Proposal: The *p* stands for Proofread: Automated Detection and Correction of Nonsignificance Misinterpretations

Raphael Merz

Department of Psychology, Ruhr University Bochum

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

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?

This proposal aims to develop an automated framework to detect, classify, and correct misinterpretations of *p* values in research papers using a mix of rule-, large language model- (LLM), and classifier-based approaches. I focus on 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 these interpretations from a previous project (Murphy et al., 2025). That said, I aim to address the bigger picture of *p* value misinterpretations, going beyond nonsignificance as the absence of an effect, in more detail in the final thesis, and will discuss how the analysis pipeline could be adapted to detect other types of misinterpretations.

1.1 *p* values criticism

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).

In my master's 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 a 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 more in the final thesis, 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 that 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 Overcoming *p* value misinterpretations

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.

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 Identifying sub-steps

The main components of this thesis are summarized in Figure 1. I conceptualize the project as progressing through three key steps to achieve the goal of full automation.

Firstly, statements about nonsignificance – whether they are interpreted correctly or not – must be reliably detected. That is, the chosen method should ideally identify all sentences that refer to nonsignificant results while minimizing false positives (i.e., sentences that do not actually interpret nonsignificance). From which sections of a manuscript these statements should be identified remains an open question. To keep the project scope manageable, the initial focus will be on statements that include an explicitly nonsignificant *p* value (for now, $p > .05$). If this approach proves successful, I plan to expand the scope to include related contextual statements.

A more ambitious, potentially out-of-scope of this thesis, extension would be to incorporate interpretations in discussion sections that do not directly report *p* values but still convey conclusions about nonsignificance.

Secondly, these statements must be classified as either correct (i.e., not suggesting the absence of an effect) or incorrect (i.e., suggesting the absence of an effect). Aczel et al. (2018) further categorized incorrect statements as either sample-level or population-level misinterpretations. For example, a sample-level misinterpretation might state that "groups were the same," while a population-level one might claim that "men and women are the same".

Figure 1

Project Components

Detection	
Sentences with p > .05	
Sentences related to specific nonsignificant effects in results section	
Interpretations of nonsignificant effects in discussion section	
Classification	
Single sentences with p > .05 as correct/incorrect	
Single sentences with p > .05 as correct/incorrect–sample/incorrect–population	
Interpretations of nonsignificant effects in the discussion section as correct/incorrect	
Interpretations of nonsignificant effects in the discussion section as correct/incorrect–sample/incorrect–population	
Implementation	
Feedback and writing assistance…	…on single statements with p > .05
	…on related statements in results section
	…on interpretations in discussion section
Apply tool to larger set of articles	Apply tool to estimate the prevalence of correct/incorrect statements over journals and time
Platform to use tool	Implement tool into a papercheck module that can be used through the papercheck Shiny app

Note. The colors indicate how likely it is that the respective component will be included in the final thesis: green – definitely included; yellow – likely included; orange – possibly included; red – unlikely to be included.

Although this distinction is theoretically meaningful, for the purposes of detection and correction both types are problematic and can be treated similarly in this context. However, the distinction may still offer valuable insights in meta-scientific contexts and will be explored as an optional classification system if time permits.

Lastly, there are three ways in which I plan to implement this tool: (A) as writing assistance, where users will receive suggested alternatives to problematic phrasings that avoid implying the absence of an effect. Initially, this will focus on sentences containing a nonsignificant *p* value, but later stages may also target surrounding contextual statements. (B) If the tool's performance is sufficient, it will be applied to a large collection of articles to estimate the prevalence of the misinterpretation across journals and time. (C) The tool will be added to the papercheck module list, allowing users to run automated checks via the existing Shiny app interface (see science-verse.github.io/papercheck/reference/papercheck_app.html; DeBruine & Lakens, 2025).

