cogitatio

MEDIA AND COMMUNICATION

Volume 12

2024

Reproducibility and Replicability in Communication Research

Open Access Journal

ISSN: 2183-2439

Edited by Johannes Breuer and Mario Haim

Media and Communication, 2024, Volume 12
Reproducibility and Replicability in Communication Research

Published by Cogitatio Press
Rua Fialho de Almeida 14, 2º Esq.,
1070–129 Lisbon
Portugal

Design by Typografia®
http://www.typografia.pt/en/

Cover image: © metamorworks from iStock

Academic Editors
Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies)
Mario Haim (LMU Munich)

Available online at: www.cogitatiopress.com/mediaandcommunication

cogitatio

Table of Contents

Media and Communication
2024 • Volume 12 • Editorial 8382
https://doi.org/10.17645/mac.8382

cogitatio

EDITORIAL

Open Access Journal

Are We Replicating Yet? Reproduction and Replication in Communication Research

Johannes Breuer [1,2] and Mario Haim [3]

[1] Department Computational Social Science, GESIS—Leibniz Institute for the Social Sciences, Germany
[2] Research Data & Methods, Center for Advanced Internet Studies, Germany
[3] Department of Media and Communication, LMU Munich, Germany

Correspondence: Johannes Breuer (johannes.breuer@gesis.org)

Submitted: 26 March 2024 **Published:** 19 June 2024

Issue: This editorial is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

The replication crisis has highlighted the importance of reproducibility and replicability in the social and behavioral sciences, including in communication research. While there have been some discussions of and studies on replications in communication research, the extent of this work is significantly lower than in psychology. The key reasons for this limitation are the differences between the disciplines in the topics commonly studied and in the methods and data commonly used in communication research. Communication research often investigates dynamic topics and uses methods (e.g., content analysis) and data types (e.g., media content and social media data) that are not used, or, at least, are much less frequently used, in other fields. These specific characteristics of communication research must be considered and require a more nuanced understanding of reproducibility and replicability. This thematic issue includes commentaries presenting different perspectives, as well as methodological and empirical work investigating the reproducibility and replicability of a wide range of communication research, including surveys, experiments, systematic literature reviews, and studies that involve social media or audio data. The articles in this issue acknowledge the diversity and unique features of communication research and present various ways of improving its reproducibility and replicability, as well as our understanding thereof.

Keywords

communication research; meta-science; open science; replicability; reproducibility

Scientific progress is based on the premise that the generation of knowledge is cumulative and that science is self-correcting. To ensure that scientific research is cumulative and self-correcting, the reproducibility and replicability of empirical research play key roles. The replicability of research, especially in the social and

behavioral sciences, has become a topic of raised academic but also public interest with the dawn of the so-called "replication crisis" in the early 2010s, sparked by failed attempts at systematically replicating several seminal findings in psychology. Subsequently, similarly low replication rates have also emerged in other disciplines, including more broadly in the social sciences (Camerer et al., 2018).

Replicability has multiple definitions. The glossary of the Framework for Open and Reproducible Research Trainin, hence, describes it as "an umbrella term, used differently across fields" (Parsons et al., 2022). A common definition that we also use in this article is that of The Turing Way Community (2022): "A result is replicable when the same analysis performed on different datasets produces qualitatively similar answers." Notably, there are different types of replications. A common distinction is between a direct or exact replication and a conceptual replication. According to Shrout and Rodgers (2018), "a direct or exact replication is a new study that employs the same procedure, materials, measures, and study population as the original study," and "a conceptual replication is intentionally different from a direct replication and is designed to assess [the] generalizability, as well as [the] veracity, of a result" (p. 492). Replication differs from *reproduction*, which means that "the same analysis steps performed on the same dataset consistently [produce] the same answer" (The Turing Way Community, 2022). In the wake of the replication crisis, researchers have also started to discuss and assess reproducibility. Although all research should, ideally, at least be reproducible, empirical investigations in the social and behavioral sciences have revealed that substantial parts of published research are not reproducible (see, e.g., Artner et al., 2021) and that the main reasons for this are a lack of shared data and code as well as errors in code execution, and insufficient documentation (Hardwicke et al., 2020; Krähmer et al., 2023; Trisovic et al., 2022).

Replicability has also been discussed in communication research, with initial studies on this subject dating back to several decades before the onset of the current replication crisis (Kelly et al., 1979). Following the identification and assessment of the replication crisis in psychology and other disciplines, in 2018, the journal *Communication Studies* published a Special Issue for Replications, collecting a series of nine replication studies from different subfields of communication research (volume 69, issue 3). However, despite these and similar singular efforts, such as the Open Science theme for the 2020 Annual Conference of the International Communication Association or a subsequent issue in the *Journal of Communication* (volume 71, issue 5), what McEwan et al. (2018) wrote in their editorial for the aforementioned special issue is still true: that "direct replications…are few and far between" (p. 236). McEwan et al. also pointed out the following important reasons for this observation: "engaging in replications is often undervalued" and "replication studies can be difficult to publish" (p. 325).

Notably, these challenges are not unique to communication science. However, in communication science, unlike in psychology, the replication crisis has not led to large-scale replication efforts or substantial cultural changes in the ways research is conducted and in the requirements it should meet. This is partly because the replication crisis originated in psychology, but this reason cannot fully explain the differences between psychology and communication science in dealing with questions of replicability. Three aspects that play a major role in this regard are the (a) methods and (b) data types that are commonly used in these fields, as well as the (c) topics that are studied in communication research. Although there are some overlaps in the methods and data types used in psychology and communication (e.g., experiments and self-reports, which are widely used in psychology, are also widely used in communication research), communication researchers also widely employ other methods and data types, such as content analyses, qualitative interviews, media content,

or social media data. This has implications for the nature and relevance of replications not only at the practical level but also at the conceptual level.

Notably, the distinction between reproduction and replication is less clear in communication research than in other fields. For example, numerous communication studies have used X (formerly Twitter) data collected via the platform's application programming interface (API). This API used to be freely and easily usable for academic research, but its terms of service allowed only the sharing of post IDs (not the full texts). Over time, if one were to use shared post IDs to recollect the original data (a process known as rehydration) in order to run the same analyses as in the original study, some posts and accounts might have already been deleted. In addition, API features might have already changed. Thus, the data will likely be different. Does this then constitute a reproduction or a replication?

Apart from this conceptual question of what replicability and reproducibility mean in and for communication research, another important question is when to expect reproducibility and/or replicability in communication research. To give yet another example, in content analyses of news reports at a specific time and on a specific topic, the findings might not be expected to be replicated in empirical efforts conducted on the same topic but several years later or in a different country. This contextual dependency of replicability in communication research starkly contrasts with the case in, for example, large parts of psychology, where successful replications are considered crucial estimates of the robustness and generalizability of research findings.

Aside from conceptual questions on reproducibility and replicability in communication research, there are also empirical questions. These empirical questions are related to (a) the prevalence and degree of reproducibility and replicability and (b) the ways of facilitating and improving reproductions and reproducibility, as well as replications and replicability. However, answering such questions requires conducting actual reproductions and replications, as well as systematic assessments of research practices in the field, and how they can affect reproducibility and replicability. As noted before, the three main factors that set communication research apart from other disciplines and are important to consider when answering conceptual as well as empirical questions related to reproducibility and replicability are: (a) the topics studied in communication research, (b) the methods employed in communication research, and (c) the various types of data that are used by communication scholars.

A key defining feature of many of the topics that are studied in communication research is their dynamic and, in part, also transient nature. Media content, technology, platforms, and usage patterns often undergo rapid transformations. In addition, studies regularly consider topics from current events. Accordingly, unlike studies in psychology that aim to understand the essential or universal mechanisms of human cognition and behavior, for large parts of communication research, cultural and/or temporal differences are expected to strongly affect study outcomes and, thus, to limit replicability.

Although communication research widely uses methods that are also used in other areas of the social and behavioral sciences, such as surveys, experiments, and interviews, one method is originally associated with the discipline: content analysis. Content analyses in communication research often investigates media content and thereby looks at timely topics, such as news reporting. Another more recent methodological development in communication research is the increasing use of computational methods, especially within

the growing subfield of computational communication science (see van Atteveldt & Peng, 2018). The use of machine learning algorithms, natural language processing techniques, and network analysis introduces new challenges for reproducibility and replicability, as these methods often involve probabilistic procedures and require documentation of additional steps and decisions, such as data preprocessing or parameter (and, in the case of machine learning, also hyperparameter) selection.

As for the methods, there are some overlaps between communication research and other fields in the social and behavioral sciences with regard to data types. Many studies rely on self-report data, which are typically gathered via surveys, interviews, or questionnaires in experimental studies. However, some data types are much more widely used in communication research, such as media content and social media data. In general, much of communication research draws upon data governed by platform-specific terms of service, proprietary content, or individual-level data that are subject to privacy concerns. This limits the capacity for data sharing due to privacy and copyright regulations (Davidson et al., 2023; van Atteveldt et al., 2020) and, thus can reduce reproducibility.

Overall, the special features of communication research necessitate a more nuanced perspective on reproducibility and replicability. Instead of asking only how reproducible or replicable communication research is, we should also ask what kinds of reproduction and replication are possible and informative, and what is needed to enable or facilitate different kinds of reproduction and replication. For example, in their analysis of communication science studies published between 2007 and 2016, Keating and Totzkay (2019) found that conceptual replications are much more common than direct replications. Considering the topics, methods, and data types that are often used in communication research, conceptual replications may generally be more appropriate than direct or exact replications. Another type of replication that may be particularly suitable for communication science is *prospective replication*, in which researchers plan "a series of replication studies that may occur simultaneously or at different times" (Steiner et al., 2019, p. 281). A helpful clarification question in addition to what type of replication is being pursued is what steps (or parts thereof) should be reproduced or replicated at all: data collection, its processing, the analysis, or the compilation and interpretation of results.

The eight articles included in this thematic issue reflect the breadth and complexity of replicability and reproducibility in communication research and address different conceptual, methodological, and empirical questions in various ways. In particular, this thematic issue combines three replication studies, three methodological articles, and two commentaries.

Knöpfle and Schatto-Eckrodt (2024), in their combined reproduction and replication study, empirically assess the challenges in working with social media data, in this case, with data from X. Although the findings from the original study could largely be reproduced using the same data, in the replication attempt, only slightly more than half of the posts could be recollected, which led to substantial differences in the results. Dubèl et al. (2024) also replicate a previous study that found that viewers are more aroused by negative than positive news, by conducting a laboratory study that combined physiological measures with self-reports. They also extend the result of the previous study by repeating the study in another country and using additional measures.

Adopting a methodological perspective, Lukito et al. (2024) present their study on the implications of tool choice and preprocessing of audio data for the reproducibility and replicability of studies that use this data

type. Although they found that the tools they tested provided accurate automated transcriptions, they note subtle yet significant differences between tools that could also impact reproducibility and replicability. Rogge et al. (2024) propose a standardized sampling method as a way of ensuring reproducibility and replicability for systematic literature reviews. Their method represents a structured multistage approach that can complement and extend existing guidelines for systematic literature reviews. The article of Vermeulen et al. (2024) offers a conceptual and practical exploration of direct replication in experimental communication science. They argue that replication studies in communication research almost always require the adaptation of at least some parts of the original design, extend existing replication typologies by adding the dimension of the motivation behind a replication study, and provide recommendations for replicators.

Xu and Zhang (2024) replicated and extended an earlier survey study on data-sharing practices among psychologists with a sample of Chinese communication scholars. Building on the theory of planned behavior, they find that various factors, including perceived risks and benefits, subjective norms, and pressure from journals, influence attitudes toward data sharing. Based on their findings, they present practical suggestions for improving research practices that can facilitate reproduction and replication in communication research. In his commentary, Ivory (2024) illustrates the urgent implications of the replication crisis for communication research. Acknowledging different perspectives on issues related to reproducibility and replicability and potential solutions, his discussion focuses on the dimensions of responsibility to the public, stewardship of resources, and membership in a community of scholars. Finally, in another commentary, Bowman (2024) stresses the continued need for replications in communication research. He reflects on key issues and recent developments and discusses replication as a key element of postpositivist approaches.

We find the collection of articles in this thematic issue, taken together, not only insightful but also highly representative of the breadth of perspectives and challenges pertaining to reproducibility and replicability in communication research. We believe that the different contributions in this issue can help in arriving at a better understanding of the nature and relevance of reproducibility and replicability in communication research, as well as of potential challenges and ways to address those.

Acknowledgments

First and foremost, we thank the authors of the articles in this thematic issue for their contributions. We also thank the reviewers of this thematic issue. Their expert feedback on the articles has helped significantly in ensuring and improving the overall quality of the issue. We are also grateful to the members of the projects in the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) priority program META-REP: A Meta-Scientific Programme to Analyse and Optimise Replicability in the Behavioral, Social, and Cognitive Sciences, with whom we had many fruitful and inspiring exchanges about reproducibility and replicability and their meaning and assessment, as well as ways of improving them. Finally, we extend our appreciation to the editorial team at *Media and Communication* for their support in the entire process of planning and publishing this thematic issue.

Funding

The editing of this thematic issue is part of the work in the project What Defines and Affects Replicability in Computational Communication Science? (project number 464291459) funded by the DFG under the META-REP priority program (project number 441890184).

Conflict of Interests

The authors declare no conflict of interests.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Bowman, N. D. (2024). On the continued need for replication in media and communication research. *Media and Communication*, *12*, Article 7935.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., Nave, G., Nosek, B. A., Pfeiffer, T., Altmejd, A., Buttrick, N., Chan, T., Chen, Y., Forsell, E., Gampa, A., Heikensten, E., Hummer, L., Imai, T., . . . Wu, H. (2018). Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour*, *2*(9), 637–644. https://doi.org/10.1038/s41562-018-0399-z

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., Van Der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*, 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Dubèl, R., Schumacher, G., Homan, M. D., Peterson, D., & Bakker, B. N. (2024). Replicating and extending Soroka, Fournier, and Nir: Negative news increases arousal and negative affect. *Media and Communication*, *12*, Article 7807.

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Crüwell, S., & Ioannidis, J. P. A. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014–2017). *Royal Society Open Science*, *7*(2), Article 190806. https://doi.org/10.1098/rsos.190806

Ivory, J. D. (2024). Remembering reasons for reform: A more replicable and reproducible communication literature without the rancor. *Media and Communication*, *12*, Article 7852.

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239. https://doi.org/10.1080/23808985.2019.1632218

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*(4), 338–342. https://doi.org/10.1111/j.1468-2958.1979.tb00646.x

Knöpfle, P., & Schatto-Eckrodt, T. (2024). The challenges of replicating volatile platform-data studies: Replicating Schatto-Eckrodt et al. (2020). *Media and Communication*, *12*, Article 7789.

Krähmer, D., Schächtele, L., & Schneck, A. (2023). Care to share? Experimental evidence on code sharing behavior in the social sciences. *PLoS ONE*, *18*(8), Article e0289380. https://doi.org/10.1371/journal.pone.0289380

Lukito, J., Greenfield, J., Yang, Y., Dalhke, R., Brown, M. A., Lewis, R., & Chen, B. (2024). Audio-as-data tools: Replicating computational data processing. *Media and Communication*, *12*, Article 7851.

McEwan, B., Carpenter, C. J., & Westerman, D. (2018). On replication in communication science. *Communication Studies*, *69*(3), 235–241. https://doi.org/10.1080/10510974.2018.1464938

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., Norris, E., O'Mahony, A., Parker, A. J., Todorovic, A., Pennington, C. R., Garcia-Pelegrin, E., Lazić, A., Robertson, O., Middleton, S. L., Valentini, B., McCuaig, J., Baker, B. J., Collins, E. . . . Aczel, B. (2022). A community-sourced glossary of open scholarship terms. *Nature Human Behaviour*, *6*(3), 312–318. https://doi.org/10.1038/s41562-021-01269-4

Rogge, A., Anter, L., Kunze, D., Pomsel, K., & Willenbrock, G. (2024). Standardized sampling for systematic literature reviews (STAMP method): Ensuring reproducibility and replicability. *Media and Communication*, *12*, Article 7836.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487–510. https://doi.org/10.1146/annurev-psych-122216-011845

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift Für Psychologie*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

The Turing Way Community. (2022). *The Turing way: A handbook for reproducible, ethical and collaborative research*. Zenodo. https://doi.org/10.5281/zenodo.3233853

Trisovic, A., Lau, M. K., Pasquier, T., & Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, *9*, Article 60. https://doi.org/10.1038/s41597-022-01143-6

van Atteveldt, W., Althaus, S., & Wessler, H. (2020). The trouble with sharing your privates: Pursuing ethical open science and collaborative research across national jurisdictions using sensitive data. *Political Communication*, *38*(1/2), 192–198. https://doi.org/10.1080/10584609.2020.1744780

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2/3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

Vermeulen, I., Masur, P. K., Beukeboom, C. J., & Johnson, B. J. (2024). Direct replication in experimental communication science: A conceptual and practical exploration. *Media and Communication*, *12*, Article 7971.

Xu, J., & Zhang, R. (2024). Attitudinal, normative, and resource factors affecting communication scholars' data sharing: A replication study. *Media and Communication*, *12*, Article 7666.

