

Statistical Inconsistencies in Experimental Linguistics

Timo B. Roettger and Dara Leonard Jenssen Etemady

Department of Linguistics & Scandinavian Studies, University of Oslo

Abstract

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

Keywords: statistics, statcheck, reproducibility

1. Introduction

What we know about human language and its cognitive underpinnings is often informed by experimental data, data based on which researchers draw theoretically relevant inferences using common statistical frameworks. This inference should be transparent in order for other researchers to critically evaluate the inference process and potentially detect and correct human errors. In the recent years, the quantitative sciences have seen repeated calls to become more transparent and reproducible by sharing data and statistical protocols (Arvan et al., 2022; Laurinavichyute et al., 2022; Roettger, 2019), sharing of statistical protocols is still rare across the language sciences (Bochynska et al., 2023). If statistical procedures cannot be critically evaluated, human errors might be left undetected and thus remain uncorrected in the publication record. If these errors affect the decision procedure of the analysis, i.e. whether a hypothesis is rejected or accepted, these errors might lead to -at best- overconfident, -at worst- false theoretical conclusions. The present paper will present evidence that the published literature in experimental linguistics contains a concerning amount of statistical errors, a state of affairs which warrants more rigorous data sharing practices.

2. Statistal reporting inconsistencies

The null-hypothesis significance testing (NHST) framework is, to date, the most dominant statistical framework that researchers use to test hypotheses in the language sciences (Sonderegger & Sós-kuthy, 2024). These statistical tests are reported in specific formats which usually contain a test statistic, the degrees of freedom of that test (if applicable), and the p-value, representing the

probability of observing the data (or more extreme data) given the null hypothesis (i.e. given that the test statistic is zero) (see example 1):

(1) $F(1, 66) = 3.88, p < .05$

Since data and statistical protocol sharing remains rare across the language sciences (Bochynska et al., 2023), interested readers are left with trusting the authors that the statistical analysis is run and reported correctly. Whether that trust is warranted can be assessed, however. The three sets of indices in (1) have a clearly defined mathematical relationship and can thus be checked for consistency. An F test with the specified degrees of freedom and a test-statistic of 3.88 should result in a p-value of 0.053 which is larger, not smaller, than 0.05. Thus, something is wrong with the statistical report in (1). Possible reasons for this inconsistency are manifold: It could be a typo of the comparison sign, i.e. the authors meant to use = or > rather than <. Alternatively, any of the numbers could be a typo. Sometimes an error might indicate erroneous rounding (e.g. 0.057 being rounded down to 0.05). Other times, human error along the data analytical pipeline might have caused an error down the line. Without access to data and scripts, it remains unclear to the reader, what has caused the inconsistency. Such inconsistencies can be particularly concerning if the calculated p-value and the reported p-values are not on the same side of the alpha threshold. In NHST, p-values below a conventionalized alpha threshold, most commonly 0.05, are interpreted as evidence that the data are sufficiently inconsistent with the null hypothesis (significant). p-values above that threshold are considered consistent with the null hypothesis and practically lead to rejecting the alternative hypothesis (not significant). In (1) above, the reported p-value suggest a significant result, the p-value derived from the degrees of freedom and the test statistic suggest a non-significant result. In the following we refer to these inconsistencies as “decision inconsistency”.

This form of inconsistency assessment can be automatically assessed if statistical tests are reported in an unambiguous format. Recently, a series of studies used such automatic assessments to evaluate the prevalence of inconsistent statistical reporting across different disciplines (Buckley et al., 2023; Green et al., 2018; Groß, 2021; e.g. Nuijten et al., 2016; Nuijten & Polanin, 2020). For example, looking at over 250000 p-values published in major psychology journals, Nuijten et al. (2016) found that around 50% of the articles with statistical results contained at least one inconsistencies and around 12.5% contained at least one “decision inconsistency”.

To assess the prevalence of statistical-reporting inconsistencies in experimental linguistics, the present paper conceptually replicates Nuijten et al. (2016) and assesses over 39000 p-values reported in eight experimental linguistic journals published between 2000 and 2023. We further assess whether the rate of inconsistency differs across journals, whether that rate has changed over the course of the last 20 years and whether there is evidence for bias in these statistical-reporting inconsistencies. We discuss the results and offer concrete recommendations for authors, reviewers, and editors to tackle this problem.