2.2 Approaches to automatic checks

In this proposal, I focus on three primary technical approaches to automatic checking of *p* value interpretations and text classification more generally. Rule-based text searches will primarily detect candidate statements (e.g., those explicitly stating nonsignificant *p* values). LLM-based approaches will help determine whether these statements truly reflect nonsignificance interpretations, identify contextually related statements without *p* values, and generate improved phrasing for incorrect interpretations. Finally, classifier-based models will categorize statements as correct or incorrect. These three approaches are complementary and will be integrated across the detection, classification, and correction steps.

2.2.1 Rule-based text searches

Rule-based methods, such as regular expressions (RegEx), allow for targeted text searches based on specific patterns. The papercheck framework already supports RegEx-based screening. RegEx is particularly useful because it accommodates flexible matching, such as optional characters (e.g., to get "wellbeing" and "well-being") or intervening words (e.g., capturing "did not […] predict").

In early project stages, I will use rule-based approaches to identify a broad set of candidate statements that might interpret nonsignificant results (e.g., because they contain a *p*

value above the 5% threshold). Because RegExes are inherently limited in semantic understanding, this approach is expected to yield many false positives but will help ensure that potential misinterpretations are not overlooked.

2.2.2 LLM-based approaches

Next, I will use LLMs, which are already integrated into papercheck, to refine the initial statement detection. For example, an LLM can assess whether sentences containing the term "well-being" refer to psychological well-being. Similarly, I will use LLMs to evaluate whether flagged sentences genuinely interpret nonsignificant findings.

In preliminary tests I conducted using data from Murphy et al. (2025), LLMs showed relatively low agreement with human coders when directly classifying statements as correct or incorrect. However, they may still play a valuable role in other stages of the process. Specifically, LLMs could help identify semantically related statements that interpret nonsignificant results (e.g., in an article's discussion section) but do not contain an explicit *p* value – an important step toward expanding beyond simple rule-based detection. Additionally, LLMs will be central to the writing assistance component of this project, where they can generate context-sensitive corrections and improved phrasings for misinterpreted statements.

2.2.3 Classifier-based models

Finally, I will explore machine learning classifiers that categorize statements based on learned patterns. Compared to generative LLMs like ChatGPT or LLaMA, fine-tuned transformer-based classifiers like BERT (Devlin et al., 2019) have often demonstrated superior performance in knowledge-intensive tasks such as scientific text classification (e.g., Bucher & Martini, 2024).

To evaluate the best-performing model, I will compare three transformer-based classifiers that vary in their training data and domain specialization. The original BERT model is a general-purpose language model pre-trained on BooksCorpus and English Wikipedia, making it suitable for a wide range of tasks but not optimized for scientific or technical language (Devlin et al., 2019). SciBERT, by contrast, was specifically trained on a large corpus of scientific articles sourced from Semantic Scholar, with a focus on biomedical and computer science domains (Beltagy et al., 2019). PubMedBERT goes further in domain specialization, having been trained

exclusively on abstracts and full-texts of biomedical articles from the PubMed database (Gu et al., 2022). These models will be tested on their ability to distinguish correct and incorrect interpretations of nonsignificant results in scientific writing.

2.3 Performance metrics and human coding

Each stage of the tool's pipeline—detection, classification, and user feedback—will be evaluated against human-coded 'ground truth' data to ensure the system functions as intended.

To begin, I will manually code a set of open access Psychological Science articles that are included in the papercheck testing dataset (provided for users to explore key functions). The final number of articles will depend on how time-intensive the manual coding proves to be, which I will assess in early piloting. These human-coded statements will serve several purposes throughout the project.

Detection. Firstly, I will evaluate whether the rule-based search approach reliably detects all relevant statements that interpret nonsignificant findings. To do this, I will manually review a selection of full-text papers and identify any relevant statements that were missed by the automated search. These insights will help refine the detection rules and improve the tool's ability to capture a broad range of relevant phrasing patterns.