About the Authors

Johannes Breuer (PhD) is a senior researcher and leader of the team Digital Society Observatory in the Department of Computational Social Science at GESIS—Leibniz Institute for the Social Sciences in Cologne, Germany, and the team Research Data and Methods at the Center for Advanced Internet Studies (CAIS) in Bochum, Germany. His research interests include the use and effects of digital media, digital trace data and computational methods, as well as open science and meta-science. More information: https://www.johannesbreuer.com

Mario Haim (PhD) is full professor for Communication Science with a focus on Computational Communication Research at the Department of Media and Communication at LMU Munich, Germany. His research interests circle around algorithmic influences, such as in political communication, journalism, health communication, or media use, as well as on (computational) methods and meta-science. More information: https://haim.it. Photo by Lena Fleischer.

Media and Communication
2024 • Volume 12 • Article 7789
https://doi.org/10.17645/mac.7789

cogitatio

ARTICLE

Open Access Journal

The Challenges of Replicating Volatile Platform-Data Studies: Replicating Schatto-Eckrodt et al. (2020)

Philipp Knöpfle [1] and Tim Schatto-Eckrodt [2]

[1] Department of Media and Communication, Ludwig Maximilian University of Munich, Germany
[2] Department of Communication and Journalism Studies, Hamburg University, Germany

Correspondence: Philipp Knöpfle (philipp.knoepfle@ifkw.lmu.de)

Submitted: 31 October 2023 **Accepted:** 26 February 2024 **Published:** 15 April 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Replication studies in computational communication science (CCS) play a vital role in upholding research validity, ensuring reliability, and promoting transparency. However, conducting such studies in CCS often proves challenging due to the data environments' dynamic nature and the complexities surrounding data and software sharing. To shed light on these challenges, we examine the replication process with CCS studies by computationally reproducing and replicating Schatto-Eckrodt et al.'s (2020) computational analysis of the X (formerly Twitter) debate about the term "gaming disorder" being added to the International Classification of Diseases 11. Our results indicate a reproduction success rate of 88.46% of the original findings. Replicating the analysis presents several obstacles, particularly in data access and availability. Five years after the original data collection, we were able to recollect only 55.08% of the initial sample, primarily due to user and platform activities, including account deletions, user suspensions, and privacy settings. Our reproduction and replication efforts revealed intricate challenges in conducting CCS research, particularly concerning data access and third-party platforms. To enhance replication in CCS, we emphasize the crucial role of data sharing, increased transparency, extensive documentation, and regulatory processes. Thus, our analysis underscores replications' critical role in enhancing CCS research validity and reliability.

Keywords

computational communication science; replicability; replication; reproducibility; Twitter

1. Introduction

Replication studies play a critical role in scientific research, functioning as a litmus test of research findings' validity and reliability. Their significance cannot be underestimated in their contribution to promoting

transparency (Munafò et al., 2017) and fostering trust in empirical results (Rosenthal, 1991). Replications should be viewed as a social science principle because they strengthen the foundation of rigorous and verifiable research, fortifying empirical findings' validity and reliability (Benoit & Holbert, 2008). In meta-science, the concepts of reproduction and replication play pivotal roles in assessing scientific research reliability and robustness. Conducting a reproduction involves faithfully recreating a study's original conditions, methodologies, and analyses, including the acquisition of original data, to verify whether the same or quantitatively similar results can be obtained. Thus, reproduction concerns verifying the original study's validity and reliability (National Academies of Sciences, Engineering, and Medicine, 2019). However, replication entails conducting a novel study that makes variations to the original study's parameters (e.g., methodology, context, sample populations, etc.) to investigate the initial results' generalizability and robustness (National Academies of Sciences, Engineering, and Medicine, 2019). Replications can span a continuum of methodological resemblance to an initial study (LeBel et al., 2017; Machery, 2020). For example, a direct replication, characterized by high methodological similarity to the original study, involves reiterating the latter using methods as closely aligned with the original as reasonably possible. The goal is to anticipate consistent results based on current understanding of the phenomenon (Nosek et al., 2012). Conversely, conceptual replications can be used as experimental procedures designed to assess generalizability and veracity by varying a study's selected operational characteristics, such as omitting or including certain variables (Hendrick, 1990). In communication science, conceptual replications are more common than direct replications (Keating & Totzkay, 2019).

Nevertheless, the pursuit of replication and reproduction studies entails overcoming multiple barriers, such as (technical) resource constraints and complex research designs (Peng, 2011). This is particularly true in the field of computational communication science (CCS), in which researchers regularly face delicate and complex data environments and experimental contexts (van Atteveldt & Peng, 2018). CCS researchers also face multiple obstacles when it comes to conducting replications, e.g., working with personal or sensitive data, dealing with copyright restrictions, and relying on third-party social media platform data. Researchers in the CCS field commonly lack control over the data they work with, as they lack authority over generation of analyzed data, e.g., setting up or documenting experimental data collection protocols. In the replicability context, this means that while reacquiring experimental data for replication studies involves reproducing original experimental conditions to collect new empirical observations, reacquisition of content data is a very different endeavor. Content data, particularly in the case of social media data, is primarily not designed for research purposes and is inherently subject to algorithmic confounding (e.g., algorithmic opacity, data sampling biases, etc.), i.e., beyond researchers' control (Haim, 2023). Such data are made accessible to both developers and researchers as a byproduct of the data's inherent availability within the platform's framework (Bruns, 2019). Consequently, most content data are not intentionally structured for seamless replication (Davidson et al., 2023). Thus, CCS researchers are subject to stringent constraints imposed by technical restrictions, as well as terms of service imposed by platform providers to control access to and distribution of their data (Puschmann, 2019). These technical restrictions create significant barriers to sharing platform data with potential replicators, presenting a notable challenge because data access frequently functions as an essential prerequisite for successful reproduction and replication of a scientific analysis (Peng & Hicks, 2021).

Large social media platforms, such as X and Reddit, have become staple suppliers of data for CCS researchers (van Atteveldt & Peng, 2018). However, this dependence on third-party platform providers

introduces significant challenges to achieving replication success (Davidson et al., 2023), particularly in the contemporary "post-API" (application programming interfaces) era, in which researchers face heightened difficulties in accessing data from social media platforms (Tromble, 2021). Re-collecting a study's original data is a necessary and pivotal step in the reproduction and replication process, thereby eliciting the question of to what extent reproductions and replications of social media analyses remain feasible in the current platform ecosystem. Our analysis investigates this by conducting a reproduction and replication of an analysis by Schatto-Eckrodt et al. (2020), who investigated discourse on the X platform concerning the introduction of the term "gaming disorder" into the International Classification of Diseases 11 (ICD-11). We reproduce the original analysis by reconducting the original analysis on Schatto-Eckrodt et al.'s data. A reproduction to us denotes researchers' capacity to re-execute or re-implement the same computational analyses and yield results that are quantitatively identical (Christensen & Miguel, 2018; Cohen-Boulakia et al., 2017; Peng, 2011). In the second step, we conduct a direct replication with rehydrated data from the original analysis. Our replication approach is to re-collect Schatto-Eckrodt et al.'s data on the X platform and conduct an identical analysis to verify to what extent a replication of original results is possible using today's data. The concepts of reproducibility and replicability are intertwined in our study. While we have access to the initial data and code necessary for our reproduction, for our replication, we must re-collect the data from the X platform. If the original authors had not provided us with their data, our reproduction approach would have necessitated a similar data re-collection procedure. Consequently, our replication also can be understood as a computational reproduction scenario comprising a situation in which the data are not initially accessible to the reproduction team. Thus, our insights into the re-collection of social media platform data are applicable to both replication and reproduction scenarios, providing additional valuable considerations for situations in which the reproduction team must recollect the original study's data in today's context. In our study, our objective is to shed light on the intricacies and challenges associated with conducting replication and reproduction studies in the CCS field, particularly in the contexts of data access and third-party platform providers, as Freiling et al. (2021) and Davidson et al. (2023) highlighted. We showcase unique obstacles that CCS researchers face and underscore recent API restrictions and cost increases' implications for replication analyses' feasibility.

2. Methodology

2.1. Reproduction: Data Collection and Analytical Strategy

In many cases, replicating scientific studies proves challenging due to research designs' complexities. As a result, reproduction has become the prevalent standard for evaluating scientific claims' credibility and significance (Peng, 2011; Peng & Hicks, 2021). Nevertheless, reproduction often suffers from constraints analogous to conducting replication analyses. One crucial step behind every reproduction is the reacquisition of the data from the original study. Overstating data sharing's importance is difficult because replicators benefit significantly when they can access data. Unfortunately, in most cases, the original study's researchers often have little incentive to share their data, e.g., sharing their materials can incur technical expenses (Longo & Drazen, 2016), and they may be under legal restrictions that prohibit them from disclosing their data (Rosenberg et al., 2020). If a study's original data are not available to reproducers, initial sample data from the original study must be re-collected, which frequently involves substantial initial expenditures, offering limited prospects for the reproducer, as exact recollection of primary data is not guaranteed. The latter is particularly problematic when reproducing content data analyses on social media

platforms, where users and providers constantly generate, edit, and/or remove content. Moreover, most social media platforms, such as X, generally prohibit public sharing of user-level data obtained from the platform in their developer agreements. For our reproduction, we acquired the primary data set directly from the original authors, allowing us to compare the re-queried data with the original data. Considering that our study utilizes components from prior research—including code, data, and documentation—to reproduce initial findings, we classified our investigation as a computational reproduction (Ziemann et al., 2023), which differs from other types of reproducibility, such as analytical reproducibility (Hardwicke et al., 2018), also known as recreated reproducibility (Dreber & Johannesson, 2023), in which the goal is to reproduce an original study's findings based on information and documentation provided in the original article without access to the raw or processed data.

Notably, even though no shortage of resources exists for promoting reproducible research practices (Alston & Rick, 2021; Munafò et al., 2017; Stodden et al., 2014), no standardized templates or structured guidelines to date exist for documenting systematic reproduction of a computational social science article. Our approach aimed to address this gap using a simple methodology: We started by cataloging all critical empirical claims, including text passages and visual elements, from Schatto-Eckrodt et al., and setting out to reproduce these claims using the data and code that the original authors provided. This comparative analysis allowed us to verify each claim's accuracy on a granular level systematically, which facilitated the assessment of reproduction success. We viewed a reproduction as "successful" when we could reproduce the original result quantitatively, ensuring that it aligns with the claim presented in the original research. Notably, different standards are used to assess a reproduction's success rate, e.g., the exact quantitative agreement of results, margin of error (i.e., whether the reproduction results in a predefined range or margin of deviation), or by comparing statistical measures' (e.g., coefficient signs, confidence intervals, and/or hypothesis decisions) congruence. We chose the exact quantitative agreement of results, as it is the strictest criterion for assessing reproduction success. We allowed for minor code modifications to accommodate changes in software and dependencies over time, ensuring that the research remains adaptable and relevant.

2.2. Data Collection Replication Strategy

The Schatto-Eckrodt et al. data set was obtained originally by querying a database generated from the Decahose X API for gaming-disorder-related search terms. The Decahose API provides access to a 10% random sample of all public posts. Our study goal was to analyze to what extent we could replicate the data set with the data available today, so we aimed to re-collect the data under today's X access opportunities. Unfortunately, the Decahose API currently is limited to enterprise-level users, rendering it unavailable for our data collection purposes.

This left us with two viable data collection options: The first approach entailed reinitiating the identical search query applied to the Decahose data set on the X API. Subsequently, all posts and their associated information were collected accordingly. However, current X API access solutions have certain limitations, particularly when conducting queries encompassing wide-ranging terms, such as "gaming/games" and "disorder," as in our case. Open-ended, extensive queries swiftly deplete the allocated monthly post query limit, rendering the process time-consuming and costly. For this project, we adopted an alternative strategy. A "loophole" for researchers is to share each post IDs in their initial sample such that replicators selectively can re-query these exact posts at the API, a process also known as *rehydration*. This gives us precise

knowledge about what specific information to request from the X API. Under current API circumstances, this allows for the most resource-efficient data collection. Moreover, the rehydration approach is well-suited for projects with a clearly defined event timeline because the chance of "missing out" on additional data outside of the primary sample is minimal when an analysis has a fixed scope and time frame. To sum up, rehydration emerged as the most cost-effective and pragmatic choice for our replication under current X access model restrictions. Notably, none of the approaches guaranteed that the initial data set could be fully re-collected in its original form.

3. Reproduction of Original Findings

3.1. Reproduction Setup and Computational Environment

The reproduction process was conducted in both the original computational environment (as documented in Schatto-Eckrodt et al.; Supplementary File A) and a contemporary updated environment (see our software bibliography in the Supplementary Material or our Open Science Framework [OSF] repository), with all *R* programming language dependencies updated to their latest versions. Schatto-Eckrodt et al. provided the computational environment and analysis code in an OSF project. The shared materials comprised individual files for each analysis, a README file with usage instructions, an *R* session info file, and all shareable data compliant with X's terms of service. For more detailed recommendations on how to make computational communication research more reproducible and accessible, see van Atteveldt et al. (2019). By reproducing the study in the original computational environment, we ensured that the results were consistent with the conditions under which the original findings were generated. Simultaneously, the examination in the updated environment ensured that the research would remain adaptable and relevant in the face of evolving open-source software landscapes. The reproduction was executed using the statistical programming language *R* (Version 4.0.0 in the original environment and Version 4.3.1 in the updated environment) in July and September 2023. Tim Schatto-Eckrodt was part of the original study, and his participation in the reproduction presented a conflict of interest, so Philipp Knöpfle conducted all reproduction activities and evaluations.

3.2. Reproduction Evaluation

In our computational reproduction, we identified 26 empirical claims in the original paper overall, including text passages, figures, and tables. We successfully reproduced 23, an 88.46% reproduction success rate. All tables from the original study were reproduced. Notably, minor challenges emerged during the reproduction of Figures 2 and 3 from the original article due to the deterministic reproducibility requirement for network illustration methods, which requires setting a seed, i.e., an aspect overlooked in the original script. To address this issue in future analyses, it is recommended that a seed be determined for illustration methods, incorporating a randomness component to ensure consistent and reproducible visualizations. Furthermore, one *t*-test statistic, as documented on page 211 of Schatto-Eckrodt et al., could not be recalculated in our reproduction. However, this non-execution did not alter the test outcome's results substantially and was used in a comparison argument, thereby not affecting the core conclusions of their research. Overall, the replication phase proceeded without larger issues. All code was executed without critical errors in its initial environment. Only minor code adjustments due to depreciated package dependencies for the exporting of results had to be made in the updated environment. Overall, this finding

was positive for stability across computing environments. Our reproduction protocol can be found in the article's supplementary files, as well as the article's OSF repository (https://osf.io/2jb9m).

4. Replication of Findings

4.1. Rehydrated Sample

The "basic" subscription model—which replaced the previous, original, free access model in March 2023—permits a rate limit of 10,000 posts per month at a cost of $100 per month. The data collection process spanned two months—August and September 2023—yielding 9,270 successfully queried posts in the rehydrated sample. Table 1 illustrates the rehydrated sample's status, revealing that only 55.08% of the initial sample remained accessible on the X platform in September 2023. The data set experienced considerable changes: Users removed 28.32% of posts (deleted); 5.58% were set to private (protected); 10.62% were removed from the X platform because they violated the X terms of service (suspended); and 0.39% were generated by users who chose to deactivate their accounts temporarily (deactivated). This data landscape mirrored findings from a similar analysis in the original study, in which the original authors investigated how much of their sample was still available in 2020 by querying the X Compliance Firehose API. More than a quarter of the data already were unavailable, just two years after the data were collected initially in 2018. While existing literature on data loss rates on social media platforms supports these findings, other scholars have reported a diverse range of deletion rates. Depending on the time frame, deletion rates varied widely, with examples including 2% within one day after posting (Almuhimedi et al., 2013), 35.14% after one week (Zhou et al., 2016), 11.11% after five weeks (Bhattacharya & Ganguly, 2021), and 3.2% after two months (Petrovic et al., 2013). Specifically analyzing suspended X users' characteristics, Wei et al. (2016) found that X suspended 7.19% of the users after three years. While the rates found in the literature depended greatly on the studied content and user group (e.g., Zhou et al., 2016, focused on individual users' "regrettable" posts by individual users), studies conducted on random samples generated similar deletion rates: 7.27% after one week and 21.65% after six months (Schatto-Eckrodt, 2022).

Our analysis substantiated this, finding that the X trend of users opting to delete or protect their posts has increased in the past three years. Table 2 compares post types in the original and rehydrated samples, indicating that 3,074 original posts and 4,487 reposts could not be re-collected in 2023, resulting in an overall absolute data erosion of 7,561 posts, comprising 44.92% of the initial sample. Overall, in absolute terms, more reposts than original posts became inaccessible over the years. Figure 1 illustrates the data loss in post composition. The data loss in the rehydrated sample occurred particularly during the most heated phases of the discussion, as can be seen in Figure 2. Altogether, 1,934 posts were lost around the time of

Table 1. Sample changes from 2020 to 2023.

Example	2020	2023
Online	73.12%	55.08%
Deleted	12.24%	28.32%
Protected	1.30%	5.58%
Suspended	13.34%	10.62%
Deactivated	—	0.39%

Table 2. Changes in post composition from 2020 to 2023.

