# From Detection to Correction: A Hybrid NLP Approach to Misinterpretations of Nonsignificant *p* Values

Raphael Merz

Department of Psychology, Ruhr University Bochum

## Author Note

Raphael Merz https://orcid.org/0000-0002-9474-3379

Correspondence concerning this article should be addressed to Raphael Merz, Email: raphael.merz@rub.de

**Abstract**

Misinterpretations of *p* values remain a highly prevalent issue in scientific reporting, despite decades of educational efforts and reform initiatives. Among the most frequent and consequential misinterpretations is the conclusion that a statistically nonsignificant result (e.g., $p > .05$) implies the absence of an effect – a claim not supported by the logic of null hypothesis significance testing (NHST). This project draws on a human factors perspective, arguing that automation can offer practical, scalable solutions to persistent statistical errors – comparable to how word processors flag potential spelling and grammar mistakes. This master's thesis project proposes the development of an automated tool to detect, classify, and correct such misinterpretations using a combination of rule-based searches, large language models (LLMs), and machine learning classifiers. Building on the existing papercheck framework – an R package created to make automated checks of academic manuscripts easier and more systematic – the project aims to identify statements interpreting nonsignificant results, determine whether these interpretations are correct, and suggest improved phrasing if they are not. Initial detection will rely on rule-based text searches to locate candidate sentences, which will then be filtered and contextualized using LLMs. Classification of interpretations as correct or incorrect will be achieved through transformer-based classifiers (BERT, SciBERT, PubMedBERT), which will be evaluated against human-coded ground truth data. In its final form, the tool will serve as a writing assistant, a research instrument for large-scale corpus analysis, and an extension of papercheck. Ultimately, the goal is to reduce misinterpretations of nonsignificant findings and contribute to more accurate and informative scientific reporting.

*Keywords:* p value, misinterpretation, automation, automated checks, RegEx, LLMs, BERT

# From Detection to Correction: A Hybrid NLP Approach to Misinterpretations of Nonsignificant *p* Values

## 1 Introduction

Over the past decades, numerous articles have addressed common misinterpretations of *p* values in the context of standard null hypothesis significance testing (NHST) (Goodman, 2008; Greenland et al., 2016; Schervish, 1996). Some go further, questioning the use of frequentist methods altogether (e.g., Edwards et al., 1963; Wagenmakers, 2007), while others propose refinements within the frequentist framework that aim to improve the informativeness of statistical inference (e.g., Isager & Fitzgerald, 2024; Lakens et al., 2018). If you are a researcher writing a paper and want to interpret your results correctly, the solution seems simple: read these educational resources and revise your manuscript accordingly. Easy, right? Still, empirical studies consistently show that these misinterpretations remain widespread (e.g., Hoekstra et al., 2006; Murphy et al., 2025). Why is that? What makes interpreting *p* values so persistently difficult? Which practical solutions or promising approaches might help? And are some of the proposed 'misinterpretation checklists' perhaps less informative than their authors would hope?

In this master's thesis, I show how rule-based approaches, combined with natural language processing (NLP) can be used to automatically detect, classify, and correct these misinterpretation. I show this for the misinterpretation of statistically nonsignificant results as the absence of an effect because it is the most extensively researched misinterpretation of *p* values (Lakens, 2021), and I have experience in classifying them from a previous project (Murphy et al., 2025). That said, the general framework I propose can be easily adapted to address other misinterpretations, also beyond *p* values.

### 1.1 Common Misinterpretations of *p* Values

The criticism of *p* values has become a prominent and recurring theme in discussions around scientific reform. From claims that they encourage dichotomous thinking (Amrhein et al., 2019; Hoekstra et al., 2006) to arguments that they offer little informational value (Wagenmakers, 2007), *p* values – and the broader framework of NHST – have been blamed for many of science's replication problems (McShane et al., 2019). On the other hand, many have

also argued that NHST per se is not to blame for these problems, but rather how researchers (mis)use and (mis)interpret this tool (Lakens, 2021). As a result, many researchers present whole collections of, in their view, common *p* value misinterpretations (see, e.g., Goodman, 2008; Greenland et al., 2016). Reviewing these, I come to the conclusion that there are four distinct types of misconceptions about *p* values that seem to be at play:

XXX

In my master thesis, I will zoom in on one specific misinterpretation: concluding *no effect* based on a statistically nonsignificant finding. Many studies have previously shown that this misinterpretation is and remains highly prevalent across time and sub-domains of psychology (e.g., Aczel et al., 2018; Hoekstra et al., 2006; Murphy et al., 2025). In fact, in a recently published article investigating articles published in 2009, 2015, and 2021 across ten different psychology journals (mainly in the field of personality and social psychology), we estimated the prevalence of this misinterpretation in articles' discussion sections to lie between 76.17% and 84.90% (Murphy et al., 2025). This study highlights that the situation seems not to have greatly improved despite many researchers exploring new analysis techniques (e.g., Lakens et al., 2018) and continuous calls to reflect on interpretations of nonsignificant results (e.g., McShane et al., 2019).

### *1.1.1 The Big-Four* p *value misinterpretations*

In preparation for this thesis project, I reviewed many of the previously reported misinterpretations of *p* values (e.g., Goodman, 2008; Greenland et al., 2016; chapter 1.7 from Lakens, 2024) and categorize them into four main groups:

- *p* values as hypothesis probabilities
- Blending statistical and practical significance
- *p* values as measures of replicability or error rates
- Technical misunderstandings about *p* values

I argue that many of these published 'misinterpretation checklists' largely reiterate similar underlying misconceptions, often merely rephrasing what is, at its core, the same fundamental issue. While I will elaborate on this reasoning in the final thesis more, for the purposes of this proposal, I focus on the first category.

This misinterpretation refers to the tendency of researchers to treat $p$ values as if they represented the probability that the null (or alternative) hypothesis is true. Researchers who follow this misinterpretation may interpret $p$ values below the conventional 5% threshold as evidence that $H_1$ is true (and $H_0$ is false), and nonsignificant $p$ values as evidence that $H_0$ is true (and $H_1$ is false). In this project, I specifically focus on the latter mistake: interpreting a nonsignificant result as proof that no effect exists. This interpretation cannot be justified within the standard NHST framework, which defines the $p$ value as the probability of observing the data, or something more extreme, assuming that the null hypothesis is true. There are, however, ways to overcome these misinterpretations, which I will discuss in the next section.

### 1.2 Research on nonsignificance as absence

XXX Aczel et al. (2018); Hoekstra et al. (2006); Murphy et al. (2025); and more

### 1.3 Possible solutions

This section will be more detailed in the final thesis, but I do want to briefly outline what I consider the most important solutions to the misinterpretation of nonsignificant results as evidence for the absence of an effect. One frequently suggested solution is to improve researchers' statistical literacy through enhanced education, such as better statistics teaching at the undergraduate and graduate levels (e.g., Lakens, 2021). However, as noted earlier, the prevalence of the misinterpretations I focus on does not seem to have substantially decreased, suggesting that calls for better education alone have not resolved the problem (Murphy et al., 2025).

A promising practical solution when conducting research involves the use of alternative analysis techniques, such as equivalence testing or minimum-effect tests. These methods allow researchers to test whether an effect is practically relevant and larger than a predefined smallest effect size of interest (SESOI) (Lakens et al., 2018). In many contexts, such approaches might be more closely aligned with the substantive questions researchers aim to answer, namely whether an effect is meaningful in practice.

### 1.4 An Automated Human-Factors Perspective

These strategies also align with the argument made by Lakens (2021) that $p$ value misinterpretations represent a human factors problem, requiring practical and

easy-to-implement solutions. Everyday examples of such solutions include cars with automatic braking systems, word processors that flag spelling and grammar mistakes, or email clients that filter out malware and phishing attempts. Analogously, and recognizing that new analytic approaches may not be adopted overnight, automated checks for statistical misinterpretations offer a highly promising route. This perspective emphasizes that many statistical errors arise not from bad intentions or ignorance, but from cognitive limitations and suboptimal workflows.