Classification. Secondly, the manually coded statements will be used to train and evaluate the BERT-based classifiers. For each model, I will report the confusion matrix values (true/false positives and negatives) and calculate the following standard performance metrics:

- *Precision* measures the proportion of true positives out of all positive predictions, reflecting how many identified nonsignificant findings were correct.
- *Recall* measures the proportion of true positives out of all actual positives, indicating how well the model identifies all relevant cases.
- The *F1 score* is the harmonic mean of precision and recall, balancing both into a single metric, which is especially useful when there is an imbalance between the number of positive and negative cases.

User feedback. Finally, I will assess the quality of LLM-generated suggestions for revised, statistically accurate phrasing. A subset of these corrections will be evaluated against human judgment to verify if they provide genuinely helpful and correct feedback. This step ensures that the final tool supports better writing without introducing new errors.

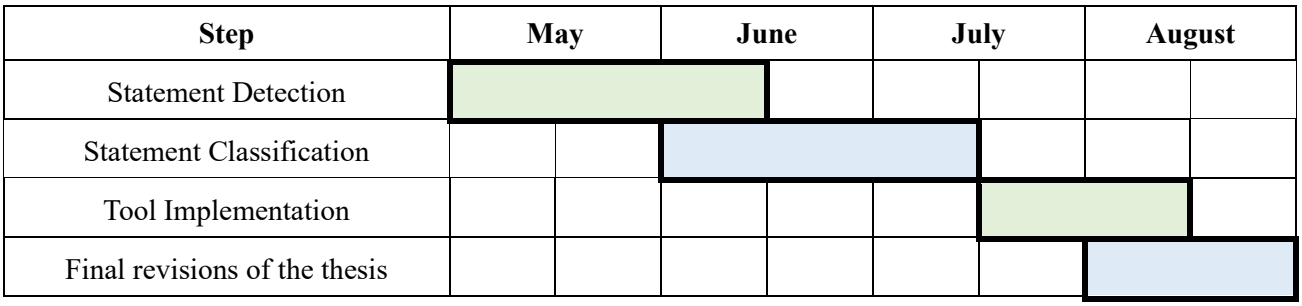
Following this validation and if the model achieves sufficient performance, the tool will be applied to a larger set of articles for analyses (see Figure 1), similar to the approach in Murphy et al. (2025). This potential phase will, however, be preregistered separately with its respective hypotheses and planned analyses.

2.4 Timeline

A tentative timeline for the development and evaluation of the tool is shown in Figure 2. As mentioned earlier, the project follows an iterative structure. I will begin by implementing the core functionality (indicated in green in Figure 1), and based on its performance, aim to extend it to the more advanced components (yellow to red in Figure 1). The final deadline for submitting the thesis is August 31, 2025.

Figure 2

Thesis Timeline



Note. Timeline of the project's components (as outlined in Figure 1), indicating when each will be conducted between May and August 2025.

3 Preliminary Results and Final Remarks

While many aspects of the project are still in development, early results from the rule-based text searches are highly encouraging. Although these searches currently focus on relatively simple terms and patterns, identifying these statements in a manuscript already provides significant value to authors writing scientific texts. The inclusion of classifiers and suggested alternative phrasings will further enhance the tool's functionality, offering users even greater support.

Regardless of the final scope, I am confident that this project will yield a robust and effective prototype. While further improvements may need to be made, the tool will be a

valuable contribution to improving the accuracy and clarity of scientific writing. Even if its performance might not be flawless, I am optimistic that it will play an important role in fostering more transparent, accurate, and effective research communication.

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

Bucher, M. J. J., & Martini, M. (2024). *Fine-Tuned 'Small' LLMs (Still) Significantly Outperform Zero-Shot Generative AI Models in Text Classification* (arXiv:2406.08660). arXiv. https://doi.org/10.48550/arXiv.2406.08660

Cummin, J., & Hussey, I. (2024). *RegCheck. Compare preregistrations with papers. Instantly.* Online app 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*. Online textbook available at https://lakens.github.io/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*. Blog post available at 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