Post type	2020 (% of the original sample)	Data Loss (% compared to the original sample	2023 (% of the rehydrated sample)
Original posts	7,555 (44.89%)	−3,074 (40.69%)	4,481 (48.34%)
Re-posts	9,276 (55.11%)	−4,487 (48.37%)	4,789 (51.66%)
Sum	16,831 (100%)	−7,561 (44.92%)	9,270 (100%)

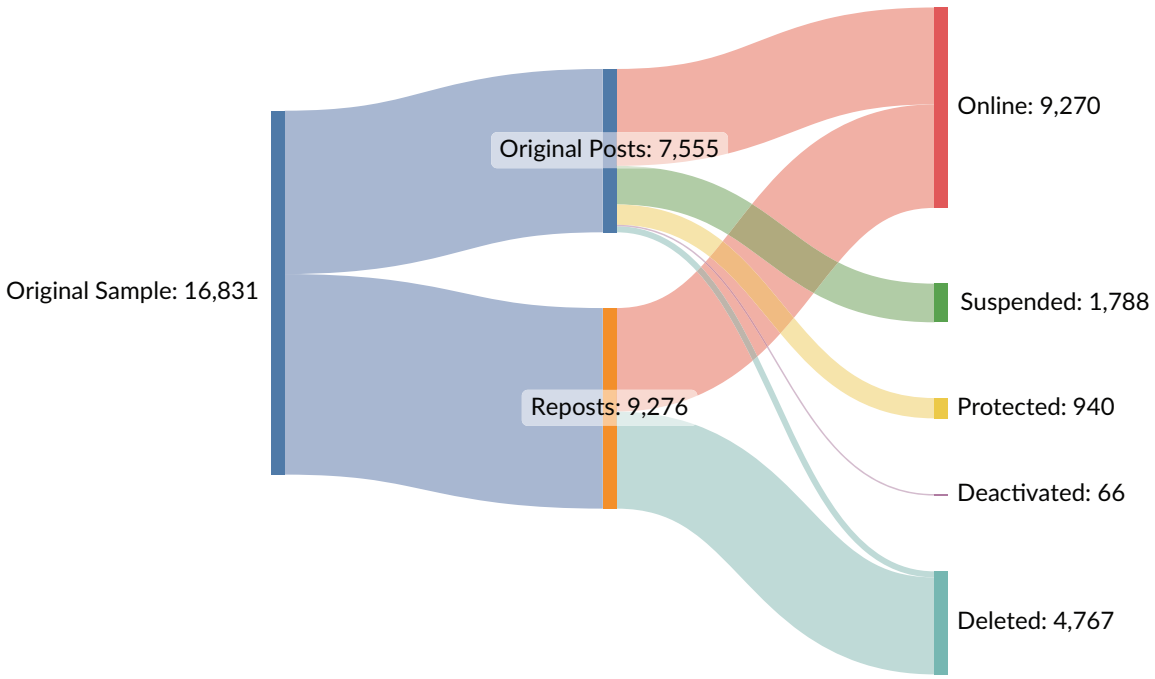

Figure 1. Sankey-diagram of the sample's structure from the original sample to the rehydrated sample.

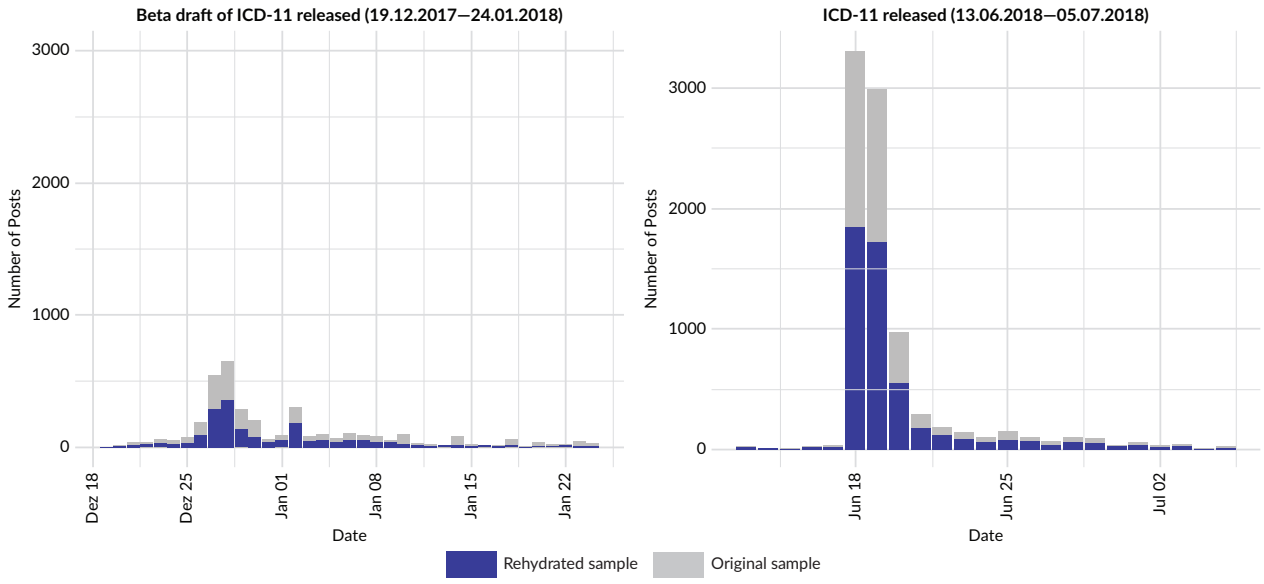

Figure 2. Sample loss over time.

the ICD-11 beta draft release (December 1, 2017–February 1, 2018) and 5,227 during the ICD-11 release (June 1, 2018–July 30, 2018), coinciding with the heated period of the discussion. Most data points were missing from the discussion's central segments on the days the ICD-11 beta draft was discussed (December 26, 2017) and on the day of the introduction (June 18, 2018). In the original sample, 15,845 unique users altogether participated in the discussion. However, in the rehydrated sample, data from only 8,621 of those individuals could be retrieved successfully.

4.2. Changes in the Analysis of Hashtags and Main Actors in the Debate

To better understand the key topics and participants in the debate, Schatto-Eckrodt et al. conducted an initial analysis that focused on influential actors and topics. Table 3, based on the original and rehydrated data, indicates a re-creation of Schatto-Eckrodt et al.'s Table 3, which highlighted top users based on the extent of their participation in the debate. Four of the top 10 participants were suspended due to X terms of service violations in the rehydrated sample. The reasons for these suspensions were not made public. The number of posts in the rehydrated data set stayed relatively stable for the remaining top users. Almost all posts from users who are still on the platform were rehydrated successfully. Most data attrition for the main actors in the debate happens on a user level. Regarding perspectives on the World Health Organization's (WHO) decision, no discernible patterns emerged in the data loss favoring or opposing the decision.

Table 4 indicates systematic differences between the original rehydrated data for the top users based on their reach, as measured as reposts. While the top 10 users with the most reposts remained the same, their ranking order shifted. In the rehydrated data set, significant data loss of approximately 30–40% was found for each user. CNN, the only traditional media outlet in the top 10, exhibited the most prominent reduction in repost support, with a decrease of 40.44%. The loss of repost data appeared to be homogenous across users. Notably, users with the highest reach registered a relative increase in sample size when comparing the original sample, suggesting that data loss disproportionately affects users with a smaller reach.

Table 3. Re-created top users according to their extent of participation.

User	Posts by the user (in the rehydrated sample)	% of all posts (in the rehydrated sample)	Status	# Followers in September 2023	View on the WHO decision	Notes
MommyNooz	32 (0)	0.19 (0)	Suspended	n/a	Supporting	Parenting blog
camerondare	31 (0)	0.18 (0)	Suspended	n/a	Supporting	Activist
AvocateforEd	25 (25)	0.15 (0.27)	Active	29,272	Supporting	Blog
Lynch39083	22 (22)	0.13 (0.24)	Active	43,298	Supporting	Scholar/activist
techedvocate	22 (22)	0.13 (0.24)	Active	22,082	Supporting	Tech blog
eplayuk	16 (16)	0.10 (0.17)	Active	284	Opposing	Gaming blog
Gamescosplay	14 (14)	0.08 (0.15)	Active	627	Opposing	Gaming blog
Gamingthemind	14 (12)	0.08 (0.13)	Active	3,758	Opposing	NGO/activists
Pairsonnalites	13 (0)	0.08 (0)	Suspended	n/a	Opposing	NGO
HealthyWrld	12 (0)	0.07 (0)	Suspended	n/a	Neutral	Health blog

We also replicated Schatto-Eckrodt et al.'s analysis of the most prominent topics in the debate, i.e., the examination of the most salient hashtags in the discourse. Table 5 presents our results. Out of the initial 3,017 hashtags, we successfully rehydrated posts with 1,691 hashtags, illuminating a consistent data erosion

Table 4. Re-created top users according to their reach.

User	Times the user was reposted (of the rehydrated sample)	% of all posts (of the rehydrated sample)	View on the WHO decision	# Followers	Notes
CNN	403 (240)	2.39 (2.59)	Neutral	61,766,081	Media
Deadmau5	385 (257)	2.29 (2.77)	Opposing	3,287,491	Musician
GaijinGoombah	318 (221)	1.89 (2.38)	Opposing	71,334	Youtube content creator
BrendoTGB	274 (153)	1.63 (1.65)	Opposing	294	Regular user
LEGIQN	245 (158)	1.46 (1.70)	Opposing	309,287	Twitch content creator
Pamaj	236 (159)	1.40 (1.72)	Opposing	1,191,142	E-Sports athlete
NoahJ456	233 (158)	1.38 (1.70)	Opposing	1,198,840	Youtube content creator
TheSmithPlays	227 (131)	1.35 (1.41)	Opposing	229,963	Youtube content creator
Boogie2988	182 (114)	1.08 (1.23)	Opposing	482,982	Youtube content creator
CaptainSparklez	170 (112)	1.01 (1.21)	Opposing	4,257,088	Youtube content creator

Table 5. Re-created top hashtags.

Hashtag	Number of occurrences (in the rehydrated sample)	% of hashtag occurrences (in the rehydrated sample)
Gaming	269 (150)	8.92 (8.87)
Gamingdisorder	123 (78)	4.08 (4.61)
Datascience	121 (2)	4.01 (0.12)
Addiction	113 (51)	3.75 (3.02)
Health	102 (50)	3.38 (2.96)
Icd11	97 (69)	3.22 (4.08)
Mentalhealth	97 (76)	3.22 (4.49)
Bbcbreakfast	70 (0)	2.32 (0)
Videogames	61 (38)	2.02 (2.25)
Parenting	59 (8)	1.96 (0.47)
Who	47 (38)	1.56 (2.25)
Gamingaddiction	46 (32)	1.52 (1.89)
Children	40 (5)	1.33 (0.30)
News	36 (19)	1.19 (1.12)
tech	32 (21)	1.06 (1.24)

across the top 15 hashtags. Remarkably, certain hashtags—such as "datascience," "parenting," and "bbcbreakfast"—experienced substantial data loss, reaching 98.35%, 86.44%, and 100%, respectively. This data attrition pattern suggests the absence of entire conversational threads within the discourse related to specific hashtags in the rehydrated data set. Otherwise, data loss occurred systematically across topical hashtags and variations thereof, as well as for individual hashtags most likely associated with journalistic reporting ("#bbcbreakfast" and "#datascience") and parenting ("#parenting"). Even though the omitted hashtags within the top 15 have been substituted by hashtags with a similar thematic tone ("edtech," "games," and "worldhealthorganization"), they represent very different conversation threads. While the rehydrated versions largely maintained unity with the original table, the process of rehydrating revealed that a considerable number of subconversations on the discussion surrounding the inclusion of "gaming disorder" in the ICD-11 led by certain hashtags was lost.

Overall, we found mixed success when replicating the explorative analyses in Schatto-Eckrodt et al. Although many of the replicated tables exhibited high congruence with their counterparts in the initial study, pronounced data loss was found across the rehydrated data set, which effectively affected the analysis of the most prominent actors and topics in the debate. Specifically, we noted that data decay in the context of original posts tends to manifest as binary absence, with all post data per user unattainable.

4.3. Changes in Sentiment Analysis

To investigate the prevailing sentiment within the discourse, the original authors conducted a sentiment analysis with 141,908 sentiment-labeled tokens, revealing an overall negative tone in the discussions. While acknowledging data pre-processing's considerable influence on sentiment analysis outcomes, particularly in the context of X data (Krouska et al., 2016), we replicated the initial sentiment analysis by adhering to the same preprocessing procedures as in Schatto-Eckrodt et al. Our replication sentiment analysis, with 78,661 sentiment tokens, reaffirmed this observation, indicating a consistently negative sentiment. Moreover, upon a detailed examination, we observed a marginal increase in the negativity of sentiments within the rehydrated data set compared with the original sample. Building on the methodology used by Schatto-Eckrodt et al., who examined weekly sentiment fluctuations to follow the discourse's evolving tone over time, our analysis revealed disparities in the absolute weekly sentiment measures between the original and rehydrated samples, reaching statistical significance at the 10% level (t [88] = 1.93, $p < 0.057$, $d = 0.12$). Notably, these differences stemmed from variations in the magnitude of sentiments, rather than the fundamental nature of the sentiments themselves. Furthermore, our analysis did not reveal any statistically significant differences in sentiment between Segment 1 (March 16, 2017–November 30, 2017) and the combined sentiment in Segments 2 (December 1, 2017–June 14, 2018) and 3 (June 15, 2018–November 15, 2018; t [48] = 1.49, $p < 0.1427$, $d = 0.36$), as was found in the original analysis. While we could not completely replicate the original study's results, our replicated sentiment analysis indicated extremely similar patterns observed by Schatto-Eckrodt et al., which is impressive considering the significant data loss of 44.57% in sentiment-labeled tokens.

4.4. Changes in the Topic Model

Finally, we calculated a structured topic model (Roberts et al., 2019) while adhering to the same pre-processing steps that the original authors used. The original sample comprised 5,378 documents, of

which we could replicate only 3,266 documents in the rehydrated sample. We used the same rule of thumb as the original authors to choose the number of topics, i.e., the elbow method on the semantic coherence and held-out likelihood. Semantic coherence measures topics' "interpretability." A higher semantic coherence suggests that topics are more interpretable by humans, as terms within a topic are related closely. The measure aids in understanding topics' meaningfulness. Held-out likelihood evaluates the model's ability to predict unseen data. A higher held-out likelihood indicates that the model captured underlying structures and patterns in the text data. Table 6 presents the results from these two measures for a topic model calculated on the rehydrated data set. In our case, a topic model with three topics ($K = 3$) had the highest held-out likelihood and semantic coherence, and it seemed to be the best choice based on the two metrics, although admittedly, the difference in semantic coherence between topic models was small. Notably, a topic model originally chosen by Schatto-Eckrodt et al. (see their Table 6), featuring $K = 5$, proved to be an ill-suited fit for our data. This discrepancy likely arose from the sample size's impact. Larger data sets tend to produce more stable and robust topic models. By reducing the sample size, we introduced more variability in topic assignments, leading to less reliable results in the replicated topic model. Moreover, we found no change in the prevalence of topics during certain phases of the sample. All three topics, highlighted in Table 7, were discussed to an equal extent over the sample period. However, Topics 1 and 2 were discussed considerably more than Topic 3. To sum up, replication of the topic model did not reflect the original study's results, as it arrived at a different optimal number of topics ($K = 3$) than the original study ($K = 5$), and we

Table 6. Topic model metrics for the rehydrated sample.

	$K = 3$	$K = 4$	$K = 5$
Held-out likelihood	−2.260403	−2.27303	−2.271619
Semantic coherence	−91.21908	−97.38662	−105.8436

Table 7. Re-created topic model: Description, top terms, and representative quote of the topics.

Topic	Description	Terms	Quote
1	This topic focuses on discussions of health aspects related to gaming disorder	health, YouTube, condition, addiction, mental, official, disease	"US Health News: If Gaming Addiction Is Now a Mental Health Disorder, How Can We Fight It? #MentalHealth" "Playing too many video games can cause a mental health disorder, says World Health Organization"
2	This topic revolves around issues related to the classification and recognition of gaming addiction	addiction, mental, condition, organization, recognize, classification, health	"WHO to classify 'Gaming Disorder' as a mental health condition in 2018" "Mental Health Experts Warn Against World Health Organization's Definition of 'Gaming Disorder'"
3	This topic involves discussions about the influence of video games, gaming habits, and their impact on individuals and society	video, classification, play, people, time, disease, official	"So, gaming disorder only is a thing if you play games so much that you are physically, mentally, and socially atrophying…like every other disorder built around doing too much of something" "Imo it is an overreaction especially when it's mostly aimed at kids. It's the time in their life for them to be able to play games as much as they do, but now it is gonna be classed as a mental disorder if you play too much"

could not replicate the topics that Schatto-Eckrodt et al. found semantically. Reasons for the latter could very well be the reduced rehydrated sample size.

5. Discussion

The reproduction exercise's results demonstrated a high success rate in reproducing empirical claims from the original study, with 88.46% of the claims successfully reproduced. Challenges arose primarily from issues related to deterministic reproducibility in network illustration methods and a minor error in reporting the correct *t*-statistic. These results emphasize the importance of following proper coding practices and documentation standards in CCS research to maintain research outcomes' integrity, reproducibility, and reliability.