In the context of research, similar automated solutions are already gaining traction. For instance, the reference manager Zotero flags references to retracted papers (Stillman, 2019). Statcheck (Nuijten & Epskamp, 2024) automatically detects inconsistencies between reported test statistics and *p* values. Other tools, such as GRIM, GRIMMER, and SPRITE, identify impossible values in reported summary statistics (Heathers et al., 2018), while Regcheck (Cummin & Hussey, 2024) verifies the consistency between manuscripts and their preregistration documents.

To make the process of checking manuscripts more systematic, DeBruine and Lakens (2025) developed papercheck, an R package and Shiny app, which allows users to run a battery of checks on research papers. These include statistical checks (e.g., identifying imprecisely reported *p* values) as well as general manuscript quality checks (e.g., verifying links to online repositories or consistency between in-text citations and reference lists). Papercheck can be used both for single articles (e.g., as writing assistance) and for batches of articles (e.g., for meta-scientific studies). Because this framework is actively maintained and continues to evolve, I plan to build my thesis project within the papercheck infrastructure.

In summary, there are many reasons why *p* values remain difficult to interpret correctly. Empirical evidence suggests that misinterpretations of nonsignificant results remain highly prevalent (Murphy et al., 2025) This persistence highlights that improved education alone may not be sufficient. Drawing on a human factors perspective (Lakens, 2021), practical solutions such as automated error-checking tools offer a promising avenue for addressing these challenges. In this project, I aim to develop an automated approach to detect misinterpretations of nonsignificant results, building on the existing papercheck framework

(DeBruine & Lakens, 2025). In the following section, I outline the methods and approaches that I will explore to achieve this goal.

## 2 Methods

### 2.1 Statement Detection, Classification and Correction

Before describing the data used in this study, it is important to understand the three steps of the proposed framework. Statements from scientific articles needed to be reliably detected, classified, and finally corrected. For each step, I applied specific methods that were best suited to achieve the respective goal.

To detect statements I searched used rule-based regular expressions (RegEx) and searched articles's results sections for them. Effectively, RegEx searchers are just more complex Ctrl+F searches, where a user can also include optional charcters (e.g., 'significant(ly)' would catch both *significant* and *significantly*) and more complex rules (e.g., 'not.{0,20}significant' allows up to 20 characters between *not* and *significant*). Papercheck (DeBruine & Lakens, 2025) has a module that detects almost all p values (see Section XXX) based on RegEx searches and I filtered these to just the ones equal to or above .05. I then expanded the extracted nonsignificant p values to the full sentence with papercheck and added +/- one sentence as context in case of extraction errors etc. (more on this in Section XXX).[1]

In the next step, these statements (labeled as correct or incorrect by me; see Section XXX) were used to train several BERT-based models. BERT (Bidirectional Encoder Representations from Transformers) is a general-purpose language model pre-trained on the BookCorpus and English Wikipedia, making it suitable for a wide range of tasks – but not specifically optimized for scientific or technical language (Devlin et al., 2019). Since its introduction, many researchers have developed domain-specific variants of BERT to enhance its performance on specialized tasks. To test whether such domain adaptation improves performance in my classification task, I trained two models in addition to standard BERT: SciBERT was trained on a large corpus of scientific articles from Semantic Scholar, particularly in the biomedical and computer science domains (Beltagy et al., 2019).

[1] In a final tool, users will be able to set the alpha level they used themselves, thus allowing other levels than the conventional 5%.

PubMedBERT goes even further, having been trained exclusively on biomedical abstracts and full-text articles from the PubMed database (Gu et al., 2022). These models were evaluated on their ability to distinguish between correct and incorrect interpretations of nonsignificant results in scientific writing.

Lastly, statements that have been labeled as being incorrect by a BERT classifier were sent to a LLM to get corrected. The full prompt is available in X was had X word plus the respective statement. In short, the model was instructed to only change any misinterpretations of nonsignificant p values as the absence of an effect and keep the rest of the statement. Do communicate with the LLM, I used papercheck (DeBruine & Lakens, 2025), which, in turn, uses the Groq API (available at https://groq.com/) to communicate with different LLMs. I used papercheck's standard LLM, 'llama-3.3-70b-versatile' (as of 07/24/2025).

