A Systematic Review of Open Science Practices in the Studies of Crime

Research Proposal

Michael Beck

2025-01-08

Master's thesis
Supervisor:
Dr. Alexander Trinidad

Contents

List of Figures

List of Tables

Intro & Motivation

Modern Science

The rise of the internet in the last decades drastically changed our lives: Our ways of looking at the world, our social lives or our consumption patterns - the internet influences all spheres of life, whether we like it or not (*Society and the Internet*, 2019). The surge interconnectivity enabled a rise in movements that resist the classic definition of intellectual property rights: open source, open scholarship access and open science (Willinsky, 2005). Modern technologies enhanced reliability, speed and efficiency in knowledge development, thereby enhancing communication, collaboration and access to information or data (Eisend, 2002; Thagard, 1997; Warden, 2010). The internet significantly facilitated formal and informal scholarly communication through electronic journals and digital repositories like Academia.edu or ResearchGate (Waite, 2021; Warden, 2010). Evidence also shows that an increase in access to the internet also increases research output (Xu & Reed, 2021). But greater output doesn't necessarily imply greater quality, progress or greater scientific discoveries. As availability and thereby the quantity of publications increased, the possible information overload demands for effective filtering and assessment of publicated results (Warden, 2010).

But how do we define scientific progress? In the mid 20st century, Thomas Kuhn characterized scientific progress as a revolutionary shift in paradigms that are accepted theories in a scientific community at a given time. According to Kuhn, normal science operates within these paradigms, "solving puzzles" and refining theories. However, when anomalies arise that cannot be explained by the current paradigm, a crisis occurs, leading to a scientific revolution (Kuhn, 1962, 1970). Opposed to that, a critical rationalist approach to scientific progress emerged that saw danger in the by Kuhn described process, as paradigms might facilitate confirmation bias and thereby stall progress. Karl Popper's philosophy of science, which emphasizes falsifiability and the idea that scientific theories progress through conjectures and refutations rather than through paradigm shifts. Popper argued that science advances by eliminating false theories, thus moving closer to the truth in a more linear and cumulative manner (Popper, 2005). Where Kuhn emphasized the development and refinement of dominant theories, Popper suggested the challenging or falsification of those theories.

Social sciences today engage in frequentist, deductive reasoning where significance testing is used to evaluate the null hypothesis, and conclusions are drawn based on the rejection or acceptance of this hypothesis, aligning with Popper's idea that scientific theories should be open to refutation. This approach is often criticized for its limitations in interpreting p-values and its reliance on long-run frequency interpretations (Dunleavy & Lacasse, 2021; Wilkinson, 2013). In contrast, Bayesian inference is associated with inductive reasoning, where models are updated with new data to improve predictions. Bayesian methods allow for the comparison of competing models using tools like Bayes factors, but they do not directly falsify models through significance tests (Doll & Jacquemin, 2019; Gelman, 2011). Overall, while falsification remains a cornerstone of scientific methodology, contemporary science often employs a pluralistic approach, integrating various methods to address complex questions and advance knowledge (Rowbottom, 2011). This pluralistic approach in contemporary science underscores the importance of integrating diverse methodologies to solve complex questions and enhance our understanding. Despite the differences between frequentist and Bayesian methods, both share a fundamental commitment to the rigorous testing and validation of scientific theories. But besides the more theoretically driven discourse about scientific discovery, there are many tangible reasons to talk about the scientific method and the publication process. A recent, highly cited article revealed that only a very small proportion in variance of the outcomes in studies based on the same data can be accounted to the choices made by researchers in designing their tests. Breznau et al. (2022) observed 77 researcher teams analyzing the same dataset to assess the same hypothesis and found that the results were extremely divers, ranging from strong positive to strong negative results. Between-team deviance could only be explained to less than 50% by assigned conditions, research decisions and researcher characteristics, the rest of the variance remained unexplained. This underlines the importance of transparent research: results are prone to many errors and biases, made intentionally or unintentionally by the researcher or induced by the publisher.

> "Only by ... repetitions can we convince ourselves that we are not dealing with a mere isolated 'coincidence', but with events which, on account of their regularity and reproducibility,

are in principle inter-subjectively testable." (Popper, 2005, p. 23)

To challenge the biases and to support the possibility of these "repetitions" or replications of research, a movement has formed within the scientific community, fuelled by the "replication crisis" that was especially prevalent within the field of psychology (Dienlin et al., 2021). The open science movement tries to establish open science practices to challenge many of the known biases that endanger the reliability of the scientific process and enable access to the scientific discourse for a broader public.