3. Method

All quantitative analyses were conducted using R [] and the r packages [LIST ALL PACKAGES from 01_Analysis.R and add to references]

3.1. Statcheck

We used the R package statcheck (Version 1.4.1-beta.2, Nuijten & Epskamp, 2023) to automatically detect statistical-reporting inconsistencies. Statcheck works as follows: After converting

pdf or html articles to plain text, statcheck searches for specific strings that correspond to a NHST result, using “regular expressions”. That way, statcheck can detect results of t tests, F tests, Z tests, χ^2 tests, correlation tests, and Q tests as long as the test result fulfills three conditions: (a) the test result is reported completely including the test statistic, degrees of freedom (if applicable), and the p-value; (b) the test result is in the body of the text, i.e. it usually misses results reported in tables; and (c) the test result is reported in American Psychological Association style (APA, 2019). Given these constraints, statcheck is estimated to detect roughly 60% of all reported NHST results (Nuijten et al., 2016). Statcheck uses the reported test statistic and degrees of freedom to recalculate the p-value, compares the reported and recalculated p-values and, if there is a mismatch, reports a comparison as an “inconsistency.” The algorithm takes into account that tests might have been performed as one-tailed by identifying the search strings “one-tailed,” “one-sided,” or “directional”. Moreover, statcheck counts $p = .000$ and $p < .000$ as inconsistent as p-values of exactly zero are mathematically impossible and the APA manual (American Psychological Association, 2020) advises to report very small p-values as $p < .001$. Validity checks suggest that inter rater reliability between manual coding and statcheck is high, i.e. 0.76 for inconsistencies and 0.89 for decision inconsistencies (Nuijten et al., 2016). The overall accuracy of statcheck is estimated between 96.2% to 99.9% (Nuijten et al., 2017). We thus consider statcheck a valid, but rough proxy of statistical reporting inconsistencies.

3.2. Sample

Focusing on experimental linguistic research, we started with Kobrock and Roettger (2023) as a point of departure who list 100 linguistic journals that had at least a hundred articles published at the time of assessment (2021) and a high ratio of articles containing the term “experiment* in title, abstract or keywords. Out of these 100 journals, we selected all journals with at least 10% of articles containing the search string “experiment* “. Out of the remaining 37 journals, we selected those journals that urged APA formatting either in the main body of the text or specifically regarding statistics in the author guidelines, resulting in nine remaining journals. Moreover, to download the pdfs, the articles had to be either accessible to us through our library license, or open access, resulting in a final list of eight journals: Applied Psycholinguistics (APS), Bilingualism: Language and Cognition (BLC), Linguistic Approaches to Bilingualism (LAB), Language and Speech (LaS), Language Learning and Technology (LLT), Journal of Language and Social Psychology (LSP), Journal of Child Language (JCL), and Studies in Second Language Acquisition (SLA).

We included only original research articles within the publication years of 2000-2023, excluding any book reviews, response articles, commentaries, editorials, corrigenda, errata, advertisements, etc. Articles from LAB spanned 2011-2023, while the rest spanned 2000-2023. This procedure resulted in 5804 research articles. All 5804 articles were submitted to analysis, using Statcheck v1.5.0. 157 articles were not analyzable by statcheck, and thus had to be removed from the pool, likely related to issues with rendering the Chi-Squared symbol being erroneously converted from .pdf to .txt.

3.3. Data availability statement

All derived data and corresponding R scripts are publicly available here: [LINK](#). The original journal articles cannot be bulk-shared due to distribution restrictions by the publishers. NEED TO FIGURE OUT A WAY

Table 1

Summary. Applied Psycholinguistics (APS), Bilingualism: Language and Cognition (BLC), Linguistic Approaches to Bilingualism (LAB), Language and Speech (LaS), Language Learning and Technology (LLT), Journal of Language and Social Psychology (LSP), Journal of Child Language (JCL), and Studies in Second Language Acquisition (SLA).

Journal	eligible articles	assessible articles	assessible results	inconsistencies	decision inconsistencies
APS	953	690	9570	1368	170
BLC	964	610	9093	1161	120
JCL	1109	529	6240	750	69
LAB	471	133	1719	234	25
LLT	421	111	919	201	61
LSP	695	376	4320	429	60
LaS	598	363	4717	552	86
SLA	593	247	2954	471	64
Total	5804	3059	39532	5166	655