Replicating the original analysis introduced unique challenges, particularly social media data's volatile nature and X API's changing access policies and limitations. The rehydrated sample, collected under strict, new X API data access restrictions, revealed that only 55.08% of the initial sample remains accessible today—a data loss that affected our replication results significantly. Data attrition becomes more pronounced during intense discussion phases and is characterized by a complete presence or absence of data for a user. We have demonstrated that data attrition can be linked to user deletions, post-privacy settings, and account suspensions. Thus, data loss over extended periods—in our case, five years—when dealing with social media data is a consequence of the evolving nature of data on social media platforms. This raises concerns about the long-term sustainability and feasibility of conducting replications based on social media analyses, particularly when working with nonrecent data. Replications may fail for several reasons, such as an initially incorrect finding by the original authors, changes in the measured phenomenon over time, or a lack of generalizability of a finding because it is population-specific (e.g., Dienlin et al., 2021). Our examination revealed that the use of volatile data (e.g., constant changes in user behaviors, account statuses, and data accessibility) also can have a significant impact on studies' replication success.

Our study is subject to several constraints. First, the employed rehydration strategy does not entail drawing a completely new sample. This would be the ideal approach for generalizing Schatto-Eckrodt et al.'s findings. Instead, we reacquired the original data, potentially perpetuating any biases or shortcomings present in the original data set. Another important aspect to note is that X currently manages post edits by deleting the original post ID and assigning a new one (authors' note: Post editing, at the time this was published, was available exclusively to X users who are subscribed to X Premium, a paid service). Consequently, the rehydration method does not capture edited posts, which may have led to a slight overestimation of data loss in our sample. Furthermore, we had to work with X API's technical and financial restrictions, so the sample size in the replication is relatively small, limiting the findings' overall generalizability.

Notably, we replicated an exploratory analysis, which possesses lower evidentiary weight, as its findings are less definitive and more preliminary in nature. Future research should examine how data erosion can affect predictive and inferential analyses. Given that inferential analyses frequently form the foundation for policy recommendations and decision-making, understanding the extent of their vulnerability to data loss in dynamic data environments is important.

Our replication serves as a hypothetical scenario illustrating how a reproduction could have been undertaken in a scenario with no access to the original data. Our replication results indicate that we would not have been

able to replicate most of Schatto-Eckrodt et al.'s findings without the data. This inability to replicate—caused predominantly by data fluctuations, rather than improper documentation—underscores data sharing's crucial role in replication efforts. Transparency and accessibility of data and code are almost prerequisites for conducting replications (Marsden & Pingry, 2018). Unfortunately, data sharing is not yet the prevailing practice in CCS, and in some instances, it is legally and technically unfeasible. This problem is worsened by the unpredictability of reacquiring data effectively, as our study has demonstrated. Furthermore, recent increases in financial and access barriers imposed by major social media platform providers have exacerbated this situation. When significant barriers hinder access to materials needed for replication, conducting replication studies becomes more unlikely. In cases in which direct replication is improbable, and conceptual replication is difficult, researchers must hope for regulatory changes in platform access, such as those that the EU's Digital Services Act (DSA) has been promising. This raises a fundamental question about the core value of a scientific analysis when technical and financial obstacles hinder or prevent other researchers from replicating the analysis. The value of a scientific analysis traditionally lies in its potential to contribute to the body of knowledge and to be subjected to testing and validation through replication (National Academies of Sciences, Engineering, and Medicine, 2019). However, when substantial barriers disrupt replication efforts, the value of the analysis itself should be brought into question. In such challenging circumstances, it is imperative for the CCS community to promote practices that mitigate these barriers actively. Encouraging widespread data sharing, advocating for greater transparency, more collaboration between researchers, and establishing standardized protocols for replication attempts can help bridge the gap between original studies and their validation (Dienlin et al., 2021). Moreover, regulatory institutions are crucial in guaranteeing access to social media platforms for CCS researchers at a time when social media platform APIs are threatening open scientific endeavors (Davidson et al., 2023). Through implementation and enforcement of policies that foster scientific access, these regulatory bodies not only can support CCS research, but can also help overcome barriers to replicability. Addressing these barriers is essential to upholding the core tenets of scientific inquiry and safeguarding the integrity of valuable research through rigorous examination.

6. Conclusion

This study highlights the importance of replication in enhancing the validity and reliability of CCS research. Our replication and computational reproduction efforts provide insights into the challenges and complexities of conducting and replicating CCS research, particularly in the context of data access and third-party platform providers. Our results emphasize the need for researchers to consider data loss and changes in data availability over time when conducting replication and reproduction studies in the CCS field, particularly when working with social media data.

While the practical recommendations drawn from the present study, urging researchers to prioritize reproducibility in the realm of volatile data access, can enhance CCS studies' overall quality and robustness, recognizing that the individual researcher is just one component within a broader framework is crucial: Regulatory bodies, social media platforms and their users, and journals and academic institutions also wield significant influence in this context. Thus, replicability in CCS is also a political and institutional issue. For example, enhancing CCS replicability can be achieved by providing researchers with increased access and control over their data, particularly in the realm of social media platform research. The DSA promises to play a central role in this context, as it offers, among many other benefits to researchers, clear regulations concerning access to data of "very large online platforms," i.e., platforms with more than 45 million monthly

active EU users. The DSA could establish the foundation for transparent social media and platform research, aiming to ensure that "socially relevant aspects of digitization can be investigated appropriately, consistently, and independently" (Klinger & Ohme, 2023, p. 3). Under Article 40 of the DSA, researchers affiliated with a research institution and independent of commercial interests will be able to request data from these platforms through a Digital Services Coordinator to conduct research on systemic risks in the EU. Systemic risks, as outlined in the DSA under Article 34(1), include negative effects on civic discourse and electoral processes, protecting public health, and dissemination of illegal content (European Centre for Algorithmic Transparency, 2023). A comprehensive scientific examination of digitization's socially relevant aspects necessitates researchers' ability to reproduce and replicate their findings within this context. Without strong supranational regulations, such as the DSA, many researchers' individual efforts to improve their work's reproducibility and replicability would be conducted in vain. In addition to regulatory actions' effects, the users who create the data being researched also should be considered: Analyzing content that users have deleted also poses ethical research questions that must be considered case by case. Examining actively harmful or anti-democratic entities' actions may justify the analysis of data deleted by users, whereas, in other instances, researchers may need to respect individual users' decisions to withdraw their content from public access. Contemplations on CCS research replicability should extend beyond individual researchers' efforts and include all pertinent stakeholders in the evaluation process. Thus, recognizing that incentive structures established by academic journals and publishers, alongside those of academic institutions, play a pivotal role in advancing the broader goal of enhancing research replicability is crucial. This includes incorporating systematic code and data review as integral components of the peer-review process, establishing infrastructure for responsible sharing of code and data in compliance with data privacy and access regulations, and allocating funding to bolster scientific results' robustness and reproducibility.

Increasing access to data is paramount in addressing reproducibility and replication challenges in CCS. Researchers' ability to access and analyze data directly impacts replication efforts' feasibility and robustness. Policies and regulatory changes, such as the DSA, play a pivotal role in facilitating such access, as they require platforms to provide data access to researchers, particularly concerning socially relevant digitization aspects. Improving data accessibility stands as a crucial measure in tackling reproducibility and replication hurdles in CCS. Moreover, overcoming barriers related to data sharing, transparency, and technical constraints is important to preserving the value of scientific analysis and promoting robust replication practices in CCS. We have demonstrated that reacquiring platform data is both resource-intensive and comes with no guarantee of fully regaining the primary sample. Thus, our study highlights vital aspects in establishing sustainable practices for reproducibility and replicability in CCS, such as data accessibility, data sharing, transparency, and comprehensive documentation of research methods and materials.

Funding

This research was supported by the German Research Foundation priority program META-REP: A Meta-Scientific Programme to Analyse and Optimise Replicability in the Behavioural, Social, and Cognitive Sciences (project number 441890184), with the subproject What Defines and Affects Replicability in Computational Communication Science? (project number 464291459).

Conflict of Interests

All reproduction analyses and the evaluation thereof were done by Philipp Knöpfle. The authors declare no conflict of interests.

Data Availability

The replication package for this study, containing all shareable data, code, and materials, is available at: https://osf.io/2jb9m

Supplementary Material

All supplementary material such as replication materials, software bibliography, reproduction protocol, etc. can be found in the OSF repository (https://osf.io/2jb9m).

References

Almuhimedi, H., Wilson, S., Liu, B., Sadeh, N., & Acquisti, A. (2013). Tweets are forever: A large-scale quantitative analysis of deleted tweets. In A. Bruckman & S. Counts (Eds.), *CSCW '13: Proceedings of the 2013 conference on computer supported cooperative work* (pp. 897–908). ACM. https://doi.org/10/gfzgsd

Alston, J. M., & Rick, J. A. (2021). A beginner's guide to conducting reproducible research. *The Bulletin of the Ecological Society of America*, *102*(2), Article e01801. https://doi.org/10.1002/bes2.1801

Benoit, W. L., & Holbert, R. L. (2008). Empirical intersections in communication research: Replication, multiple quantitative methods, and bridging the quantitative–qualitative divide. *Journal of Communication*, *58*(4), 615–628. https://doi.org/10.1111/j.1460-2466.2008.00404.x

Bhattacharya, P., & Ganguly, N. (2021). Characterizing deleted tweets and their authors. *Proceedings of the International AAAI Conference on Web and Social Media*, *10*(1), 547–550. https://doi.org/10.1609/icwsm.v10i1.14803

Bruns, A. (2019). After the 'APIcalypse': Social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, *22*(11), 1544–1566, https://doi.org/10.1080/1369118X.2019.1637447

Christensen, G., & Miguel, E. (2018). Transparency, reproducibility, and the credibility of economics research. *Journal of Economic Literature*, *56*(3), 920–980. https://doi.org/10.1257/jel.20171350

Cohen-Boulakia, S., Belhajjame, K., Collin, O., Chopard, J., Froidevaux, C., Gaignard, A., Hinsen, K., Larmande, P., Bras, Y. L., Lemoine, F., Mareuil, F., Ménager, H., Pradal, C., & Blanchet, C. (2017). Scientific workflows for computational reproducibility in the life sciences: Status, challenges and opportunities. *Future Generation Computer Systems*, *75*, 284–298. https://doi.org/10.1016/j.future.2017.01.012

Davidson, B. I., Wischerath, D., Racek, D., Parry, D. A., Godwin, E., Hinds, J., Van Der Linden, D., Roscoe, J. F., Ayravainen, L., & Cork, A. G. (2023). Platform-controlled social media APIs threaten open science. *Nature Human Behaviour*, *7*, 2054–2057. https://doi.org/10.1038/s41562-023-01750-2

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.

Dreber, A., & Johannesson, M. (2023). *A framework for evaluating reproducibility and replicability in economics*. SSRN. http://doi.org/10.2139/ssrn.4458153

European Centre for Algorithmic Transparency. (2023). *FAQs: DSA data access for researchers*. https://algorithmic-transparency.ec.europa.eu/news/faqs-dsa-data-access-researchers-2023-12-13_en

Freiling, I., Krause, N. M., Scheufele, D. A., & Chen, K. (2021). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*, *71*(5), 686–714. https://doi.org/10.1093/joc/jqab032

Haim, M. (2023). *Computational communication science: Eine Einführung*. Springer. https://doi.org/10.1007/978-3-658-40171-9

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Mohr, A. H., Clayton, E., Yoon, E. J., Tessler, M. H., Lenne, R. L., Altman, S., Long, B., & Frank, M. C. (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. https://doi.org/10.1098/rsos.180448

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: are they important? *Journal of Social Behavior and Personality*, *5*(4), 41–49.

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239.

Klinger, U., & Ohme, J. (2023). *What the scientific community needs from data access under Art. 40 DSA: 20 Points on infrastructures, participation, transparency, and funding* (Weizenbaum Policy Paper No. 8). Weizenbaum Institute for the Networked Society; The German Internet Institute. https://doi.org/10.34669/WI.WPP/8.2

Krouska, A., Troussas, C., & Virvou, M. (2016). The effect of preprocessing techniques on Twitter sentiment analysis. In N. Bourbakis, G. Tsihrintzis, M. Virvou, & D. Kavraki (Eds.), *2016 7th International Conference on Information, Intelligence, Systems & Applications (IISA)* (p. 144). IEEE. https://doi.org/10.1109/IISA.2016.7785373

LeBel, E. P., Berger, D., Campbell, L., & Loving, T. J. (2017). Falsifiability is not optional. *Journal of Personality and Social Psychology*, *113*(2), 254–261. https://doi.org/10.1037/pspi0000106

Longo, D. L., & Drazen, J. M. (2016). Data sharing. *New England Journal of Medicine*, *374*(3), 276–277. https://doi.org/10.1056/NEJMe1516564

Machery, E. (2020). What is a replication? *Philosophy of Science*, *87*(4), 545–567. https://doi.org/10.1086/709701

Marsden, J. R., & Pingry, D. E. (2018). Numerical data quality in IS research and the implications for replication. *Decision Support Systems*, *115*, A1–A7. https://doi.org/10.1016/j.dss.2018.10.007

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*, Article 0021. https://doi.org/10.1038/s41562-016-0021

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. National Academies Press.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, *7*(6), 615–631. http://doi.org/10.1177/1745691612459058

Peng, R. D. (2011). Reproducible research in computational science. *Science*, *334*(6060), 1226–1227. https://doi.org/10.1126/science.1213847

Peng, R. D., & Hicks, S. C. (2021). Reproducible research: A retrospective. *Annual Review of Public Health*, *42*(1), 79–93. https://doi.org/10.1146/annurev-publhealth-012420-105110

Petrovic, S., Osborne, M., & Lavrenko, V. (2013). *I wish i didn't say that! Analyzing and predicting deleted messages in Twitter*. arXiv. http://arxiv.org/abs/1305.3107

Puschmann, C. (2019). An end to the wild west of social media research: A response to Axel Bruns. *Information, Communication & Society*, *22*(11), 1582–1589. https://doi.org/10.1080/1369118X.2019.1646300

Roberts, M. E., Stewart, B. M., & Tingley, D. (2019). stm: An R package for structural topic models. *Journal of Statistical Software*, *91*(2), 1–40. https://doi.org/10.18637/jss.v091.i02

Rosenberg, D. E., Filion, Y., Teasley, R., Sandoval-Solis, S., Hecht, J. S., van Zyl, J. E., McMahon, G. F., Horsburgh, J. S., Kasprzyk, J. R., & Tarboton, D. G. (2020). The next frontier: Making research more

reproducible. *Journal of Water Resources Planning and Management*, *146*(6), Article 1820002. https://doi.org/10.1061/(ASCE)WR.1943-5452.0001215

Rosenthal, R. (1991). Replication in behavioral research. *Journal of Social Behavior and Personality*, *4*(4), 1–30.

Schatto-Eckrodt, T. (2022). Hidden biases—The effects of unavailable content on Twitter on sampling quality. In J. Jünger, U. Gochermann, C. Peter, & M. Bachl (Eds.), *Grenzen, Probleme und Lösungen bei der Stichprobenziehung* (pp. 178–195). Herbert von Halem Verlag.

Schatto-Eckrodt, T., Janzik, R., Reer, F., Boberg, S., & Quandt, T. (2020). A computational approach to analyzing the Twitter debate on gaming disorder. *Media and Communication*, *8*(3), 205–218. https://doi.org/10.17645/mac.v8i3.3128

Stodden, V., Leisch, F., & Peng, R. D. (2014). *Implementing reproducible research*. CRC Press.

Tromble, R. (2021). Where have all the data gone? A critical reflection on academic digital research in the post-API age. *Social Media + Society*, *7*(1). https://doi.org/10.1177/2056305121988929

van Atteveldt, W., & Peng, T.-Q. (2018). When communication meets computation: Opportunities, challenges, and pitfalls in computational communication science. *Communication Methods and Measures*, *12*(2/3), 81–92. https://doi.org/10.1080/19312458.2018.1458084

van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Computational communication science: Toward open computational communication science—A practical road map for reusable data and code. *International Journal of Communication*, *13*, 3935–3954. https://ijoc.org/index.php/ijoc/article/view/10631

Wei, W., Joseph, K., Liu, H., & Carley, K. M. (2016). Exploring characteristics of suspended users and network stability on Twitter. *Social Network Analysis and Mining*, *6*(1), 1–18. https://doi.org/10/gc92c9

Zhou, L., Wang, W., & Chen, K. (2016). Tweet properly: Analyzing deleted tweets to understand and identify regrettable ones. In J. Bourdeau, J. A. Hendler, R. N. Nkambou (Eds.), *WWW '16: Proceedings of the 25th International Conference on World Wide Web* (pp. 603–612). ACM. https://doi.org/10/gf4hpq

Ziemann, M., Poulain, P., & Bora, A. (2023). The five pillars of computational reproducibility: Bioinformatics and beyond. *Briefings in Bioinformatics*, *24*(6), Article bbad375. https://doi.org/10.1093/bib/bbad375

About the Authors

Philipp Knöpfle is a research associate at the Department of Media and Communication at Ludwig-Maximilian University of Munich (LMU). He obtained his MSc degree in Finance and Information Management from the Technical University of Munich, University of Augsburg, and University of Bayreuth. Currently, he is working at the Computational Communication Research unit under the supervision of Professor Haim at LMU. His primary research interests encompass meta science, open science, and computational methods.

Tim Schatto-Eckrodt is a research associate at the Department of Journalism Studies and Communication at the University of Hamburg, Germany. He holds an MA degree in Communication Studies from the University of Münster. He is currently working as the chair for Digital Communication and Sustainability at Hamburg University. His further research interests include computational methods and online propaganda.