Banks et al. (2019) establish a definition of open science as a broad term that refers to many concepts including scientific philosophies embodying communality and universalism, specific practices operationalizing these norms including open science policies like sharing of data and analytic files, redefinition of confidence thresholds, pre-registration of studies and analytical plans, engagement in replication studies, removal of pay-walls, incentive systems to encourage the above practices and even specific citation standards. This typology is in line with the work of many other authors from diverse disciplines (e.g. Dienlin et al., 2021; and Greenspan et al., 2024). The ongoing debate of the last decades were especially focused on two open science practices.

First, the **publishing of materials, data and code** or *open data* that enables replication of studies. Replication thereby makes it possible to assess the pursued research in detail, find errors, bias or simply support the results of the replicated work (Dienlin et al., 2021). While many researchers see challenges in the publication of their data and materials due to a potentially higher workload, legal concerns or just lack of interest, many of these concerns could be ruled out by streamlined processes or institutional support (Freese, 2007; Freese et al., 2022). As open data reduces p-hacking, facilitates new research by enabling reproduction, reveals mistakes in the analytical code and enables a diffusion of knowledge on the research process, it seems that many scientists, journals and other institutions start to adopt open data in their research to an increasing extent (Dienlin et al., 2021; Fink & Marcus, n.d.; Freese et al., 2022; Mattern et al., 2024; Zenk-Möltgen et al., 2018).

Second, **preregistration** involves thoroughly outlining and documenting research plans and their rationale in a repository. These plans can be made publicly accessible when the researcher decides to share them. The specifics of preregistration can vary based on the research type and may encompass elements such as hypotheses, sampling strategies, interview guides, exclusion criteria, study design, and analysis plans (Manago, 2023). Within this definition, a preregistration shall not prevent exploratory research. Deviations from the research plan are still allowed but have to be communicated transparently (Manago, 2023; Nosek & Lakens, 2014). Preregistration impacts research in multiple ways: it helps performing exploratory and confirmatory research independently, protects against publication bias as journals typically commit to publish registered research and counters "researchers' degrees of freedom" in data analysis by reducing overfitting through cherry-picking, variable swapping, flexible model selection and subsampling (*False-Positive Psychology*, n.d.; Mertens & Krypotos, 2019). This minimizes the risk of bias by promoting decision-making that is independent of outcomes. It also enhances transparency, allowing others to evaluate the potential for bias and adjust their confidence in the research findings accordingly (Hardwicke & Wagenmakers, 2023).

My initial plan for my master's thesis was to study the effect of open science practices on reported effect sizes in published papers. During my initial literature review, it appeared to me that there were very few publications that used pre-registration in data-driven Criminology and Legal Psychology. Instead of assessing effect sizes, this raised the question how open science practices have been adapted within criminology. I therefore intend, motivated by the expected positive impact of open science practices and in line with the research of Scoggins & Robertson (2024), to assess the two research questions in Criminology and Legal Psychology:

> RQ_1: What proportion of papers that rely on statistical inference make their data and code public?

> RQ_2: What proportion of experimental studies were preregistered?

Scoggins & Robertson (2024) did an extensive analysis of nearly 100,000 publications in political science and international relations. They observed an increasing use of preregistration and open data, with levels still being relatively low. The extensive research not only revealed the current state of open science in political science, but also generated rich data to perform further meta research.

I intend to apply similar methods in the field of Criminology and Legal Psychology: gather data about papers in a subset of Criminology and Legal Psychology journals, classify those papers by application of open source practices using sophisticated machine learning methods and explore the patterns over time to take stock of research practices in the disciplines. In the following section I describe the intended data collection and research methods that are highly based on Scoggins & Robertson (2024) research.

Data and Method

The study will focus on papers in criminal psychology that use data and statistical methods. The aim is to evaluate the prevalence of key open science practices, including open access, pre-registration and open data. The research process will follow three steps: collection, classification and analysis. In line with preregistration guidelines, the outlined research plan may be subject to reconsideration during the research process that will be reported transparently (Manago, 2023; Nosek & Lakens, 2014).

Data Collection

The process of data collection will closely follow Scoggins & Robertson (2024) and begin with identifying relevant journals in criminal psychology. I will consult the Clarivate Journal Citation Report to obtain a comprehensive list of journals within the fields by filtering for the top 100 journals. The Transparency-and-Openness-Promotion-Factor[1] (TOP-Factor) according to Nosek et al. (2015) will be used to then assess the journal's admission of open science practices and by including it in the journal dataset. Once the relevant journals are identified, I will use APIs such as Crossref, Scopus, and Web of Science to download metadata for all papers published between 2013 to 2023.