4. Results

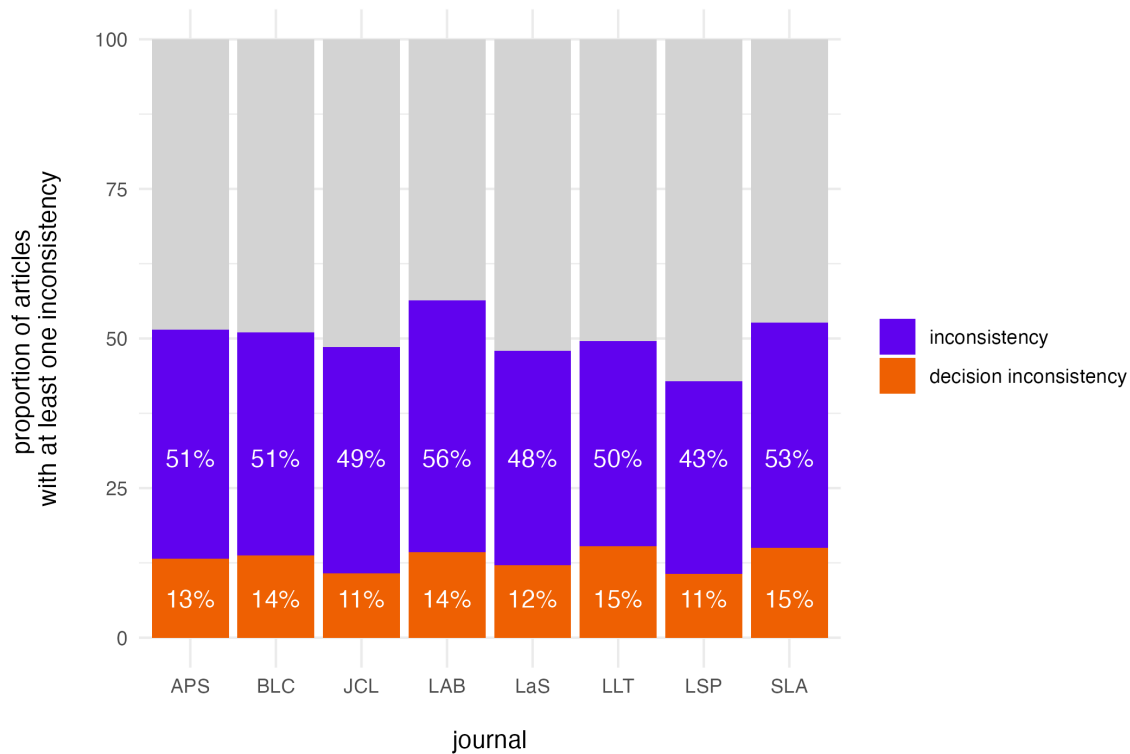
4.1 Prevalence of inconsistencies

The results are summarized in Table 1. Out of 5804 articles, 3059 articles contained statistical tests that statcheck could assess (53%), amounting to 39532 assessable p-values. 5166 p-values were flagged as inconsistent (13.1%) and 655 of which were considered decisions inconsistencies (1.7%) (see Table 1).

The proportion of inconsistencies ranged from 10 to 22% across journals (1 to 7% for decisions inconsistencies) (see Figure 1). These rates appear to be stable across year of publication (see Figure 2). On average, 50% of assessable articles contained one or more inconsistencies (journals range from 43 to 56%) and 13% contained one or more decisions inconsistencies (journals range from 11 to 15%).