### 2.2 Validation Process and Performance Metrics

MAYBE (RE)MOVE The articles that were used in this study were part of papercheck's sample library of 250 open access article from the journal Psychological Science, published between 2013 and 2024 (Median = 2021; IQA = [2018; 2022]). MAYBE (RE)MOVE

To assess how well each of these three automated approaches worked, I compared each one to human ground truth and calculated appropriate measures of reliability between automated and human results.

Firstly, to ensure that the statement detection process actually caught all statements with nonsignificant p values, I manually extracted all of these from 25 (10%) of the papercheck sample library's 250 open access article from the journal Psychological Science. Articles were published between 2013 and 2024 (Median = 2021; IQA = [2018; 2022]). I then coded whether a statements I found were also extracted with the automated RegEx search.

For the training of the BERT models and to assess their final performance I labeled all automatically extracted statements that were detected to be from an article's Results section. This resulted in 960 statements in total. Of these, 420 were classified as correct and 352 were classified as incorrect. The remaining 188 statements were classified as not neither correct nor incorrect because they interpreted the nonsignificant effect as (marginally) significant (83),

because the statements were not complete enough to check their correctness (20), because they interpreted model fit indices and not the p value (20), because they were 'false flags' of nonsignificant p values (19), or due to a combination of these or other reasons (46).

Lastly, I also also went through the 100 statements that were sent to an LLM to be corrected, to see if the 'corrected' statements were actually correct. Of these 100 statements, 80 had previously been labeled incorrect and 20 correct by me, to check how the LLM deals with possible false positives from the automated classification.

In addition to these validity checks, there are also performance metrics specific to the trained classifiers. For training purposes, the labeled data was split into three parts: a test set (20%) used for the final evaluation of the best model; a training set (72%, or 90% of the remaining 80%) that the model uses to learn underlying patterns and adjust its parameters; and a validation set (8%, or 10% of the 80%) used to calculate evaluation metrics after each epoch (i.e., one full cycle of the model processing the training data).

During training, I computed the training loss (sum of errors between model predictions and actual labels in the training set) and the validation loss (same for validation set). The best-performing model was selected based on the lowest validation loss. For final evaluation, I computed the fraction of correctly predicted positive cases among all predicted positives (precision), the fraction of correctly predicted positive cases among all actual positives (recall), and their harmonic mean (F1 score), separately for each class label. To summarize overall performance across the two classes, I calculated the unweighted average of the two F1 scores (macro-F1 score).

### 2.4 Open Science

All scripts and data to reproduce and use the trained BERT models (Python), analyse the results and validity checks (mostly R) and recreate this manuscript (Quarto Markdown in R Studio; I did change few formatting, not content, related things manually in the exported Word document) are available in this GitHub repository, together with instructions on how to set it up: XXX.com.

This thesis was not preregistered as no inferential statistical tests were performed.

# 3 Results

## 3.1 Detection Accuracy

To see if my rule-based approach of using RegEx searches reliably detects all sentences with nonsignificant p values, I went through 25 random articles manually and copied each into a spreadsheet. I then compared how many of these were also caught by the automated approach. In total, I detected 179 statements with a nonsignificant p value. Of these, the automated RegEx searches got 130 (73 %) completely, and 6 partially due to extraction errors (e.g., because of pdf formatting like page breaks, figures or footnotes). In 3 cases the automated approach yielded false positives in the sense that it extracted 'statements' from tables or figure notes. Note, however, that the large majority of the total 49 missed statements were due to specific ways of writing (or not writing) the p value: 8 were missed because the authors wrote 'n.s.' instead of the nonsignificant p value, and 31 were missed because the p value was written as '$p_s$'. Without these two mistakes (since the p value was written in a strange way and this would be relatively easy to fix) the overall agreement of automated and manual approach would be 93 %.

Most of the other misses were due to pdf formatting issues like figures, tables, footnotes and page breaks or unusual characters inside the statement that messed with the statement extraction (11 in total).[2]

## 3.2 Classification Performance

The different performance metrics can be found in Table 1.