Media and Communication
2024 • Volume 12 • Article 7807
https://doi.org/10.17645/mac.7807

ARTICLE

Open Access Journal

Replicating and Extending Soroka, Fournier, and Nir: Negative News Increases Arousal and Negative Affect

Roeland Dubèl [1], Gijs Schumacher [2], Maaike D. Homan [3], Delaney Peterson [1], and Bert N. Bakker [1]

[1] Amsterdam School of Communication Research, University of Amsterdam, The Netherlands
[2] Department of Political Science, University of Amsterdam, The Netherlands
[3] Organizational Behavior Group, University of Utrecht, The Netherlands

Correspondence: Bert Bakker (b.n.bakker@uva.nl)

Submitted: 2 November 2023 **Accepted:** 9 February 2024 **Published:** 11 April 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

The negativity bias hypothesis in political communication contends that people are more aroused by negative vs. positive news. Soroka et al. (2019) provide evidence for this negativity bias in a study in 17 countries across six continents. We find suggestive evidence for Soroka et al.'s (2019) central finding that negativity causes an increase in skin conductance levels in a conceptually close, well-powered, and preregistered replication. We extend Soroka et al. (2019) in three ways. First, we theorise, test, and confirm that negative (vs. positive) news causes an increase in activity of the corrugator major muscle above the eyebrow (using facial electromyography activity) and is associated with a negative affect. Second, we find people self-reporting negative news causes negative affect but that positive (instead of negative) news increases self-reported arousal. Third, we test Soroka et al.'s (2019) argument in another context, the Netherlands. Our article suggests that negative news is, especially, causing negative affect. Doing so, we contribute to the negativity bias argument in political communication research and, at the same time, show the importance of replication in empirical communication research.

Keywords

corrugator; negative news; negativity bias; physiology; skin conductance

1. The Negativity Bias in Political Communication Research

In a path-breaking study published in the *Proceedings of the National Academy of Sciences*, Soroka et al. (2019; SFN from here on) found that negative news is more arousing and attention-grabbing compared to positive

news. The study was conducted among 1,156 subjects in 17 countries on six continents and revealed that "reactions to video news content reveal a mean tendency for humans to be more aroused by and attentive to negative news" (Soroka et al., 2019, p. 18891). The finding, which deepened and broadened an earlier pilot study by Soroka and McAdams (2015) conducted in Canada, has important implications for political communication as a field. According to Soroka et al. (2019, p. 18888), "the importance of negativity biases for news is relatively clear" as "negativity biases affect news selection, and thus also news production, as well as citizens' attitudes about current affairs."

The work by SFN introduced a new theory in (political) communication research about responses to negative news. In an earlier study, Soroka and McAdams (2015, p. 4) summarised the argument for a negativity bias as follows:

> Humans have a reasonably well-established tendency to react more strongly to negative than to positive information; it follows that news content, created by humans, with the goal of getting attention from other humans, will tend to be biased toward the negative. Critical to this account is evidence that news content does indeed tend to generate stronger reactions and/or greater attentiveness when it is negative.

This innovative argument brings together research in evolutionary psychology—arguing that there are evolutionary benefits to being attuned to negativity—with literature in media systems suggesting that there might be cross-cultural differences in negativity bias. Moreover, SFN advanced the study of responses to news by turning to physiological measures of arousal and attention. This was combined with a unique cross-cultural data collection across six continents, thereby addressing concerns about the focus on the Western context in empirical social science research. The article, although relatively recent, made an impact on the field: It has been widely cited (311 citations in Google Scholar, as of January 11, 2024) and was viewed 167,135 times on the journal's website. The article was also covered in the press by dozens of news outlets including prominent outlets such as *The Washington Post*, *Medium*, *Al Jazeera*, *The Guardian*, *Newsweek*, and *Handelsblatt*, while being referenced on six Wikipedia pages.

The work by SFN has been foundational and, to us, inspirational. At the same time, replications are scarce in communication research (Keating & Totzkay, 2019). Especially direct replications—which closely follow the design and procedures of the original study—are sparse: In the period 2007 to 2016, 0.5% of the published papers were direct replications of a previously published study in the field of communication research (Keating & Totzkay, 2019). Our goal is to increase the percentage of published replication studies (a tiny bit). Therefore, we set out to directly replicate one of the two central findings of SFN. We preregistered that we expect that negative news compared to positive news increases skin conductance levels (SCLs) as an indicator of arousal (H1). It is important to replicate this finding because by closely following the design, stimuli, and measures, we can provide crucial information on the repeatability and robustness of SFN's findings (Chambers, 2017).

Note that we do not replicate the other finding of SFN, that negative news vs. positive news captures more attention, as measured with an increase in heart rate variability. Data collection was part of an omnibus study in which we did not collect heart rate to keep the protocol within the limits of an hour (i.e., attaching instruments for each physiological measure cost time).

Aside from replicating SFN, we follow King (1995) and extend the work of SFN in three ways. First, we theorise that negative news causes an increase in negative affect. SFN theorised and tested the effect of negative news on arousal. Yet, theories in psychology and neuroscience on the structure of affect indicate that affect consists of (at least) two dimensions. Arousal captures the intensity of the affective experience, but this intensity can be both negative and positive. Valence is the second dimension of affect and captures the directionality of the affect, ranging from positive to negative (Schiller et al., 2022). SCLs—which SFN used—capture arousal but do not distinguish the valence of the affective experience (Dawson et al., 2007). To capture valence, we turned to facial electromyography (fEMG) which registers rapid, partially automatic affective responses in the face while participants view stimulus material. fEMG is primarily used as an indicator of emotional valence (Larsen et al., 2003). We focus on the corrugator supercilii muscle region which is the muscle above the eyebrow that draws the brow down and pulls the brows together. The corrugator is used as an indicator of negative affect, also in political communication research. For instance, in response to political communication, messages incongruent with held beliefs produced more corrugator activity (negative affect) than congruent communication (Bakker, Schumacher, & Rooduijn, 2021; Boyer, 2023). We extend SFN and preregistered that we expect that negative news compared to positive news increases the activity of the corrugator muscle (H2).

Second, we broaden SFN by also measuring self-reported valence and arousal. SFN did not collect measures of either but did ask about the "tone of the message" which was: "Please rank this story on the following dimensions….Negativity." That is, however, not the same as measuring the affective response to the message. We agree with SFN that physiological measures have the advantage that they capture affective responses during exposure. At the same time, inspired by ongoing discussions in psychology and neuroscience (e.g., Arceneaux et al., in press; Barrett, 2017; Evers et al., 2014; LeDoux & Pine, 2016), it is valuable to capture different parts of the affective response: the more unconscious physiological responses during exposure and the more reflective evaluations of the affective experience after exposure. Physiological and self-reported affect are often loosely related (Bradley & Lang, 2000). Yet, we follow the implications of SFN and expect effects in similar directions as those reported with physiological measures. Specifically, for arousal, we expect that negative compared to positive news leads to a higher level of self-reported arousal (H3), while for valence we expect that negative compared to positive news leads to a higher level of self-reported negative valence (H4).

Third, and finally, we conducted our study in the Netherlands. Conducting this study in the Netherlands allows us to test the repeatability of SFN in another context. The Netherlands is the 18th country where SFN's argument is tested. By pooling our results with those of SFN, we get one (small) step closer to the population-based estimate of the effect of negativity on arousal.

2. Methods

2.1. Transparency and Ethics

We preregistered the study on the Open Science Framework on September 25, 2023 (https://osf.io/w2rkq). This was before we conducted the analyses but after data collection was completed. Deviations from preregistration are clearly flagged and all other tests and procedures should be considered as confirmatory and in line with preregistration. The raw data and code to reproduce the results in R can be found on our OSF page as well. The stimuli materials of SFN cannot be publicly shared due to copyright issues. The study

received ethical approval from our ethical review board 2022-AISSR-15445. Participants completed an informed consent form before the start of the study.

2.2. Sample and Procedure

In Autumn 2022 and Spring 2023, we recruited a total of 104 participants in an omnibus laboratory study. We conducted the study in the Netherlands at the university laboratory of our university. The study took one hour. In line with the guidelines of our ethics board, participants either received research credits or 15 euros in return for their participation. After preregistered preprocessing steps (see Section 2.3), we have 97 participants in the sample.

Upon signing the informed consent participants completed a survey on Qualtrics. Afterwards, a trained research assistant connected the equipment for skin conductance and fEMG. We recorded physiological responses using the Versatile Stimulus Response Registration Program 1998 (Vsrrp98) software with a sampling rate of 1,000 Hz. The lab equipment was able to reliably and validly capture fEMG activity and skin conductance in earlier work in other domains (see, for instance, Gazendam et al., 2013; Nohlen et al., 2016; Sevenster et al., 2015).

SCLs are measured by passing a small current through two electrodes placed on the skin. The electrodes were attached with adhesive tape on the medial phalanges surfaces of the index and ring finger of the non-dominant hand (Dawson et al., 2007). By keeping the current constant it is possible to measure the flow of the current—what we call skin conductance expressed in microsiemens (Dawson et al., 2007). SCL is a validated measure of arousal (for an introduction see, Settle et al., 2020) and is applied in (political) communication research (e.g., Carlson et al., 2020; Wang et al., 2014)

Activity of the corrugator major is associated with the experience of negative affect (Larsen et al., 2003). We measured this using two 7 mm Ag/AgCl mini-electrodes that we filled with electrolyte gel (Signa, Parker Laboratories). Using double-stick adhesive tape, we placed the two electrodes just above the eyebrow, at the place where the muscle is located (Fridlund & Cacioppo, 1986). A third electrode was placed on the middle of the forehead (just below the hairline) and served as a ground measure. The corrugator has no overlapping muscle groups, has a very limited representation in the motor cortex, and "tends to be bilateral innervated" (Larsen et al., 2003, pp. 776–777). The measurement of the corrugator is therefore less subject to disruptions from the (voluntary) movement of other muscles.

The original protocol of SFN was 25 minutes long. We included the stimuli for this article in a larger omnibus study and did not have space for 25 minutes of news clips. However, as the videos were shown in a random order and the focus was not on specific news items, we decided to select a subset of the videos. We did this in consultation with Soroka (the S in SFN). We picked four videos, two negative and two positive, that were shown in all 17 countries included in SFN. Moreover, the selected stories were consistently rated as negative (in the case of the negative stimuli) and positive in the case of the positive videos (see Figure S3 of the supporting materials in SFN). The negative videos were *Peru*, which describes "the small town of Chimbote burns down," and *Niger*, which describes "current food shortages in Niger." The positive videos were *Gorillas*, which describes "gorillas from a zoo are released into the wild," and *Young Director*, in which "an 11-year-old makes stop-motion films." For details on the validation of the stimuli, we refer to SFN who

provide evidence that the videos are valid. Like SFN, we randomised the order in which the four videos were shown.

After each video, participants were asked to self-report the arousal and valence (both on a nine-point scale) that they experienced while watching the videos. We used self-assessment manikins to measure arousal and valence, which is a valid and reliable way to measure self-reported arousal and valence (Bradley & Lang, 1994).

In addition to these stimuli, participants in the protocol evaluated candidates for political office and participated in an immigration framing experiment. These other stimuli are part of other papers. The protocol to replicate SFN was placed in the middle of the omnibus study.

2.3. Measures

After data collection, we conducted a visual inspection of our data for indications of (a) broken or malfunctioning electrodes, identified by a signal that drops dramatically, or (b) a loose electrode, identified by an unusual, stable up-and-down pattern, and (c) cross-referenced these unusual events from the log book (our trained lab assistants made notes about unexpected events). After this, we removed cases with clear distortions. We did this blind to the results and the experimental conditions. This procedure is highly similar to SFN. Following this, seven participants were excluded from the skin conductance analysis: three participants were excluded due to non-response, one participant was excluded due to under-responsiveness (both can occur due to equipment failure, e.g., loose sensors or disconnected cables), and lastly, three participants were excluded due to wrong synchronisation (i.e., the durations of baseline and/or experimental periods were wrongly recorded).

As for preprocessing the raw SCLs, similar steps as those taken by SFN were performed. A rolling average was applied within each participant by attributing slightly larger weights to the middle three values. This serves to smooth the raw scores, reducing the presence of possible outliers. To address the missing values that are produced due to the procedure at the beginning and end of each participant's time series, the closest available score was used to populate these gaps. Consequently, the normalised SCL (nSCL) was centred on the basis of the mean score of the preceding baseline period. SFN calculate the mean score after excluding the first and last five seconds of the baseline period. Given that our baseline period only spans eight seconds, we preregistered to not apply this step in our data but to take the full eight seconds baseline. We also preregistered to re-analyse the results using the last five seconds of the baseline and our results are not affected by this modelling decision.

The preprocessing for the raw fEMG data is as follows: First, the raw fEMG data is band-pass filtered between 20 and 400 Hz, with an additional 50-Hz notch filter and subsequently integrated (van Boxtel, 2010). fEMG data is further corrected using a Hampel filtering algorithm to filter for cross-talk (Bhowmik et al., 2017) and exclude outliers (observations higher or lower than four standard deviations from the mean). We take an individual time-specific baseline measure by taking the median of all fEMG observations during the interstimulus interval prior to the treatment. We then take the processed fEMG signal and divide it by the baseline with 100. This way the fEMG value represents the microvolt increase or decrease compared to an individual's fEMG readings prior to the treatment.

2.4. Analyses and Power

We rely on frequentist statistics and use the p-value of $p < 0.05$ as the cut-off for statistical significance. We engage in the more conservative two-sided tests. The analysis strategy per hypothesis is discussed in the results section. We present z-standardised coefficients to ease interpretation.

For all our analyses we conducted a power analysis after we completed data collection (the stopping rule for the data collection was decided by other parts of this omnibus study). For details, we refer to the pre-analysis plan (https://osf.io/w2rkq). The power analyses demonstrate that we are sufficiently powered ($\beta = 0.8$, α (two-sided) $= 0.05$) to detect small effects—based upon previous studies (Lakens, 2022)—for our tests of SCLs (arousal; H1) and corrugator activity (valence; H2) captured with physiological measures as well as valence captured with self-reports (H4). When it comes to self-reported arousal, it depends on the population-based effect size one would expect to conclude whether we are sufficiently powered. We expect small effects as the impact of news negativity on self-reported arousal is more ambiguous (the cognitive counterpart of the SCLs; Keib et al., 2018; Ravaja et al., 2015). It is therefore unclear whether our study has enough power to detect an arguably small treatment effect of negativity (vs. positivity) on self-reported arousal.

3. Results

The results section is structured as follows: We discuss the preregistered modelling strategy that belongs to the test of the hypothesis, followed by the results and, where appropriate, exploratory extensions. In each section, we flagged deviations from the preregistration.

3.1. Negative News Is Causing Increases in SCL (H1)

In line with SFN, the first hypothesis stated that negative news compared to positive news increases SCLs as an indicator of arousal. Participants (97 included in the analyses after preprocessing) saw four videos, lasting approximately one minute and 45 seconds. At the second-per-second level, this yields a total of 40,470 observations. The primary model that we used to test H1 is the fixed-effects within panel estimation of nSCL (for the report of these results, see Table S4 of the supporting information of SFN). The estimation was performed via the plm package (Version 2.6–3; Croissant & Millo, 2008) in R. In this model, the change in nSCL is predicted by a second-by-second negativity score of the video (provided by SFN), while controlling for the logarithm of seconds per video as well as the interaction with the negativity score, the lag of the dependent variable and story order. Note that a reanalysis of the original (preprocessed) data of SFN in R (as well as Stata), using this modelling strategy yielded the same outcomes as presented in the article and appendix by SFN; we thereby verified the results of SFN which is a good starting point for a replication study (Nuijten et al., 2018). A fixed-effects within panel estimation is equivalent to performing a dummy least squares (an ordinary least squares which includes dummies for each unit). This is confirmed by the original data and 2,000 simulated datasets. To conclude, the model we use to test H1 is the following:

$$\begin{aligned}\text{Changes in nSCL} = {} & \text{Negativity} + \text{Time (seconds, logged)} + \text{Negativity} \times \text{Time (seconds, logged)} \\ & + \text{Lagged nSCL} + \text{Story order}\end{aligned}$$

For H1 to be confirmed, we expected a positive and statistically significant effect of negativity. We find a positive but not statistically significant effect of negativity on SCLs ($\beta = 0.019$, $t = 0.984$, $p = 0.325$). This is a small effect in the hypothesised and preregistered direction but it is not statistically significant at the preregistered p-value cut-off. This small effect of negativity on SCLs also becomes apparent when inspecting the average SCLs over time (Figure 1), as there are small differences in SCL per condition.

Yet, when comparing the standardised negativity estimate from the article by SFN (2019; $\beta = 0.042$, CI = [0.032, 0.052]) and our replication ($\beta = 0.019$, CI = [−0.019, 0.056]), the estimates are in the same direction and while the confidence intervals overlap, our effect is 2.5 times smaller than SFN. For ease of comparison, Table 1's first column provides our replication model and SFN's model in parallel to each other.