Figure 1

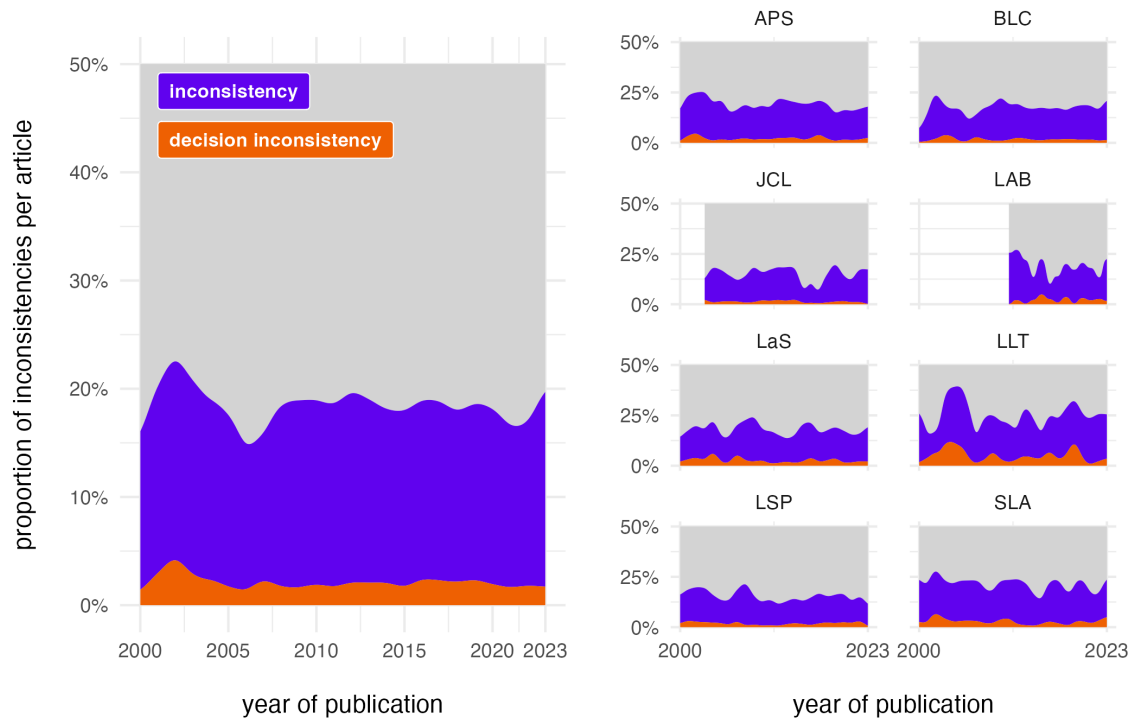
Proportion of articles containing at least one inconsistency / decision inconsistency. Applied Psycholinguistics (APS), Bilingualism: Language and Cognition (BLC), Linguistic Approaches to Bilingualism (LAB), Language and Speech (LaS), Language Learning and Technology (LLT), Journal of Language and Social Psychology (LSP), Journal of Child Language (JCL), and Studies in Second Language Acquisition (SLA).



When examining reported against recalculated p-values (see Figure 3) for all inconsistencies, we can identify certain spatial patterns (only inconsistencies that are based on ‘=’ comparisons were included to make them easier to interpret, $n = 5154$ out of 5166). First, the center of density of points is located in the bottom left corner with more values reported closer to the alpha level. This is not surprising, since there is a documented publication bias in published quantitative articles, with hypotheses being much more often confirmed (and thus $p < .05$) than not (Franco et al., 2014; Sterling, 1959). Second, there is a linear density of points along the diagonal axis which corresponds to numerically small inconsistencies, some of which might be related to simple rounding errors. However, comparing the diagonal to the black line, which represents a linear model predicting recalculated by reported p-values, we can see a clear divergence of what is expected if inconsistencies were equally likely in both directions (i.e. recalculated p-values were as likely to be smaller than the reported p-value as larger). The slope of the regression line is flatter than the diagonal axis which means that, on average, reported p-values are lower than their recalculated counterparts. In other words, inconsistencies have a tendency to lead to smaller p-values. This potential bias can also be observed in decision inconsistencies. Of all decision inconsistencies ($n = 655$), 71% represent cases in which a reported significant result ($p < 0.05$) is recalculated as non-significant value ($p > 0.05$).

Figure 2

Proportion of inconsistencies / decision inconsistencies from 2000 to 2023. Left: global rates with all journals pooled; Right: rates per journal. Applied Psycholinguistics (APS), Bilingualism: Language and Cognition (BLC), Linguistic Approaches to Bilingualism (LAB), Language and Speech (LaS), Language Learning and Technology (LLT), Journal of Language and Social Psychology (LSP), Journal of Child Language (JCL), and Studies in Second Language Acquisition (SLA).



4.2 Manual inspection of decision inconsistencies

To further understand the nature of decision inconsistencies, we manually inspected all of them ($n=655$). We (a) evaluated whether statcheck has extracted the information correctly, and (b) assessed whether the text actually suggest an erroneous inferential decision, i.e. whether the paper uses a reported significant result to claim a significant effect or a reported non-significant result to claim a null result.