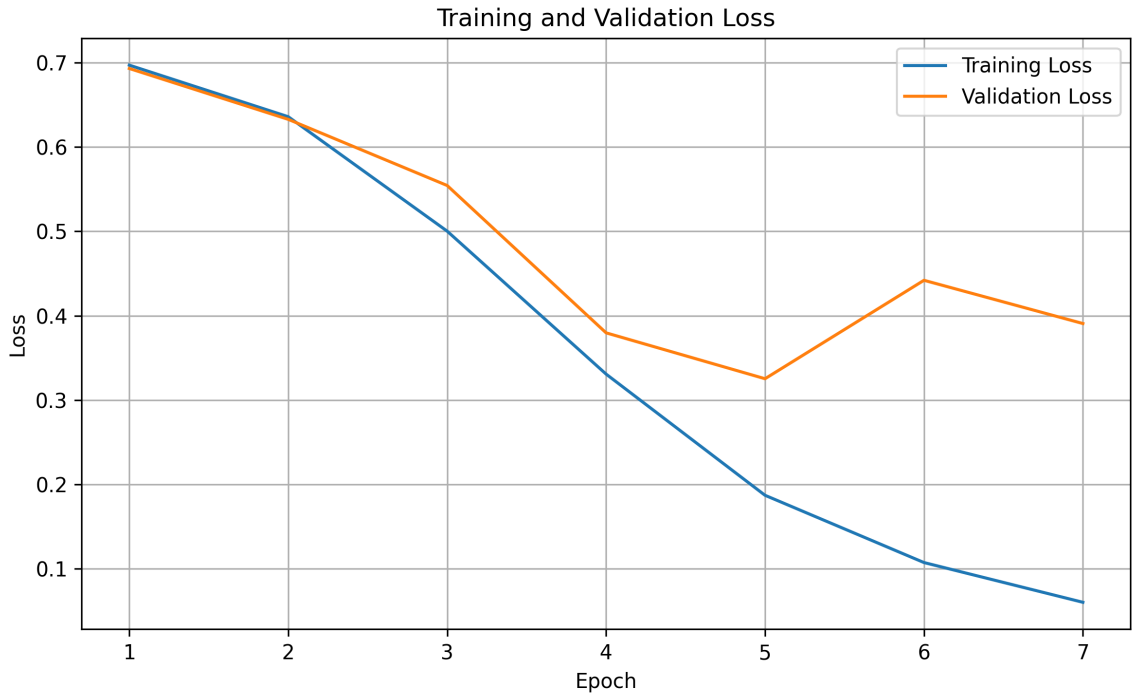
## 3.3 Correction Evaluation

A total of 100 statements from the classifier training data (since I had human labels for them and knew whether they were correct or incorrect) were sent to an LLM with papercheck using the groq API (XXX Insert model here XXX). 80 of these were statements that were classified as incorrect by me (see XXX section). An additional 20 correct statement were also included, to check how the LLM would deal with false positives from the automated

[2] I could not find one statement that was extracted automatically. My current theory is that this was an artifact from when the pdf was compiled and might be from a different article even, once again highlighting how impractical the pdf format is in times of increasing automation.

**Figure 1**

*Training and Validation Loss Curve*




*Note.* Curve of the training and validation loss across the all 7 epochs. The model after epoch 5 was chosen as the final model.