Next, we performed an exploratory non-preregistered test to assess whether the negativity coefficient found by SFN differs from our outcome. We pooled the data from SFN and our replication sample. Consequently, we ran the pre-specified model on the pooled data adding two predictors: whether the respondent originated from the sample by SFN or the replication, and an interaction of the sample origin with the negativity coefficient. The latter would indicate whether the negativity coefficient differs significantly between samples. Given that fixed-effects within estimation cannot account for time-invariant features, which is the case for the sample origin as a respondent always belongs to the same sample, we ran a random-effects model instead. We find no support for the hypothesis that our coefficient is statistically significantly different than the one found by SFN: The interaction effect between the negativity indicator

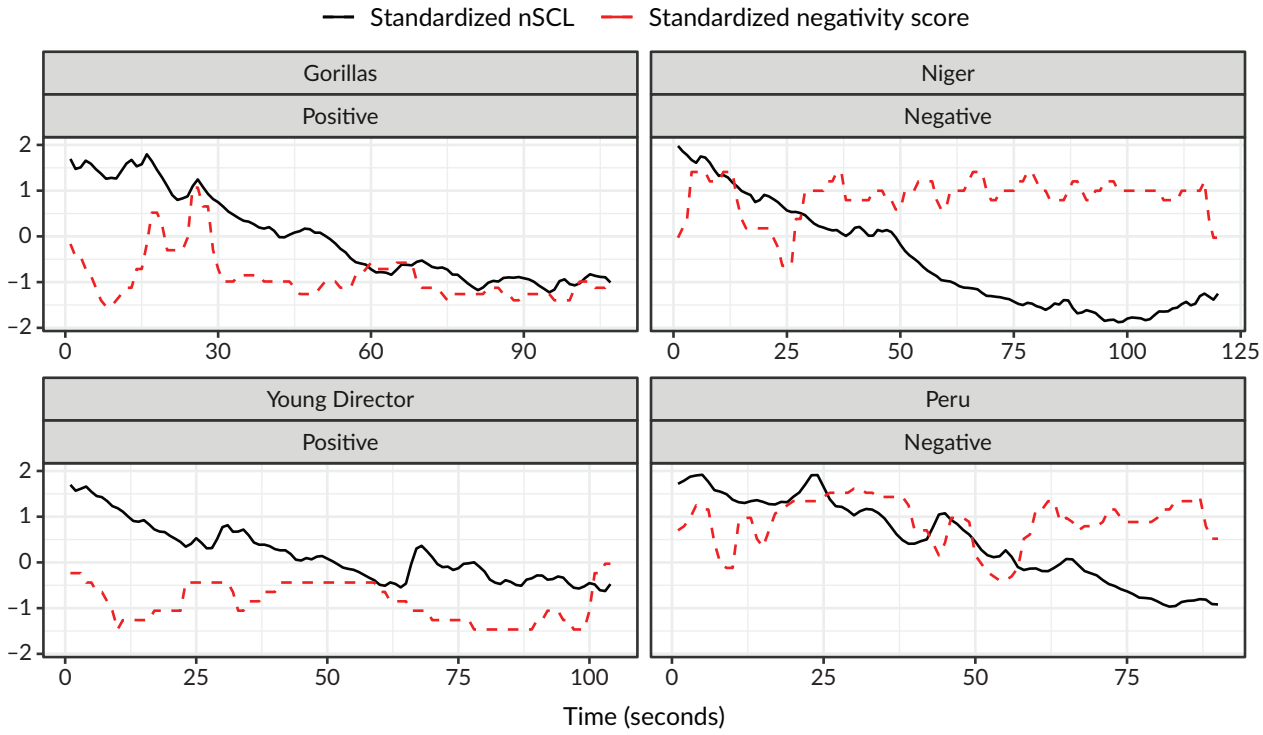

Figure 1. Average nSCL time series per story. Notes: The panels plot the average SCL per clip; the top of each panel indicates the name of the panel and whether its valence is positive or negative; the black line in each panel projects the z-standardised mean SCL activity over the seconds (x-axis) per video treatment (panel); the red line plots the z-standardised negative score of the video as scored per second and provided by SFN; the negativity score ranges from negative (positive values on the y-axis) to positive (negative values on the y-axis).

and the dummy variable capturing whether it was SFN's sample (0) or our replication sample (1) was very close to zero and not statistically significant, with $\beta = -0.002$, $t = -0.448$, and $p = 0.654$ (see Model 4, the pooled model with sample controls, in Table 1).

Comparing our effect to SNF's, we find no statistically significant difference. This, however, raises the question of the practical significance of our effect. For this, we turn to exploratory non-preregistered equivalence testing to determine whether our effect is practically equivalent to 0. When we use the parameters package (Version 0.21.3; Lüdecke et al., 2020) and assume the default bounds our negativity effect falls within the bounds of what is practically equivalent to 0, with CI 90% [−0.01, 0.05], range of practical equivalence [−0.1, 0.1], and $p < 0.001$. This suggests that our effect can be labelled as negligible.

However, we can also use the heuristic by Simonsohn (2015). This states that the smallest effect size of interest is the minimum effect size that the original study could have detected with 33% power. In the case of SNF, this is a β of 0.0018. Using this effect size as the upper (0.0018) and lower (−0.0018) bound, we conclude that our effect is not equivalent to 0. To put it differently, our effect is different from zero and falls within the bounds of SFN, CI 90% [−0.01, 0.05], range of practical equivalence [−0.0018, 0.0018], and $p = 0.954$.

To summarise, our study failed to replicate the statistical significance of the effect reported by SFN. Exploratory equivalence tests suggest that our SCL effect falls within the bounds of SFN, but its practical significance depends on the bounds chosen in the equivalence test. When we pool our data with SFN, we get a smaller yet statistically significant estimate of the effect of negativity on SCL (see Models 3 and 4 in Table 1).

Table 1. Effect of negativity on standardised change in nSCL.

	Standardised change in nSCL			
	Replication	SFN	Pooled	Pooled with sample controls
Time (seconds, logged)	−0.006 (0.006)	−0.004*** (0.001)	−0.001 (0.001)	−0.001 (0.001)
Time × Negativity	−0.006 (0.005)	−0.011*** (0.001)	−0.010*** (0.001)	−0.010*** (0.001)
Lagged nSCL	−0.100*** (0.006)	−0.082*** (0.001)	−0.059*** (0.001)	−0.059*** (0.001)
Order	0.010** (0.004)	−0.004*** (0.0003)	−0.003*** (0.0003)	−0.003*** (0.0003)
Negativity	0.019 (0.019)	0.042*** (0.005)	0.039*** (0.005)	0.040*** (0.005)
Negativity x Replication sample				−0.002 (0.004)
Replication sample				−0.023*** (0.005)
Observations	40,346	641,287	681,633	681,633
R^2	0.008	0.005	0.003	0.003
Adjusted R^2	0.005	0.003	0.003	0.003
F statistic	63.630***	628.749***	2,163.547***	2,182.103***

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

3.2. Negative News Is Causing Increases in Corrugator Activity (H2)

The second hypothesis stated that negative news would elicit more corrugator activity in the supercilii (as expressed through the fEMG measure). Similarly, as for nSCL, we use the fixed effects within panel estimation to analyse the effect of negativity on fEMG. The change in the corrugator variable is regressed on the second-by-second negativity score of the video while controlling for the logarithm of seconds per video as well as the interaction with the negativity score, the lag of the dependent variable, and story order:

$$\text{Changes in corrugator} = \text{Negativity} + \text{Time (seconds, logged)} + \text{Negativity} \times \text{Time (seconds, logged)} + \text{Lagged corrugator} + \text{Story order}$$

The interaction effect between negativity and time is in the predicted positive direction and close to the preregistered p-value cut-off $p < 0.05$ ($\beta = 0.038$, $t = 1.926$, $p = 0.054$; see the first column of Table 2). Importantly, the interaction term picks up the effect of a one-unit increase in negativity when time is one second (given that we control for the logarithm of time, where a logarithm of 0 equals 1). Moreover, when inspecting the average time series (Figure 2) of corrugator activity per story, the negative videos elicited greater corrugator activity compared to positive activity. This confirms there is a main effect of negativity on corrugator activity. A more simplified model—which we did not preregister—in which the interaction between corrugator activity and time is excluded produces a positive and statistically significant main effect of negativity on the corrugator ($\beta = 0.060$, $t = 13.569$, $p < 0.001$; see the second column of Table 2). This means that once negativity increases, corrugator activity increases. The standardised effect is small but

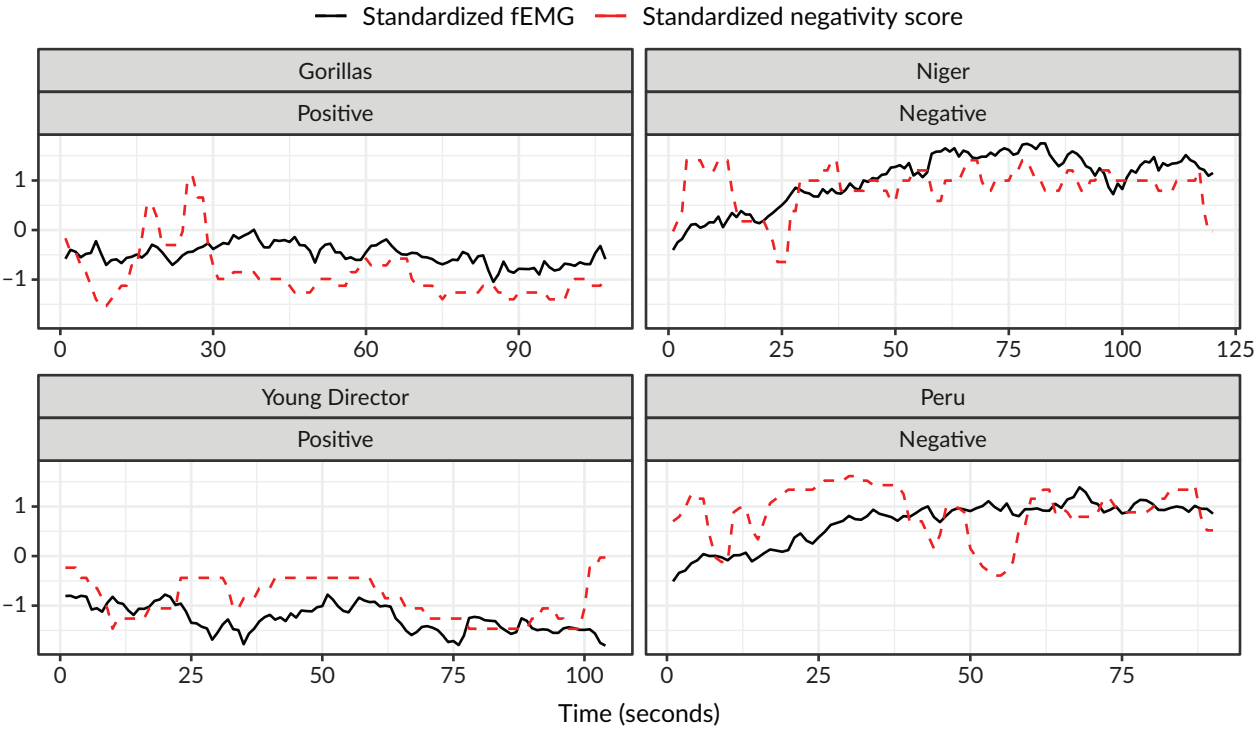

Figure 2. Average corrugator (fEMG) activity from time series per story. Notes: The black line projects the z-standardised mean corrugator (fEMG) activity over the seconds (x-axis) per video treatment (panel); the red line plots the z-standardised tone of the video; the negativity score ranges from negative (positive values on the y-axis) to positive (negative values on the y-axis).

Table 2. Effect of negativity on standardised change in fEMG.

	Standardised change in fEMG	
	Preregistered model	Exploratory model
Time (seconds, logged)	0.005 (0.006)	0.005 (0.006)
Time × Negativity	0.006 (0.005)	
Lagged fEMG	−0.004*** (0.0001)	−0.004*** (0.0001)
Order	−0.004 (0.005)	−0.004 (0.005)
Negativity	0.038* (0.019)	0.060*** (0.004)
Observations	37,844	37,844
R^2	0.030	0.030
Adjusted R^2	0.028	0.028
F statistic	235,679***	294.249***

Note: * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

larger than SFN's SCL effect and in line with other studies that rely upon the corrugator. Taken together we accept H2: Negative news increases corrugator activity.

3.3. Extensions Using Self-Reported Measures of Arousal (H3) and Valence (H4)

Finally, we hypothesised that the cognitive counterparts of our physiological measures would be affected in the same manner: Negative news compared to positive news would be perceived as both more negatively valenced and arousing. Both arousal and valence were measured using the manikins on a nine-point scale, in which 1 indicates *very positive* or *no arousal* and 9 indicates *very negative* or *very arousing*. Instead of discussing the outcomes of the preregistered fixed-effects models (see pre-analysis plan on OSF), Welch's two-sample *t*-tests will be discussed. These were performed with base R (Version 4.3.0; R Core Team, 2023). The Welch's variant was used given that the variances differ significantly across conditions. The interpretations do not change when relying on the preregistered models—Results can be derived from the replication files.

Negative news is indeed perceived as more negatively valenced ($M = 7.55$, $SD = 1.15$) compared to positive news ($M = 2.66$, $SD = 1.49$), $t = 36.013$, $p < 0.001$. This is in line with our preregistered hypothesis. Figure 3 visualises this pattern by showing the aggregate perceived valence per story. This indicates that the negative videos are rated much higher compared to the positive ones. This pattern replicates the findings for tone that SFN reported.

Contrary to our expectations, it is positive news ($M = 6.45$, $SD = 2.17$) instead of negative news ($M = 4.76$, $SD = 1.80$) which is perceived as more arousing, with $t = -8.28$ and $p < 0.001$. This pattern can also be deduced when inspecting the aggregates per story (Figure 3), highlighting that the centrality measures (the mean and median) are consistently higher for the positive videos compared to the negative videos. This is in line with Ravaja et al. (2015), who also found that positive news was perceived to be slightly more arousing as compared to negative news.

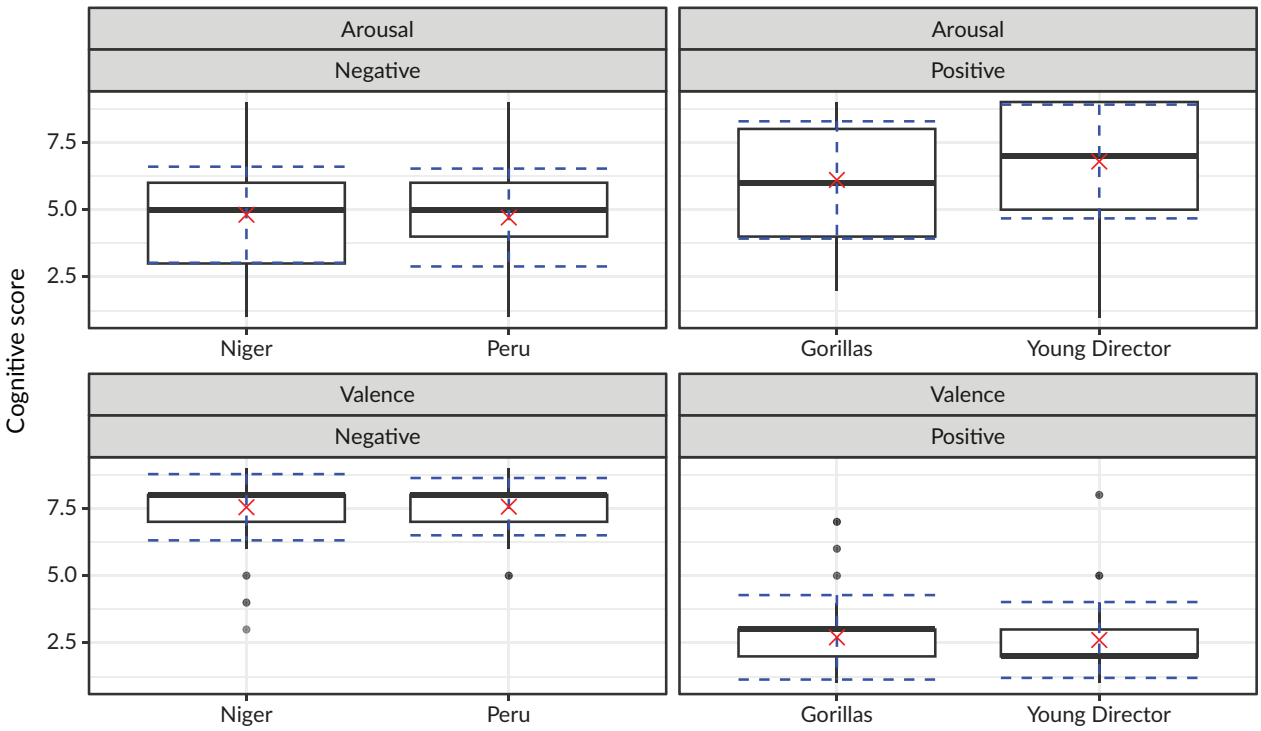

Figure 3. Valence and arousal in response to negative and positive news clips. Notes: The black lines depict the distribution via a boxplot; the red cross represents the mean; the blue dotted lines indicate the standard deviation from the mean; the top panel projects the self-reported arousal, from *low* (1) to *high* (9) in response to negative clips (left-hand panel) and positive clips (right-hand panel); the bottom panel projects self-reported valence, from *positive* (1) to *negative* (9) in response to negative clips (left-hand panel) and positive clips (right-hand panel)

3.4. Exploring the Alignment Between Physiological and Self-Reported Responses to News

Self-reported valence and the physiological measure of valence produce the same conclusions, while the results for arousal are mixed. Yet, there is ongoing discussion about the extent to which physiological responses and self-reported measures should, or should not, align in psychology and neuroscience (for a discussion and illustration, see Arceneaux et al., in press). Therefore, we explore to what extent these two measures correlate (note that this was not preregistered). If corrugator activity (SCL) and self-reported valence (arousal) are weakly correlated with each other, then this provides evidence that (political) communication evokes effects at the more conscious (self-reports) and unconscious (physiology) levels which are relatively independent of each other (Bakker, Schumacher, & Rooduijn, 2021). Yet, if the cognitive self-reported measure of valence (arousal) is strongly correlated with its physiological counterpart, then it might provide an argument to move away from using fEMG to measure valence, as it is a more costly and time-consuming task.