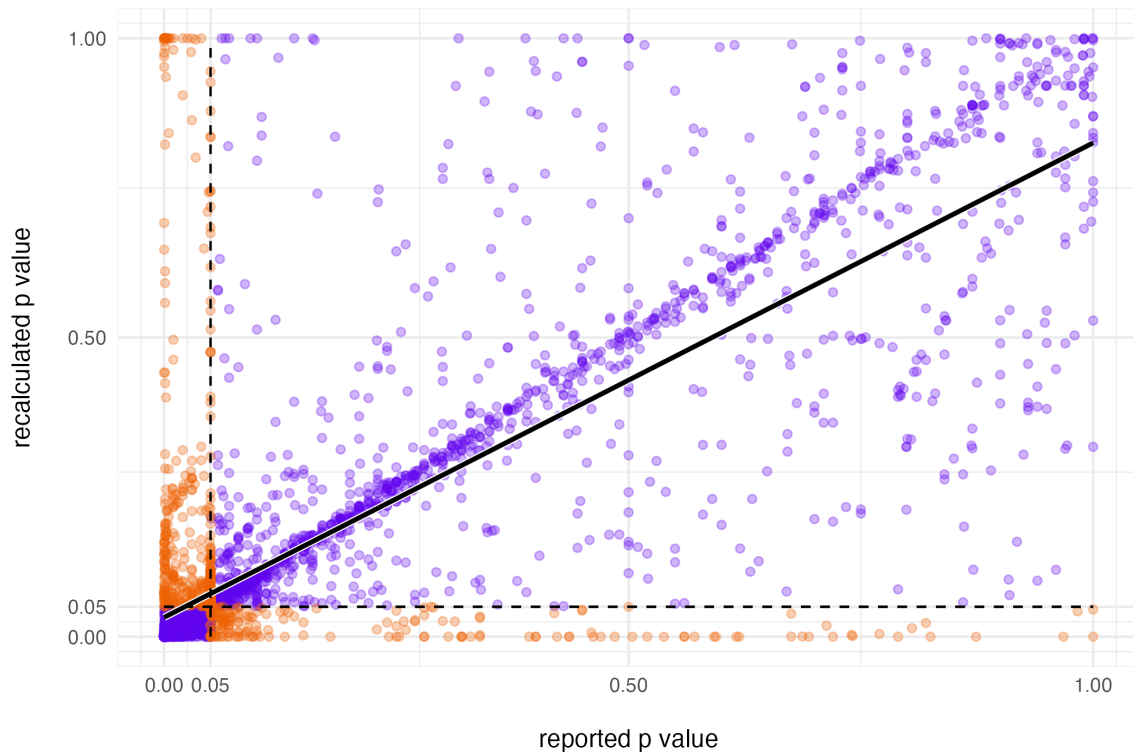
5. Discussion and Recommendations

5.1. Statistical reporting inconsistencies are prevalent

The present study found a concerning amount of statistical reporting inconsistencies across a sample of 5804 experimental linguistic articles, containing 3059 assessable p-values. 13.1% of all p-values were flagged as inconsistent and 1.7% were flagged as decision inconsistencies, i.e. the reported p-value is on the opposite side of the alpha threshold than the recalculated p-value. On

Figure 3

Reported vs. recalculated p-values for inconsistencies that are based on '=' comparisons. Black line descriptively indicates the linear relationship, indicating a bias towards lower reported p-values.



average, 50% of assessable articles contained at least one inconsistency and 13% contained at least one decision inconsistency.

The present study can be considered a conceptual replication of previous studies investigating statistical-reporting inconsistency in psychology (Bakker & Wicherts, 2011, 2014; Caperos & Pardo, 2013; Claesen et al., 2023; Green et al., 2018; Nuijten et al., 2016; Nuijten & Polanin, 2020; Veldkamp et al., 2014; Wicherts et al., 2011), medical sciences (García-Berthou & Alcaraz, 2004,; Van Aert et al., 2023), psychiatry (Berle & Starcevic, 2007), cyber security studies (Groß, 2021), technological education research (Buckley et al., 2023), and experimental philosophy (Colombo et al., 2018). These studies report on inconsistency rates between 4% and 14%, with between 10% and 63% of articles containing at least one inconsistency and between 3% and 21% decision inconsistencies.

Even if the prevalence of these inconsistencies could be largely attributed to inconsequential typos or rounding errors (an assumption we cannot test without access to the data), the sheer amount of these errors that have made it through peer-review should concern us. They are human errors. If such a substantial amount of errors is found in plain site, the question naturally arises how many errors that happen during the data analysis itself remain undetected. If the tip of the iceberg above water is so large, how large is the iceberg underneath the surface?

The present examination also looked at whether the rates of inconsistency changed over the

last 23 years of publications. Descriptive assessment did not indicate any noticeable trend. Moreover, the rates of both inconsistencies and decision inconsistencies seem to remain stable across journals.

Observed inconsistencies were characterized by reported p-values being on average lower than their recalculated counterparts and the prevalence of decision inconsistencies was higher for p-values reported as significant than for those reported as non-significant. This could indicate a systematic bias in favor of lower p-values in general and a bias towards significant results in particular. Our data do not speak to the causes of these biases, but possible reasons include the following: Researchers might intentionally round down p-values because they think a lower p-value is more convincing to reviewers and/or readers. This practice has been admitted to by 1 in 5 surveyed psychological researchers (John et al., 2012). Given that quantitative linguists have been reported to commit questionable research practices (and even fraud) more often than we would like to (Isbell et al., 2022), we cannot exclude the possibility that some of the inconsistencies were intentional. It is our strong belief, however, that the majority of inconsistencies are unintentional.