**Table 1**

*Model Performance*

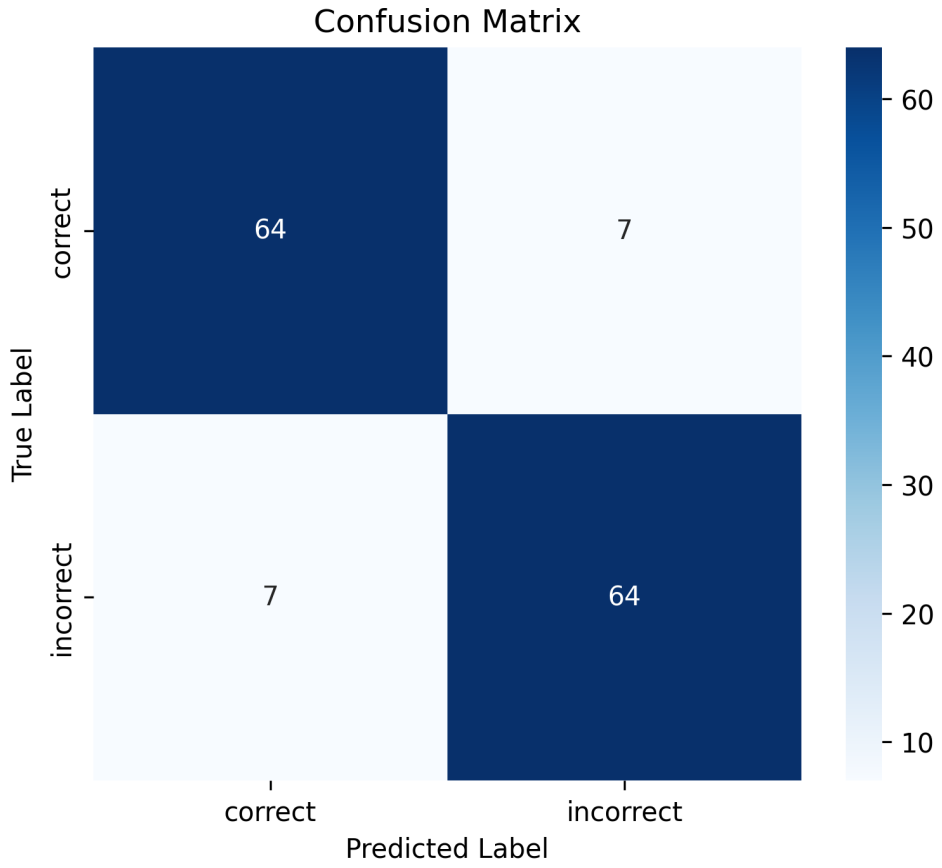
| Label | Precision | Recall | F1 Score |
|---|---|---|---|
| Correct | 0.90 | 0.90 | 0.90 |
| Incorrect | 0.90 | 0.90 | 0.90 |

*Note.* Table of precision, recall and F1 score per category.

classification. In total, 85 % of these were correct. Interestingly, 2 of the previously correct statements got turned incorrect by the LLM. 18, on the other hand, remained correct. Similarly, the LLM actually corrected 67 of the 80 previously incorrect statements, whereas 13 remained incorrect. Note, however, that the LLM was instructed to change as much as necessary, but as little as possible about the original statement. For some statements, this meant that they could not be corrected without major rephrasing. Examples for some really

**Figure 2**

*Confusion Matrix*




*Note.* Confusion matrix of the performance of the final model.

good/bad corrections can be found in Table 2.

## 4 Discussion

### 4.1 Summary of Findings

### 4.2 Strengths of the Approach

### 4.3 Limitations and Challenges

### 4.4 Implication of the Tool

### 4.5 Future Directions and Improvements

## 5 Conclusion

## References

Aczel, B., Palfi, B., Szollosi, A., Kovacs, M., Szaszi, B., Szecsi, P., Zrubka, M., Gronau, Q. F., Van Den Bergh, D., & Wagenmakers, E.-J. (2018). Quantifying Support for the Null Hypothesis in Psychology: An Empirical Investigation. *Advances in Methods and Practices in Psychological Science*, *1*(3), 357–366. https://doi.org/10.1177/2515245918773742

Amrhein, V., Greenland, S., & McShane, B. (2019). Scientists rise up against statistical significance. *Nature*, *567*(7748), 305–307. https://doi.org/10.1038/d41586-019-00857-9

Beltagy, I., Lo, K., & Cohan, A. (2019). *SciBERT: A Pretrained Language Model for Scientific Text*. https://doi.org/10.48550/ARXIV.1903.10676

Cummin, J., & Hussey, I. (2024). *RegCheck. Compare preregistrations with papers. Instantly.* Available at https://regcheck.app/.

