	0.0176	-0.3347	-0.2311	0.0126
P-value	0.2594	0.0202	0.7753	0.9377	0.2292	0.8099	0.7495
Std.Error	0.1926	0.186	9.9055	0.2244	0.277	0.9587	0.0394
$\alpha < .05$	No	Yes	No	No	No	No	No

Table 2.

Members' Contribution

	Idea	Data selection	Pre-processing	Code	Evaluation	Writing paper	Proof-reading
Aurelio Sabini						X	X
Noah Smeets	X	X			X	X	X
Sanam Vaswani	X						X
Marco Wede-meyer	X	X	X	X	X	X	X
Sen Yang						X	X