Researchers might scrutinize non-significant results more than significant results or fail to double check significant results more often because results that confirm their hypothesis feed into their confirmation bias (Nickerson, 1998). For example, Fugelsang et al. (2004) let researchers evaluate data that are either consistent or inconsistent with their prior expectations. They showed that when researchers encounter results that disconfirm their expectations, they are likely to blame the methodology while results that confirmed their expectations were rarely critically scrutinized. Alternatively, the observed bias might be a reflection of publication bias (Franco et al., 2014; Sterling, 1959) with (erroneously) reported significant p-values being more likely to be published than non-significant ones.

5.2. Limitations of our study

While we believe our work offers an important contribution to improving statistical reporting practices in experimental linguistics, there are, of course, a number of limitations. The present assessment and the conclusions we can draw from them are limited. First, our sample is limited to only a subset of experimental linguistic journals. Our sample is based on a crude criterion of what constitutes an experimental linguistic journal (see Kobrock & Roettger, 2023) and we restricted ourselves further to (for us) accessible journals which explicitly require APA statistical formatting in the author guidelines. Thus it is possible that a different selection of journals would have resulted in different results. However, given that the inconsistency rates of our study are comparable to similar studies from other disciplines and that the inconsistency rates are both comparable across the eight journals and stable across the most recent 23 years, suggests that our findings are relevant for experimental linguistics at large.

Second, given the constraints on automatically detecting test statistics, statcheck misses reported values that either diverge from APA reporting standards or are reported in tables. However, inconsistency rates have been shown to be similar for results in APA format vs. results that diverge from APA formatting (Bakker & Wicherts, 2011; Nuijten et al., 2016). Third, statcheck slightly overestimates inconsistency rates, because it might not accurately detect corrections for multiple comparisons (Schmidt, 2017). Nuijten et al. (2017), however, show that not only were there only a small proportion of flagged inconsistencies related to multiple comparisons, but also that these multiple comparisons themselves were often erroneously reported. They conclude that “[a]ny reporting inconsistencies associated with these tests and corrections could not explain the high prevalence of reporting inconsistencies” (Nuijten et al., 2017, p. 27). More elaborate automatic tools for the

extraction of statistical information might allow a more detailed and more accurate assessment of statistical reporting in the future (e.g. [Kalmbach et al., 2023](#)). But even statcheck is limited and only provides a rough proxy of true inconsistency rate in the published literature, we hope the reader agrees that it is a state of affairs that should be tackled.

5.3. Recommendations for the field

There are concrete actionable steps the field of experimental linguistics can make to reduce statistical reporting inconsistencies. In order to avoid simple copy and paste errors related to working in two separate programs for writing the manuscript and conducting the statistical analysis, authors should consider ‘literate programming’, i.e. an integration of analysis code and prose into a single, dynamic document ([Casillas et al., 2023](#); [Knuth, 1984](#)). Several implementations of literate programming are freely available to researchers including common R markdown files (Rmd) and Quarto markdown files (qmd). Literate programming can ensure that values derived by the statistical analysis is automatically integrated into the manuscript document, avoiding errors that might happen during a transfer from one program to the other.

Authors should also consider sharing their derived data (i.e. the anonymized data table that was analyzed) as well as a detailed description of their statistical protocol, ideally in form of reproducible scripts. Sharing reproducible analyses with reviewers allows the reviewers to reproduce the authors’ analyses, possibly detect errors or even inappropriate statistical choices before publication, thus improving the quality of their work. Moreover, publicly sharing their analyses has numerous benefits to the authors themselves beyond error detection: Open data and materials can facilitate collaboration ([Boland et al., 2017](#)), increase efficiency and sustainability ([Lowndes et al., 2017](#)), and are cited more often ([Colavizza et al., 2020](#)).

If authors do not share data and scripts, reviewers can check at least the statistical reporting consistency in the manuscript by using tools such as statcheck ([Nuijten & Epskamp, 2023](#), <http://statcheck.io>) or p-checker ([Schönbrodt, 2015](#), <http://shinyapps.org/apps/p-checker/>). Reviewers could consider requesting data and scripts during peer review. Such requests might be particularly justified when inconsistencies are apparent. Explicitly requesting to share data might already instill additional care and quality checks when authors prepare their materials, but also allows the reviewers to carefully reproduce the results, and critically evaluate all choices made in the statistical analysis ([Sakaluk et al., 2014](#)). Recent evidence suggest experimental linguistics are still characterized by a pluralism of statistical approaches, even when trying to answer the same research question ([Coretta et al., 2023](#)). Some of these approaches are might be more appropriate than others ([Sonderegger & Sósokuthy, 2024](#); [Vasishth, 2023](#)), so more thorough evaluations of how researchers arrive at their statistical conclusions might elevate their analytical robustness.