DeBruine, L., & Lakens, D. (2025). *Papercheck: Check scientific papers for best practices.* R package version 0.0.0.9033, available at https://scienceverse.github.io/papercheck/.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, *70*(3), 193–242. https://doi.org/10.1037/h0044139

Goodman, S. (2008). A Dirty Dozen: Twelve P-Value Misconceptions. *Seminars in*

*Hematology*, *45*(3), 135–140. https://doi.org/10.1053/j.seminhematol.2008.04.003

Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, P values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, *31*(4), 337–350. https://doi.org/10.1007/s10654-016-0149-3

Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2022). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare*, *3*(1), 1–23. https://doi.org/10.1145/3458754

Heathers, J. A., Anaya, J., Van Der Zee, T., & Brown, N. J. (2018). *Recovering data from summary statistics: Sample Parameter Reconstruction via Iterative TEchniques (SPRITE)*. https://doi.org/10.7287/peerj.preprints.26968v1

Hoekstra, R., Finch, S., Kiers, H. A. L., & Johnson, A. (2006). Probability as certainty: Dichotomous thinking and the misuse of p values. *Psychonomic Bulletin & Review*, *13*(6), 1033–1037. https://doi.org/10.3758/BF03213921

Isager, P. M., & Fitzgerald, J. (2024). *Three-Sided Testing to Establish Practical Significance: A Tutorial*. https://doi.org/10.31234/osf.io/8y925

Lakens, D. (2021). The Practical Alternative to the *p* Value Is the Correctly Used *p* Value. *Perspectives on Psychological Science*, *16*(3), 639–648. https://doi.org/10.1177/1745691620958012

Lakens, D. (2024). *Improving Your Statistical Inferences*.

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., Baguley, T., Becker, R. B., Benning, S. D., Bradford, D. E., Buchanan, E. M., Caldwell, A. R., Van Calster, B., Carlsson, R., Chen, S.-C., Chung, B., Colling, L. J., Collins, G. S., Crook, Z., … Zwaan, R. A. (2018). Justify your alpha. *Nature Human Behaviour*, *2*(3), 168–171. https://doi.org/10.1038/s41562-018-0311-x

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon Statistical Significance. *The American Statistician*, *73*(sup1), 235–245. https://doi.org/10.1080/00031305.2018.1527253

Murphy, S. L., Merz, R., Reimann, L.-E., & Fernández, A. (2025). Nonsignificance misinterpreted as an effect's absence in psychology: Prevalence and temporal analyses. *Royal Society Open Science*, *12*(3), 242167. https://doi.org/10.1098/rsos.242167

Nuijten, M. B., & Epskamp, S. (2024). *Statcheck: Extract statistics from articles and recompute p-values. R package version 1.5.0.* Web implementation at https://statcheck.io.

Schervish, M. J. (1996). *P* Values: What They are and What They are Not. *The American Statistician*, *50*(3), 203–206. https://doi.org/10.1080/00031305.1996.10474380

Stillman, D. (2019). *Retracted item notifications with Retraction Watch integration*. https://www.zotero.org/blog/retracted-item-notifications/.

Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. https://doi.org/10.3758/BF03194105

**Table 2**

*Examples of LLM Corrections*

| Example | Statement Type | Statement | Label |
|---|---|---|---|
| | Original | Only in the twin sample did a lower age of onset predict higher levels of melody discrimination (β = -0.09, p = .23 in musicians; β = -0.05, p = .001 in twins). | 1 |
| Example 1 to 0 | LLM-Corrected | Only in the twin sample was there a significant association between a lower age of onset and higher levels of melody discrimination (β = -0.09, p = .23 in musicians; β = -0.05, p = .001 in twins), whereas in musicians, the association was not statistically significant. | 0 |
| | Original | However, when the target was uncommitted, participants no longer preferred the seeker, p = .849, d = -0.03, rating both targets equally. | 1 |
| Example 1 to 1 | LLM-Corrected | However, when the target was uncommitted, participants showed no significant preference for the seeker over the other target, p = .849, d = -0.03, rating both targets equally. | 1 |
| | Original | The years-of-experience single effect was not significant (β = 0.001, 95% CI = [-0.07, 0.07], p = .976). | 0 |
| Example 0 to 0 | LLM-Corrected | The years-of-experience single effect was not significant (β = 0.001, 95% CI = [-0.07, 0.07], p = .976), indicating that the data do not provide sufficient evidence to reject the null hypothesis of no effect. | 0 |
| | Original | By contrast, leader narcissism levels were not significantly related to follower-perceived leader authority, b = 0.04, 95% CI = [-0.04, 0.13], β = 0.10, p = .304, and effectiveness, b = 0.04, 95% CI = [-0.06, 0.14], β = 0.08, p = .398. | 0 |

# Appendix

XXX