After obtaining the metadata, I will proceed to download the full-text versions of the identified papers. Whenever possible, I will prioritize downloading HTML versions of the papers due to their structured format, which simplifies subsequent text extraction. For papers that are not available in HTML, I will consider downloading full-text PDFs. Tools such as PyPaperBot or others[2] can facilitate this process, although I will strictly stick to ethical and legal guidelines, avoiding unauthorized sources like Sci-Hub or Anna's Archive and only using sources that are either included in my institutions campus license or available via open access. If access to full-text papers becomes a limiting factor, I will assess alternative strategies such as collaborating with institutional libraries to request specific papers or identifying open-access repositories that may provide supplementary resources. Non-available texts will be considered with their own category in the later analysis. Once all available full-text papers are collected, I will preprocess the data by converting HTML and PDF files into plain text format using tools such as SciPDF Parser or others[3]. This preprocessing step ensures that the text is in a standardized format suitable for analysis.

The proposed data collection is resource-intensive but serves multiple purposes. However, resource constraints could pose challenges, such as limited access to computational tools, DDoS-protection[4], API-rate limits or delays in obtaining full-text papers. To mitigate these risks, I plan to prioritize scalable data collection methods, limit data collection to a manageable extent and use existing institutional resources, including library services and open-access repositories. Additionally, I will implement efficient preprocessing workflows ensuring that the project remains feasible within the given timeline and resources.

Classification

The classification process will begin with operationalizing the key open science practices that I aim to study. This involves the definition of clear criteria for identifying papers that fall into the categories I plan to classify: Papers that use statistical inference, papers that applied preregistration, papers that applied open data practices, papers that offer open materials and papers that are available via open access.

[1] The TOP-Factor according to Nosek & Lakens (2014) is a score that assesses the admission of open science practices can be obtained from topfactor.org.

[2] ferru97/PyPaperBot, monk1337/resp

[3] GitHub - titipata/scipdf_parser, GitHub - aaronsw/html2text, html2text · PyPI, GitHub - jsvine/pdfplumber, GitHub - cat-lemonade/PDFDataExtractor, GitHub - euske/pdfminer

[4] DDoS: Distributed Denial of Service, see Wang et al. (2015).

Classification of open access papers will be performed using the available metadata. The other classes will be identified using machine learning models trained on a preclassified training dataset. The models will categorize papers using generated document feature matrices (DFM's) in line with Scoggins & Robertson (2024).

To train machine learning models capable of classifying the papers, I will manually categorize a subset of papers. The sample size will be determined using weighted fitting of learning curves according to Figueroa et al. (2012) which need an initial hand-coded sample size of 100-200. To ensure the representativeness of this subset, I will sample papers proportionally from different journals, publication years, and subfields within Criminology and Legal Psychology. The stratified sampling approach will help mitigate biases and ensure that the training data reflects the diversity of the overall dataset. If the necessary sample size exceeds my time constraints, I will try to use clustering based text classification to extend the training sample (Zeng et al., 2003) and will also consider the use of large language models like ScienceOS for the generation of the training data by using such a model to preclassify papers.

The sampled subset will serve as a "labeled" dataset for supervised learning. Different classification methods were considered but deemed as not suitable for the task as those were either found to be designed for document topic classification or too time intense for a master's thesis (Jandot et al., 2016; e.g. Kim & Gil, 2019; Sanguansat, 2012).

Instead, I will follow the approach of Scoggins & Robertson (2024), using document feature matrices (DFMs) created from open science specific dictionaries. For instance, the frequencies of terms like "pre-registered," "open data," or "data availability statement" could indicate adherence to pre-registration or open data practices. Similarly, phrases such as "materials available on request" or "open materials" could signify the use of open materials. Scoggins & Robertson (2024) freely available data will form the foundation of keyword dictionaries for identifying relevant papers during the classification phase. Using these dictionaries, DFM's will be generated for all full-text papers gathered. To facilitate this, I will additionally develop own keyword dictionaries for each category, identifying terms and phrases commonly associated with these practices before consulting Scoggins & Robertson (2024).

I will then train various machine learning models, including Naive Bayes, Logistic Regression, Support Vector Machines, and Gradient Boosted Trees. The performance of each model will be evaluated to identify the best-performing classifier for each category of open science practices. Once the optimal models are selected, I will use them to classify the entire dataset of papers.