Starting with the correlation between the average physiological arousal, as measured through nSCL activity (mean per clip) and self-reported arousal (in response to the clip), it is practically absent and not statistically significant, $\rho(381) = 0.03$, $p = 0.482$. Thus, this suggests that these measurements capture different aspects of the affective experience. When we correlate the average fEMG per story with the perceived valence of the story, the correlation is statistically significant but not very strong, $\rho(357) = 0.35$, $p < 0.001$.

Our take-home message from this correlation is that fEMG and self-reported valence pick up different aspects of the affective experience but the results are pointing in the same direction. At the same time, we acknowledge that this test is not perfect: Physiological measures are measured continuously while the self-reports are only measured after each clip. In an ideal world, our self-reported measures of valence and arousal would have been collected at a higher interval level but (a) this would deviate substantially from SFN's design and (b) we did not think about this option in the design stage.

4. Discussion

This study replicates and extends the foundational work by SFN in four ways. First, negative news seems arousing, as captured in changes using SCLs. The effect seems to fall within the bounds of SFN but it remains to be seen if it is a substantively meaningful effect as it might not differ from 0. Second, we extend SFN by showing that negative news is causing negative affect as captured with corrugator activity and self-reported negative affect. Third, we find no evidence that negative news causes increases in self-reported arousal; in fact, we find that positive news is more arousing. Fourth, our replication adds the 18th country (the Netherlands) in which the negativity bias is studied. In what follows, we discuss the implications of our findings for political communication research on the negativity bias and reflect on the importance of replication studies in empirical (political) communication research.

We acknowledge that our result for SCL is in the expected positive direction but not statistically significant, a lot smaller than the estimate of SFN and depending on the chosen bounds of the equivalent test, suggests our result may or may not be negligible. For some readers, this might be a reason to interpret our findings as being a "failed replication" of SFN. But, following Gelman and Stern (2006), we also assessed whether our results differ from those of SFN. Exploratory tests hint at the fact that our result is not different from SFN's pooled results across 17 countries. Moreover, close inspection of the results per country in SFN's study also demonstrates that, in nine out of the 17 countries, the effect was statistically significant, while this was not the case in the other eight countries. To summarise, our results for SCL should be seen as in line with SFN's results. Pooling our results with those of SFN, we got one step closer to the population-based effect size of negative news on physiological arousal as captured with skin conductance. It is a small but positive effect.

Our work has implications for the negativity bias argument in political communication: We need to move away from the focus on (physiological) arousal. SFN limited their theorising and tests of the negativity bias to one dimension of affect, arousal measured with skin conductance. Yet, affect consists of at least two (some argue three or more) dimensions (Russell, 1980; Schiller et al., 2022). We extended SFN's argument and turned to valence as the second dimension of affect. Building upon SFN's logic, we hypothesised and confirmed that negative news increases negative valence as captured with the activity of the corrugator muscle. Our work thereby refines SFN's theory and shows that negative news causes negative affect. Thereby, we align the negativity bias theory better with developments in psychology and neuroscience that discuss the importance of studying multiple dimensions of affect (Schiller et al., 2022).

Our work also has important societal implications. The prolonged experience of negative affect in response to news could negatively affect both mental (Ford & Feinberg, 2020) and political health (Smith et al., 2023). Moreover, negative affect could cause news fatigue and ultimately the avoidance of news altogether (de Bruin

et al., 2021), which could have detrimental negative consequences for democracy (Blekesaune et al., 2012). Therefore, it is important to understand the affective responses people experience in response to the news.

We identify multiple options for future research. First, the features of the study design could be improved. SFN manipulated the valence of the stimuli but not arousal. As a reviewer pointed out, it might be worth it to manipulate arousal *independent* of valence to see if more arousing content indeed causes arousal. This would help to determine the exact boundaries of the negativity bias argument. Related to this, future research might also want to establish the equivalence of different treatments that are negative (and positive) to avoid stimuli-specific confounds driving the effects. A more dynamic modelling approach with attention to events in the treatment could be useful (e.g., Schumacher et al., 2022), but, ideally, this equivalence is achieved in the design stage by careful pilot testing and validation of materials. Second, we have studied the effects of negativity bias, while negativity bias might actually be the cause of political information consumption. Stylised experiments outside of the domain of political communication (New et al., 2007; Nissens et al., 2017) could lead to the hypothesis that negative news, just like moral content (Gantman & Van Bavel, 2014), captures more attention. Designs that allow participants to select in (or out) of negative (vs. positive) news might show that some people are more affected by negative news than others (for inspiration of such designs, see Arceneaux et al., 2013). Third, scholars could take inspiration from work in constructive and solution-based journalism, which have been proposed as ways of covering both negative and positive valenced news in a "better" way. We could envision future research that manipulates these and other forms of journalism to see if negative affect (and positive affect) are stronger or weaker depending on the way negative (or positive) news is covered.

Our study, like any other, comes with some limitations. First, our study was a pretty close replication of SFN, but we used, for instance, different equipment, slightly different procedures, and added a new context. Do these differences affect the results? Here we reflect on some of these concerns. First, does it matter that we conducted our study in the Netherlands? No, this is unlikely. SFN found little indication that country characteristics explained variation in the negativity bias, thereby addressing concerns that the negativity bias is conditional upon context (e.g., Van Bavel et al., 2016). Second, does the specific place where we conducted the study matter? No, this is unlikely. Our lab has, like any lab, a unique physical makeup. A recent evaluation of data collected at different locations (a university lab and different lab-in-the-field settings) showed that the results were robust across contexts (Schumacher et al., 2024). Third, does the equipment matter? No, this is unlikely. We used different equipment than SFN. At the same time, we think it is unlikely that the equipment would cause any differences in the results. Our equipment has been used in the past, by other teams, to successfully capture skin conductance and fEMG activity (Gazendam et al., 2013; Nohlen et al., 2016; Sevenster et al., 2015). Fourth, does the type of physiological measures used affect the results? No, this is unlikely. SFN measured SCL and heart rate variability which are relatively unobtrusive measures. We turned to corrugator activity and had to connect wires to a participant's face. We think it is unlikely that our SCL results are affected by the use of the corrugator measure. Moreover, we think our corrugator measure is valid as we (a) do not tell people about the purpose of the measure and (b) in other studies we have successfully shown to measure corrugator activity in response to political and non-political stimuli (e.g., Bakker, Schumacher, & Rooduijn, 2021; Homan et al., 2023). Last, could the effects of the stimuli have become weaker over time? No, this is unlikely. The stimuli from SFN are a few years old; Soroka (personal communication, October 30, 2023) indicated that they are from the period 2015–2016. One could argue that the stimuli became less powerful because, for instance, journalistic reporting changes over time and the world might have become more negative. We think this is unlikely as our results for

valence are consistent across physiological and self-reported measures. A second concern regarding the selection of stimuli is that we used only two negative and two positive clips. As discussed, we did this because our omnibus study had limited space for this replication project. Yet, if the results would hinge on the exact number of stimuli, then the negativity bias is not a robust effect. Therefore, we think this is unlikely to bias our results. To summarise, we disregard these limitations. At the same time, we are the first to acknowledge that similar procedures and equipment are desired when doing close replications. Yet, it is important to remember that if the negativity bias is a robust finding then the small differences between study designs discussed here should not affect the results.

Where should we go from here when it comes to doing replication studies in (political) communication research? We have four concrete recommendations. Recommendation 1: Following Nuijten et al. (2018), we reanalysed and reproduced SFN's results at the start of the study. It allowed us to get the effect size (standardised SCL effect) and showed that the results were robust to the modelling specification we proposed. Recommendation 2: Following King (1995), we directly replicated SFN's central finding and extended it. In doing so, we assess the replicability of the finding (Freese & Peterson, 2017) and generate new knowledge (i.e., negative news also causes negative affect). Recommendation 3: To verify the results (see recommendation 1) and directly replicate SFN (see recommendation 2), we needed open data and materials. We were pleased that SFN provided their data publicly and were able to share the stimuli with us. This illustrates the point made by Bowman and Spence (2020) about the importance of open data and materials. Recommendation 4: We need academics like Soroka with a welcoming mindset to replication. Soroka has been responsive to our questions throughout the project and, most of all, has encouraged us to pursue this replication project. We recommend scholars take Soroka's mindset as an example of how to "deal" with scholars who want to replicate and extend their work.

For those interested in physiological responses to political communication, we have one specific suggestion: Let us collaborate more. Laboratory studies are time intensive (it takes months to complete a study), they are costly to start (lab equipment, lab space) and they are costly for participants (one participant per hour). As a consequence, samples are often small (too small to be sufficiently powered), and novelty is privileged over replications. Like Chambers (2017), we think replications should, however, be part of the cycle through which we generate knowledge. Therefore, we urge scholars to make replications standard practice. It is tempting, also for us, to continue to function as moles (University Library Vrije Universiteit Amsterdam, 2018): Each mole (a researcher or team of researchers) digs its own tunnel and generates its own body of research. Yet, the tunnels of different moles are not in contact with each other. Thereby we are not accumulating knowledge across research teams and are not getting closer to understanding the robustness and replicability of the phenomena of interest (Chambers, 2017). Instead, Rene Bekkers (University Library Vrije Universiteit Amsterdam, 2018) suggests that scientists could take inspiration from ants: They work together and get a lot done in close cooperation. We encourage multi-laboratory collaborations where scholars free up space in their protocol for others to test their ideas. Scholars with more resources could especially consider providing space in their protocols to junior scholars and minority researchers. This could be considered an "act of solidarity" (Lukito, 2024) by those with more resources. Our lab is open to this opportunity and we encourage interested readers to contact us.

We hope close replications become standard in empirical (political) communication research. Yes, we are aware that replication studies are still relatively uncommon in quantitative communication research (Bakker, Jaidka,

et al., 2021; Keating & Totzkay, 2019; McEwan et al., 2018). At the same time, communication researchers hold positive attitudes towards replication studies and say that replications improve the discipline (Bakker, Jaidka, et al., 2021; Bowman et al., 2022). Thematic issues, like this one, are ways to stimulate replication studies. Yet, real change happens when more people start doing replication studies and when editors, reviewers, grant committees, and hiring committees become more welcoming to them. This requires structural change. At the same time, this does not excuse individual researchers from trying to change the system. Replications only become more common if more people start doing them. Therefore, we hope that our study stimulates others to also engage in replication studies.

Acknowledgments

We thank Stuart Soroka for providing the stimuli and discussions throughout this project. Following the CRedIT taxonomy roles: conceptualisation (BNB), data curation (RD, GS), formal analysis (RD), funding acquisition (GS), project administration (BNB, GS), investigation (MD, DP), methodology (RD, BNB, GS), validation (BNB), writing original draft (RD, BNB), writing (review and editing; RD, GS, MD, DP, BNB), and supervision (BNB, GS).

Funding

This publication is part of the project Under Pressure: How Citizens Respond to Threats and Adopt the Attitudes and Behaviors to Counter Them (Project No. VI.Vidi.211.055, awarded to Bert Bakker) of the research programme NWO Talent Programme VIDI which is financed by the Dutch Research Council and POLEMIC, a project that has received funding from the European Research Council under the European Union's Horizon 2020 research and innovation programme (Grant Agreement No 759079).

Conflict of Interests

The authors declare no conflict of interests.

Data Availability

The data is publicly available online at: https://osf.io/jwszx. The code to reproduce our results can also be found on this OSF page.

References

Arceneaux, K., Bakker, B. N., & Schumacher, G. (in press). Being of one mind: Does alignment in physiological responses and subjective experiences shape political ideology? *Political Psychology*.

Arceneaux, K., Johnson, M., & Cryderman, J. (2013). Communication, persuasion, and the conditioning value of selective exposure: Like minds may unite and divide but they mostly tune out. *Political Communication*, *30*(2), 213–231.

Bakker, B. N., Jaidka, K., Dörr, T., Fasching, N., & Lelkes, Y. (2021). Questionable and open research practices: Attitudes and perceptions among quantitative communication researchers. *Journal of Communication*, *71*(5), 715–738.

Bakker, B. N., Schumacher, G., & Rooduijn, M. (2021). Hot politics? affective responses to political rhetoric. *American Political Science Review*, *115*(1), 150–164.

Barrett, L. F. (2017). *How emotions are made: The secret life of the brain*. Pan Macmillan.

Bhowmik, S., Jelfs, B., Arjunan, S. P., & Kumar, D. K. (2017). Outlier removal in facial surface electromyography through Hampel filtering technique. In C. McGregor & S. Mozar (Eds.), *2017 IEEE Life Sciences Conference (LSC)* (pp. 258–261). IEEE.

Blekesaune, A., Elvestad, E., & Aalberg, T. (2012). Tuning out the world of news and current affairs—An empirical study of Europe's disconnected citizens. *European Sociological Review*, *28*(1), 110–126.

Bowman, N. D., Rinke, E. M., Lee, E.-J., Nabi, R., & de Vreese, C. H. (2022). How communication scholars see open scholarship. *Annals of the International Communication Association*, *46*(3), 205–230.

Bowman, N. D., & Spence, P. R. (2020). Challenges and best practices associated with sharing research materials and research data for communication scholars. *Communication Studies*, *71*(4), 708–716.

Boyer, M. M. (2023). Aroused argumentation: How the news exacerbates motivated reasoning. *The International Journal of Press/Politics*, *28*(1), 92–115.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, *25*(1), 49–59.

Bradley, M. M., & Lang, P. J. (2000). Measuring emotion: Behavior, feeling, and physiology. In R. Lane & L. Nadel (Eds.), *Cognitive neuroscience of emotion* (pp. 242–276). Oxford University Press.

Carlson, T. N., McClean, C. T., & Settle, J. E. (2020). Follow your heart: Could psychophysiology be associated with political discussion network homogeneity? *Political Psychology*, *41*(1), 165–187.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Croissant, Y., & Millo, G. (2008). Panel data econometrics in R: The plm package. *Journal of Statistical Software*, *27*(2), 1–43. https://doi.org/10.18637/jss.v027.i02

Dawson, M. E., Shell, A., & Filion, D. L. (2007). The electrodermal system. In J. T. Cacioppo, L. Tassinary, & G. C. Berntson (Eds.), *Handbook of psychophysiology* (pp. 159–181). Cambridge University Press.

de Bruin, K., de Haan, Y., Vliegenthart, R., Kruikemeier, S., & Boukes, M. (2021). News avoidance during the Covid-19 crisis: Understanding information overload. *Digital Journalism*, *9*(9), 1286–1302.

Evers, C., Hopp, H., Gross, J. J., Fischer, A. H., Manstead, A. S., & Mauss, I. B. (2014). Emotion response coherence: A dual-process perspective. *Biological Psychology*, *98*, 43–49.

Ford, B. Q., & Feinberg, M. (2020). Coping with politics: The benefits and costs of emotion regulation. *Current Opinion in Behavioral Sciences*, *34*, 123–128.

Fridlund, A. J., & Cacioppo, J. T. (1986). Guidelines for human electromyographic research. *Psychophysiology*, *23*(5), 567–589.

Gantman, A. P., & Van Bavel, J. J. (2014). The moral pop-out effect: Enhanced perceptual awareness of morally relevant stimuli. *Cognition*, *132*(1), 22–29.

Gazendam, F. J., Kamphuis, J. H., & Kindt, M. (2013). Deficient safety learning characterizes high trait anxious individuals. *Biological Psychology*, *92*(2), 342–352.

Gelman, A., & Stern, H. (2006). The difference between "significant" and "not significant" is not itself statistically significant. *The American Statistician*, *60*(4), 328–331.

Homan, M. D., Schumacher, G., & Bakker, B. N. (2023). Facing emotional politicians: Do emotional displays of politicians evoke mimicry and emotional contagion? *Emotion*, *23*(6), 1702–1713.

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239.

Keib, K., Espina, C., Lee, Y., Wojdynski, B. W., Choi, D., & Bang, H. (2018). Picture this: The influence of emotionally valenced images, on attention, selection, and sharing of social media news. *Media Psychology*, *21*(2), 202–221.

King, G. (1995). Replication, replication. *PS: Political Science & Politics*, *28*(3), 444–452.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, *8*(1), Article 33267.

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785.

LeDoux, J. E., & Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: A two-system framework. *American Journal of Psychiatry*, *173*(11), 1083–1093.

Lüdecke, D., Ben-Shachar, M., Patil, I., & Makowski, D. (2020). Extracting, computing and exploring the parameters of statistical models using R. *Journal of Open Source Software*, *5*(53), Article 2445. https://doi.org/10.21105/joss.02445

Lukito, J. (2024). Scholarly solidarity: Building an inclusive field for junior and minority researchers. *Political Communication*, *41*(1), 152–161.

McEwan, B., Carpenter, C. J., & Westerman, D. (2018). On replication in communication science. *Communication Studies*, *69*(3), 235–241.

New, J., Cosmides, L., & Tooby, J. (2007). Category-specific attention for animals reflects ancestral priorities, not expertise. *Proceedings of the National Academy of Sciences*, *104*(42), 16598–16603.