Journal editors could explicitly recommend consistency checks with apps such as statcheck during peer review, a practice that has been taken up on by several journals from neighboring disciplines (Psychological Science¹, *Advances in Methods and Practices in Psychological Science*², *Stress & Health* [Barber, 2017](#)). Editors could also demand, recommend or at least encourage data sharing for publication in their journal. Data sharing policies have been shown to substantially increase the reproducibility of analyses ([Hardwicke et al., 2018](#); e.g. [Laurinavichyute et al., 2022](#)).

¹http://www.psychologicalscience.org/publications/psychological_science/ps-submissions; accessed on July 15, 2024.

²<https://www.psychologicalscience.org/publications/ampps/ampps-submission-guidelines>; accessed on July 15, 2024.

References

- American Psychological Association. (2020). *Publication manual of the American Psychological Association* (7th ed.). Author. <https://doi.org/10.1037/0000173-000>
- Arvan, M., Pina, L., & Parde, N. (2022). Reproducibility in computational linguistics: Is source code enough? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2350–2361.
- Bakker, M., & Wicherts, J. M. (2011). The (mis) reporting of statistical results in psychology journals. *Behavior Research Methods*, 43, 666–678.
- Bakker, M., & Wicherts, J. M. (2014). Outlier removal and the relation with reporting errors and quality of psychological research. *PloS One*, 9(7), e103360.
- Barber, L. K. (2017). Meticulous manuscripts, messy results: Working together for robust science reporting. *Stress & Health*, 33(2), 89–91.
- Berle, D., & Starcevic, V. (2007). Inconsistencies between reported test statistics and p-values in two psychiatry journals. *International Journal of Methods in Psychiatric Research*, 16(4), 202–207.
- Bochynska, A., Keeble, L., Halfacre, C., Casillas, J. V., Champagne, I.-A., Chen, K., Röthlisberger, M., Buchanan, E. M., & Roettger, T. (2023). Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics*, 2(1).
- Boland, M. R., Karczewski, K. J., & Tatonetti, N. P. (2017). Ten simple rules to enable multi-site collaborations through data sharing. In *PLoS computational biology* (1; Vol. 13, p. e1005278). Public Library of Science San Francisco, CA USA.
- Buckley, J., Hyland, T., & Seery, N. (2023). Estimating the replicability of technology education research. *International Journal of Technology and Design Education*, 33(4), 1243–1264.
- Caperos, J. M., & Pardo, A. (2013). Consistency errors in p-values reported in spanish psychology journals. *Psicothema*, 25(3), 408–414.
- Casillas, J. V., Constantin-Dureci, G., Rascón, I. A., Shao, J., Rodríguez, S. A., Gadamsetty, A., Minetti, A., Laungani, K., Thatcher, J., Gardere, R.-T., et al. (2023). *Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research*.
- Claesen, A., Vanpaemel, W., Maerten, A.-S., Verliefde, T., Tuerlinckx, F., & Heyman, T. (2023). Data sharing upon request and statistical consistency errors in psychology: A replication of wicherts, bakker and molenaar (2011). *Plos One*, 18(4), e0284243.
- Colavizza, G., Hrynaskiewicz, I., Staden, I., Whitaker, K., & McGillivray, B. (2020). The citation advantage of linking publications to research data. *PloS One*, 15(4), e0230416.
- Colombo, M., Duev, G., Nuijten, M. B., & Sprenger, J. (2018). Statistical reporting inconsistencies in experimental philosophy. *PloS One*, 13(4), e0194360.
- Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., Al-Tamimi, J., Alotaibi, N. E., AlShakhori, M. K., Altmiller, R. M., Arantes, P., Athanasopoulou, A., Baese-Berk, M. M., Bailey, G., Sangma, C. B. A., Beier, E. J., Benavides, G. M., Benker, N., BensonMeyer, E. P., ... Roettger, T. B. (2023). Multidimensional Signals and Analytic Flexibility: Estimating Degrees of Freedom in Human-Speech Analyses. *Advances in Methods and Practices in Psychological Science*, 6(3), 25152459231162567. <https://doi.org/10.1177/25152459231162567>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345(6203), 1502–1505.
- Fugelsang, J. A., Stein, C. B., Green, A. E., & Dunbar, K. N. (2004). Theory and data interactions