The automated classification will enable me to categorize a large amount papers automatically based on their adoption of open science practices. Automating the classification process mitigates the inefficiency of manual data collection, allowing for the analysis of a significantly larger dataset than would otherwise be feasible. This classification will provide the foundation for subsequent analyses of temporal trends and other patterns within the data.

Analysis

In the analysis phase of the research, an exploratory analysis will be conducted to explore temporal trends in the adoption of open science practices over the past decade. This involves comparing the adoption rates of practices such as pre-registration, open data, open materials, and open access across the disciplines of Criminology and Legal Psychology, as well as among different journals. The goal is to identify possible differences or similarities in how these practices have been embraced over time. This evaluation aims to uncover insights into the methodological rigor and transparency within the fields, providing a comprehensive understanding of the current landscape and potential areas for improvement in research practices. By building on the methods developed by Scoggins & Robertson (2024), I hope to generate data and insights that will support future efforts to promote transparency and reproducibility in criminal psychology.

References

Banks, G. C., Field, J. G., Oswald, F. L., O'Boyle, E. H., Landis, R. S., Rupp, D. E., & Rogelberg, S. G. (2019). Answers to 18 Questions About Open Science Practices. *Journal of Business and Psychology*, *34*(3), 257–270. https://doi.org/10.1007/s10869-018-9547-8

Breznau, N., Rinke, E. M., Wuttke, A., Nguyen, H. H. V., Adem, M., Adriaans, J., Alvarez-Benjumea, A., Andersen, H. K., Auer, D., Azevedo, F., Bahnsen, O., Balzer, D., Bauer, G., Bauer, P. C., Baumann, M., Baute, S., Benoit, V., Bernauer, J., Berning, C., … Żółtak, T. (2022). Observing many researchers using the same data and hypothesis reveals a hidden universe of uncertainty. *Proceedings of the National Academy of Sciences*, *119*(44), e2203150119. https://doi.org/10.1073/pnas.2203150119

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B. K., Huskey, R., Schneider, F. M., Breuer, J., Parry, D. A., Vermeulen, I., Fisher, J. T., Banks, J., Weber, R., Ellis, D. A., … de Vreese, C. (2021). An Agenda for Open Science in Communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

Doll, J. C., & Jacquemin, S. J. (2019). Bayesian Model Selection in Fisheries Management and Ecology. *Journal of Fish and Wildlife Management*, *10*(2), 691–707. https://doi.org/10.3996/042019-JFWM-024

Dunleavy, D. J., & Lacasse, J. R. (2021). *The Use and Misuse of Classical Statistics: A Primer for Social Workers*. https://journals.sagepub.com/doi/10.1177/10497315211008247.

Eisend, M. (2002). The Internet as a new medium for the sciences? The effects of Internet use on traditional scientific communication media among social scientists in Germany. *Online Information Review*, *26*(5), 307–317. https://doi.org/10.1108/14684520210447877

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant - Joseph P. Simmons, Leif D. Nelson, Uri Simonsohn, 2011. (n.d.). https://journals.sagepub.com/doi/10.1177/0956797611417632.

Figueroa, R. L., Zeng-Treitler, Q., Kandula, S., & Ngo, L. H. (2012). Predicting sample size required for classification performance. *BMC Medical Informatics and Decision Making*, *12*(1), 1–10. https://doi.org/10.1186/1472-6947-12-8

Fink, L., & Marcus, J. (n.d.). Replication code availability over time and across fields: Evidence from the German Socio-Economic Panel. *Economic Inquiry*, *n/a*(n/a). https://doi.org/10.1111/ecin.13267

Freese, J. (2007). Replication Standards for Quantitative Social Science: Why Not Sociology? *Sociological Methods & Research*, *36*(2), 153–172. https://doi.org/10.1177/0049124107306659

Freese, J., Rauf, T., & Voelkel, J. G. (2022). Advances in transparency and reproducibility in the social sciences. *Social Science Research*, *107*, 102770. https://doi.org/10.1016/j.ssresearch.2022.102770

Gelman, A. (2011). Induction and Deduction in Baysian Data Analysis. *Rationality, Markets and Morals*.

Greenspan, R. L., Baggett, L., & B. Boutwell, B. (2024). Open science practices in criminology and criminal justice journals. *Journal of Experimental Criminology*. https://doi.org/10.1007/s11292-024-09640-x

Hardwicke, T. E., & Wagenmakers, E.-J. (2023). Reducing bias, increasing transparency and calibrating confidence with preregistration. *Nature Human Behaviour*, *7*(1), 15–26. https://doi.org/10.1038/s41562-022-01497-2

Jandot, C., Simard, P. Y., Chickering, D. M., Grangier, D., & Suh, J. (2016). Interactive Semantic Featuring for Text Classification. *ArXiv*.

Kim, S.-W., & Gil, J.-M. (2019). Research paper classification systems based on TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, *9*(1), 30. https://doi.org/10.1186/s13673-019-0192-7

Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago.

Kuhn, T. S. (1970). Reflections on my Critics. In A. Musgrave & I. Lakatos (Eds.), *Criticism and the Growth of Knowledge: Proceedings of the International Colloquium in the Philosophy of Science, London, 1965* (Vol. 4, pp. 231–278). Cambridge University Press. https://doi.org/10.1017/CBO9781139171434.011

Manago, B. (2023). Preregistration and Registered Reports in Sociology: Strengths, Weaknesses, and Other Considerations. *The American Sociologist*, *54*(1), 193–210. https://doi.org/10.1007/s12108-023-09563-6

Mattern, J. B., Kohlburn, J., & Moulaison-Sandy, H. (2024). Why academics under-share research data: A social relational theory. *Journal of the Association for Information Science and Technology*, *75*(9), 988–1001. https://doi.org/10.1002/asi.24938

Mertens, G., & Krypotos, A.-M. (2019). Preregistration of Analyses of Preexisting Data. *Psychologica*

Belgica, *59*(1). https://doi.org/10.5334/pb.493

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., ... Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374

Nosek, B. A., & Lakens, D. (2014). Registered Reports. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Popper, K. (2005). *The Logic of Scientific Discovery* (2nd ed.). Routledge. https://doi.org/10.4324/9780203994627

Rowbottom, D. P. (2011). Kuhn vs. Popper on criticism and dogmatism in science: A resolution at the group level. *Studies in History and Philosophy of Science Part A*, *42*(1), 117–124. https://doi.org/10.1016/j.shpsa.2010.11.031

Sanguansat, P. (2012). Feature matricization for document classification. *2012 IEEE International Conference on Signal Processing, Communication and Computing (ICSPCC 2012)*, 745–749. https://doi.org/10.1109/ICSPCC.2012.6335622

Scoggins, B., & Robertson, M. P. (2024). Measuring transparency in the social sciences: Political science and international relations. *Royal Society Open Science*, *11*(7), 240313. https://doi.org/10.1098/rsos.240313

Society and the Internet: How Networks of Information and Communication are Changing Our Lives. (2019). [Computer software]. Oxford University Press. https://doi.org/10.1093/oso/9780198843498.001.0001

Thagard, P. (1997). *Internet Epistemology: Contributions of New Information Technologies to Scientific Research*.

Waite, V. (2021). INTERNET KNOWLEDGE EXCHANGE AND CO-AUTHORSHIP AS FACILITATORS IN SCIENTIFIC RESEARCH. *Journal of Teaching English for Specific and Academic Purposes*, *0*, 043–050. https://doi.org/10.22190/JTESAP2101043W

Wang, B., Zheng, Y., Lou, W., & Hou, Y. T. (2015). DDoS attack protection in the era of cloud computing and Software-Defined Networking. *Computer Networks*, *81*, 308–319. https://doi.org/10.1016/j.comnet.2015.02.026

Warden, R. (2010). The Internet and science communication: Blurring the boundaries. *Ecancermedicalscience*, *4*, 203. https://doi.org/10.3332/ecancer.2010.203

Wilkinson, M. (2013). Testing the null hypothesis: The forgotten legacy of Karl Popper? *Journal of Sports Sciences*, *31*(9), 919–920. https://doi.org/10.1080/02640414.2012.753636

Willinsky, J. (2005). The unacknowledged convergence of open source, open access, and open science. *First Monday*. https://doi.org/10.5210/fm.v10i8.1265

Xu, X., & Reed, M. (2021). The impact of internet access on research output - a cross-country study. *Information Economics and Policy*, *56*, 100914. https://doi.org/10.1016/j.infoecopol.2021.100914

Zeng, H.-J., Wang, X.-H., Chen, Z., Lu, H., & Ma, W.-Y. (2003). CBC: Clustering based text classification requiring minimal labeled data. *Third IEEE International Conference on Data Mining*, 443–450. https://doi.org/10.1109/ICDM.2003.1250951

Zenk-Möltgen, W., Akdeniz, E., Katsanidou, A., Naßhoven, V., & Balaban, E. (2018). Factors influencing the data sharing behavior of researchers in sociology and political science. *Journal of Documentation*, *74*(5), 1053–1073. https://doi.org/10.1108/JD-09-2017-0126