- of the scientific mind: Evidence from the molecular and the cognitive laboratory. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale*, 58(2), 86.
- García-Berthou, E., & Alcaraz, C. (2004). Incongruence between test statistics and p values in medical papers. *BMC Medical Research Methodology*, 4, 1–5.
- Green, C. D., Abbas, S., Belliveau, A., Beribisky, N., Davidson, I. J., DiGiovanni, J., Heidari, C., Martin, S. M., Oosenbrug, E., & Wainewright, L. M. (2018). Statcheck in canada: What proportion of CPA journal articles contain errors in the reporting of p-values? *Canadian Psychology/Psychologie Canadienne*, 59(3), 203.
- Groß, T. (2021). Fidelity of statistical reporting in 10 years of cyber security user studies. *Socio-Technical Aspects in Security and Trust: 9th International Workshop, STAST 2019, Luxembourg City, Luxembourg, September 26, 2019, Revised Selected Papers 9*, 3–26.
- Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsson, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., et al. (2018). Data availability, reusability, and analytic reproducibility: Evaluating the impact of a mandatory open data policy at the journal cognition. *Royal Society Open Science*, 5(8), 180448.
- Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., Schnur, E., Zhang, M., & Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, 106(1), 172–195.
- John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, 23(5), 524–532.
- Kalmbach, T., Hoffmann, M., Lell, N., & Scherp, A. (2023). On the rule-based extraction of statistics reported in scientific papers. *International Conference on Applications of Natural Language to Information Systems*, 326–338.
- Knuth, D. E. (1984). Literate programming. *The Computer Journal*, 27(2), 97–111.
- Kobrock, K., & Roettger, T. (2023). Assessing the replication landscape in experimental linguistics. *Glossa Psycholinguistics*, 2(1), 1–28.
- Laurinavichyute, A., Yadav, H., & Vasisht, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the journal of memory and language under the open data policy. *Journal of Memory and Language*, 125, 104332.
- Lowndes, J. S. S., Best, B. D., Scarborough, C., Afflerbach, J. C., Frazier, M. R., O'Hara, C. C., Jiang, N., & Halpern, B. S. (2017). Our path to better science in less time using open data science tools. *Nature Ecology & Evolution*, 1(6), 0160.
- Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220.
- Nuijten, M. B., Assen, M. A. van, Hartgerink, C., Epskamp, S., & Wicherts, J. M. (2017). *The validity of the tool “statcheck” in discovering statistical reporting inconsistencies*.
- Nuijten, M. B., & Epskamp, S. (2023). *Statcheck: Extract statistics from articles and recompute p-values(1.4. 1-beta. 2)[r]*.
- Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, 48, 1205–1226.
- Nuijten, M. B., & Polanin, J. R. (2020). “Statcheck”: Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research Synthesis Methods*, 11(5),

- 574–579.
- Roettger, T. B. (2019). Researcher degrees of freedom in phonetic research. *Laboratory Phonology*, 10(1).
- Sakaluk, J., Williams, A., & Biernat, M. (2014). Analytic review as a solution to the misreporting of statistical results in psychological science. *Perspectives on Psychological Science*, 9(6), 652–660.
- Schmidt, T. (2017). *Statcheck does not work: All the numbers. Reply to nuijten et al.(2017)*.
- Schönbrodt, F. D. (2015). *P-checker: One-for-all p-value analyzer*. <http://shinyapps.org/apps/p-checker/>.
- Sonderegger, M., & Sóskuthy, M. (2024). *Advancements of phonetics in the 21st century: Quantitative data analysis*.
- Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association*, 54(285), 30–34.
- Van Aert, R. C., Nuijten, M. B., Olsson-Collentine, A., Stoevenbelt, A. H., Van Den Akker, O. R., Klein, R. A., & Wicherts, J. M. (2023). Comparing the prevalence of statistical reporting inconsistencies in COVID-19 preprints and matched controls: A registered report. *Royal Society Open Science*, 10(8), 202326.
- Vasishth, S. (2023). Some right ways to analyze (psycho) linguistic data. *Annual Review of Linguistics*, 9(1), 273–291.
- Veldkamp, C. L., Nuijten, M. B., Dominguez-Alvarez, L., Van Assen, M. A., & Wicherts, J. M. (2014). Statistical reporting errors and collaboration on statistical analyses in psychological science. *PloS One*, 9(12), e114876.
- Wicherts, J. M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PloS One*, 6(11), e26828.

Appendix

Title for Appendix