Nissens, T., Failing, M., & Theeuwes, J. (2017). People look at the object they fear: Oculomotor capture by stimuli that signal threat. *Cognition and Emotion*, *31*(8), 1707–1714.

Nohlen, H. U., van Harreveld, F., Rotteveel, M., Barends, A. J., & Larsen, J. T. (2016). Affective responses to ambivalence are context-dependent: A facial EMG study on the role of inconsistency and evaluative context in shaping affective responses to ambivalence. *Journal of Experimental Social Psychology*, *65*, 42–51. https://doi.org/10.1016/j.jesp.2016.02.001

Nuijten, M. B., Bakker, M., Maassen, E., & Wicherts, J. (2018). *Verify original results through reanalysis before replicating: A commentary on "Making Replication Mainstream" by Rolf A. Zwaan, Alexander Etz, Richard E. Lucas, & M. Brent Donnellan*. PsyArXiv. https://doi.org/10.31234/osf.io/fuzkh

Ravaja, N., Aula, P., Falco, A., Laaksonen, S., Salminen, M., & Ainamo, A. (2015). Online news and corporate reputation. *Journal of Media Psychology*, *27*(3), 118–133.

R Core Team. (2023). *R: A language and environment for statistical computing* [Computer software]. R Foundation for Statistical Computing.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), Article 1161.

Schiller, D., Yu, A. N.C., Alia-Klein, N., Becker, S., Cromwell, H. C., Dolcos, F., Eslinger, P. J., Frewen, P., Kemp, A. H., Pace-Schott, E. F., Raber, J., Silton, R. L., Stefanova, E., Williams, J. H. G., Abe, N., Aghajani, M., Albrecht, F., Alexander, R., Anders, S., . . . Rizzo, A. (2022). *The human affectome*. *Neuroscience & Biobehavioural Reviews*, *158*, Article 105450. https://doi.org/10.1016/j.neubiorev.2023.105450

Schumacher, G., Homan, M. D., Rebasso, I., Fasching, N., Bakker, B. N., & Rooduijn, M. (2024). Establishing the validity and robustness of facial electromyography measures for political science. *Politics and the Life Sciences*. Advance online publication. https://doi.org/10.1017/pls.2023.26

Schumacher, G., Rooduijn, M., & Bakker, B. N. (2022). Hot populism? Affective responses to antiestablishment rhetoric. *Political Psychology*, *43*(5), 851–871.

Settle, J. E., Hibbing, M. V., Anspach, N. M., Carlson, T. N., Coe, C. M., Hernandez, E., Peterson, J., Stuart, J., & Arceneaux, K. (2020). Political psychophysiology: A primer for interested researchers and consumers. *Politics and the Life Sciences*, *39*(1), 101–117. https://doi.org/10.1017/pls.2020.5

Sevenster, D., Hamm, A., Beckers, T., & Kindt, M. (2015). Heart rate pattern and resting heart rate variability mediate individual differences in contextual anxiety and conditioned responses. *International Journal of Psychophysiology*, *98*(3), 567–576.

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science*, *26*(5), 559–569.

Smith, K., Weinschenk, A., & Panagopoulos, C. (2023). On pins and needles: Anxiety, politics and the 2020 US presidential election. *Journal of Elections, Public Opinion and Parties*. Advance online publication. https://doi.org/10.1080/17457289.2023.2189258

Soroka, S., Fournier, P., & Nir, L. (2019). Cross-national evidence of a negativity bias in psychophysiological reactions to news. *Proceedings of the National Academy of Sciences*, *116*(38), 18888–18892.

Soroka, S., & McAdams, S. (2015). News, politics, and negativity. *Political Communication*, *32*(1), 1–22.

University Library Vrije Universiteit Amsterdam. (Host). (2018, November 1). Fair-data and the end of competitive science [Audio podcast episode]. In *VU Library Live*. SoundCloud. https://soundcloud.com/vu-library-live/ep2-fair-data-and-the-end-of-competative-science

Van Bavel, J. J., Mende-Siedlecki, P., Brady, W. J., & Reinero, D. A. (2016). Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences*, *113*(23), 6454–6459.

van Boxtel, A. (2010). Facial EMG as tool for inferring affective states. In A. Spink, F. Grieco, O. Krips, L. Loijens, L. Noldus, & P. Zimmerman (Eds.), *Proceedings of Measuring Behavior 2010* (pp. 104–108). Noldus Information Technology.

Wang, Z., Morey, A. C., & Srivastava, J. (2014). Motivated selective attention during political ad processing: The dynamic interplay between emotional ad content and candidate evaluation. *Communication Research*, *41*(1), 119–156.

About the Authors

Roeland Dubèl is a PhD candidate at the Amsterdam School of Communication Research, University of Amsterdam. His research primarily focuses on how journalists and citizens deal with and understand the issue of trustworthiness, as well as the impact of strategies aiming to increase media trust. Furthermore, he studies the effects of news consumption and the workings of news consumption behaviour patterns.

Gijs Schumacher is an associate professor of political science at the University of Amsterdam. He is the co-director of the Hot Politics Lab. He has published extensively in the fields of political psychology and comparative politics. His work has appeared in, for example, *Nature Human Behaviour*, *American Political Science Review*, *American Journal of Political Science*, *Political Communication*, *Political Psychology*, and *Emotion*.

Maaike D. Homan is a postdoctoral researcher in the Organisational Behaviour Group, at the Department of Social, Health and Organisational Psychology at Utrecht University. Her research focuses on the role of emotions in society and politics. More specifically, she examines the interplay between individuals' physiological changes in their bodies and their subjective feelings and attitudes in response to societal challenges, political issues, and politicians.

Delaney Peterson is a PhD candidate with the Amsterdam School of Communication Research and the Hot Politics Lab. Her research focuses on understanding how societal threat perceptions are formed and how they influence political behaviour. Specifically, Delaney investigates how one's psychology and physiology are implicated in the political trends of the 21st century.

Bert N. Bakker is an associate professor at the Amsterdam School of Communication Research, University of Amsterdam. He co-directs the Hot Politics Lab and studies how people form their beliefs about politics using insights from psychology, communication science, and political science. His work appeared in leading journals across different disciplines such as the *Journal of Communication*, *Nature Human Behaviour*, *American Political Science Review*, the *Journal of Politics*, and *Emotion*. Currently, his work is supported by the Dutch Science Foundation and the European Commission.

Media and Communication
2024 • Volume 12 • Article 7851
https://doi.org/10.17645/mac.7851

cogitatio

ARTICLE

Open Access Journal

Audio-as-Data Tools: Replicating Computational Data Processing

Josephine Lukito [1], Jason Greenfield [2], Yunkang Yang [3], Ross Dahlke [4], Megan A. Brown [5], Rebecca Lewis [4], and Bin Chen [1,6]

[1] School of Journalism and Media, University of Texas at Austin, USA
[2] Center for Social Media and Politics, New York University, USA
[3] Department of Communication & Journalism, Texas A&M University, USA
[4] Department of Communication, Stanford University, USA
[5] School of Information, University of Michigan, USA
[6] Journalism and Media Studies Centre, University of Hong Kong

Correspondence: Josephine Lukito (jlukito@utexas.edu)

Submitted: 16 November 2023 **Accepted:** 22 February 2024 **Published:** 6 May 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

The rise of audio-as-data in social science research accentuates a fundamental challenge: establishing reproducible and reliable methodologies to guide this emerging area of study. In this study, we focus on the reproducibility of audio-as-data preparation methods in computational communication research and evaluate the accuracy of popular audio-as-data tools. We analyze automated transcription and computational phonology tools applied to 200 episodes of conservative talk shows hosted by Rush Limbaugh and Alex Jones. Our findings reveal that the tools we tested are highly accurate. However, despite different transcription and audio signal processing tools yield similar results, subtle yet significant variations could impact the findings' reproducibility. Specifically, we find that discrepancies in automated transcriptions and auditory features such as pitch and intensity underscore the need for meticulous reproduction of data preparation procedures. These insights into the variability introduced by different tools stress the importance of detailed methodological reporting and consistent processing techniques to ensure the replicability of research outcomes. Our study contributes to the broader discourse on replicability and reproducibility by highlighting the nuances of audio data preparation and advocating for more transparent and standardized practices in this area.

Keywords

audio-as-data; computational methods; conservative talk shows; data processing; reproduction; talk radio

1. Audio-as-Data Tools: Reproducing Computational Data Processing

Like other disciplines, communication and media researchers have increasingly been concerned with the replicability of the field's research (Benoit & Holbert, 2008; McEwan et al., 2018). Replicability and reproducibility are critical for research: Without them, it is unclear whether a particular finding is a consequence of nuanced research decisions or an actual finding. However, methodological obfuscation of data collection, preparation, and analysis (intentionally or otherwise) continues to plague replication and reproduction efforts. Open science efforts—particularly those tailored to our field—provide new avenues for producing more empirically grounded research (Dienlin et al., 2020), but blind spots remain.

One blind spot of interest to us, specifically in computational communication research, is data preparation: the steps for cleaning and wrangling the data for analysis. While challenges to reproducing data collection methods and sharing data persist (for more, see Van Atteveldt et al., 2019), data preparation is often glossed over or subsumed as part of the data analysis process, particularly if researchers are relying on computers (Plesser, 2018). However, replicating and reproducing data preparation processes is critical as the results of two studies may vary because of data cleaning, even when holding the analysis or collection strategies constant.

This study explores reproducing data preparation practices more concretely by focusing on audio-as-data tools for data preparation. We chose this type of data for three reasons. First, there is increasing scholarly interest in audio data (likely driven by the growing popularity of digital audiovisual content), particularly spoken language. Second, researchers often transform audio-as-data into other data forms as a part of the data preparation process, such as when researchers turn spoken language into a transcription for text-as-data approaches. Third, there has yet to be an empirical study that compares whether different audio-as-data processing tools produce similar results.

To assess whether results from audio-as-data processing tools will reproduce, we consider two important data preparation practices for audio data: automated transcription and audio signal processing for computational phonology. Using two datasets of right-wing talk shows, we compare the results of four tools, two for transcription and two for detecting pitch and intensity. Our results find that, while these tools produce similar results, the nuances of each tool produce subtle differences that highlight why replication and reproduction studies must account for variations in data preparation.

2. Literature Review

2.1. Conceptualizing Audio-as-Data

We define "audio-as-data" as approaches for computationally processing and analyzing auditory communication (including, but not limited to, music, speech, and noise) to address important questions in political and social life. The field of communication is currently dominated by the "text-as-data" approach, often because of convenience and size (Lukito, Brown, et al., 2023). Studies involving multimedia content such as radio, television, or podcasts often deal with audio data by textual transcription, during which the auditory features (e.g., pitch, loudness, tone) are lost (Dietrich et al., 2019; Knox & Lucas, 2021).

Despite audio data becoming more prominent in digital spaces, there are relatively few audio-as-data studies (Piñeiro-Otero & Pedrero-Esteban, 2022). It is not just the content of speech but its delivery that informs us; auditory cues in political speech, for instance, can subtly yet significantly indicate emotion and stance and change opinions (Dietrich et al., 2019; Klofstad et al., 2012; Knox & Lucas, 2021). It is worth noting that none of the above three studies used any additional software to validate the output of audio features produced by one software: Dietrich et al. (2019) and Klofstad et al. (2012) both used praat only to measure pitch (Jadoul et al., 2018), whereas Knox and Lucas (2021) used the "communication" R library to produce audio features (Lucas, 2022).

Spoken language, as a type of audio-as-data, can processed with two approaches: a reductive approach and an additive approach (Lukito, 2023). Reductive approaches remove information from a dataset because that information is irrelevant to a particular project or research question. In the context of audio data, the most common reductive approach transforms audio data into text, removing auditory cues and facilitating text-as-data approaches through automated transcription. Automated transcriptions have the advantage of speed and scale: Automated transcriptions take a fraction of the time compared to manual transcription. One common concern is accuracy: Older automated transcription tools often made mistakes, and manual tools were necessary to correct automated transcription (e.g., Luz et al., 2008). However, the ubiquity of digital video content has motivated demand for automated transcription, resulting in considerable improvements over the last decade, both in terms of accuracy (Bokhove & Downey, 2018) and in the variety of languages considered (Wisniewski et al., 2020).

Whereas reductive approaches remove unnecessary (in the context of a study) information, additive approaches involve highlighting or annotating information so that these features can be more easily studied. Additive approaches enrich our understanding by highlighting auditory features, employing audio signal processing for both speech and music to identify characteristics like pitch and intensity, two of the most popular auditory features studied (Gold et al., 2011; Purwins et al., 2019). Pitch refers to the frequency of the wavelength of a sound. Higher-pitched sounds tend to be shriller, with more frequent oscillations. By contrast, lower-pitched sounds are deeper. To use a musical example, sopranos are higher-pitched, and baritones are lower-pitched. Pitch algorithms have advanced the study of accents and vocal nuances like sarcasm (Iosad, 2015; Larrouy-Maestri et al., 2023). Previous studies have found that pitch can impact a political candidate's electability; Klofstad (2016), for example, found that candidates perceived to have a lower pitch generally received more votes than their higher-pitched opponents, though this may vary by gender.

Whereas pitch is described concerning highness and lowness, intensity refers to the number of sound waves passing through an area per second. Intensity and loudness are related, as a more intense sound will be perceived as louder by the human ear (for this reason, intensity and loudness measures are often highly correlated; this was also the case for our study). Intensity measures have been applied to emotional analysis and acoustic engineering (Chen et al., 2012; Indrayani et al., 2020; Larsen & Aarts, 2005). In other auditory analyses, auditory attributes such as duration (how long an auditory note is held) and timbre (distinctions between two instruments or two voices) are also considered. However, these are studied more regarding music (e.g., Krumhansl & Iverson, 1992) rather than spoken language (e.g., Dietrich et al., 2019).

Despite their growing use in disciplines such as computer engineering and linguistics, these tools have comparatively few applications in media and communication research (for an exception, see Shah et al.,

2023). This scholarship may be scant because of perceived applicability, as few studies have shown how researchers can leverage these tools to study media. However, we hope these audio-as-data methods become more accessible to our field. As communication and media scholars have long studied audio media (e.g., Christenson & Lindlof, 1983; Spinelli & Dann, 2019), it stands to reason that adopting these methods will become more widespread in the literature. If so, communication researchers must compare these tools to understand how computational phonology and audio signal processing tools should be applied.

2.2. Replicable and Reproducible Audio Data Preparation

Over the past few decades, social scientists have raised concerns about the replicability and reproducibility of research. While related, these two terms are conceptually distinct (National Academies of Sciences, Engineering, and Medicine, 2019), so it is important to define this terminology. The National Academies of Sciences, Engineering, and Medicine (2019) defines reproducibility as using the same input data and data processing and producing similar results—They describe this as "computational reproducibility" (p. 6). In contrast, replication refers to finding the same results from two separate data collections. Other definitions are more ambiguous; for example, Nosek and Errington (2020) define replication as "a study for which any outcome would be considered diagnostic evidence about a claim from prior research" (p. 2). Adding further confusion, other researchers have presented conflicting definitions; for example, Plesser (2018) defined replicability as when two different teams apply the same research design and produce similar results, whereas reproducibility refers to different research teams applying different research designs yet producing similar results (Plesser also defines "repeatability" as being able to repeat one's own research and produce similar results).

One gap in this literature is the reductive treatment of the research design, which often includes multiple steps such as data collection, data processing, and data analysis; however, collection and analysis are often emphasized compared to data processing. For example, Plesser's distinction between research repeated by the same team versus research replicated or reproduced by a different team emphasizes differences in data collection (e.g., location of the research team). Similarly, many definitions of replicability and reproducibility focus on the outcome of the research (Howell, 2020; National Academies of Sciences, Engineering, and Medicine, 2019; Nosek & Errington, 2020). While it is important to define these concepts, focusing solely on definition misses a key issue with research on replicability and reproducibility: What part of any given research process becomes unreplicable or unreproducible? Furthermore, what does replicable and reproducible mean in data preparation? Suppose a team conducts a "reproduction study," but uses different software to conduct optical character recognition, are changes in the results of substantive changes in the context or case studied, or a result of the different software used to process the data?

In the context of computational data processing, we define "reproduced processing" as using the same data and the same methods to produce the same data outcomes, despite using different software to conduct the methods. This definition draws conceptually from the definition of reproduction as using the same data and methods and recognizes software as a research tool that should (but may not) produce similar results.

Assessing the reproducibility of transformations to data is not new. There is a large body of research concerning the sensitivity and consistency of text analyses based on choices in preprocessing (Hegazi et al., 2021; Naseem et al., 2021; Tabassum & Patil, 2020). Different choices in text preprocessing have even been

shown to change the performance of downstream models and analyses (Alakrot et al., 2018; Juneja & Das, 2019). For example, Denny and Spirling (2018) show how different choices in pre-processing data for unsupervised learning (such as punctuation removal, lowercasing, stemming, and stopword removal) could produce different latent Dirichlet allocation topic modeling results such that important keywords were only present in the top-20 important terms for a topic for some pre-processed data and not others.

In audio data, the proliferation of methods and tools for analysis poses the same challenges as text. More specifically, there is a need for methods to validate audio analyses. This manuscript proposes a method for validating audio-as-data processing tools, a step in the audio data analysis pipeline.

2.3. Conservative Talk Radio

We apply these strategies to conservative talk radio. Though talk radio has long existed as a form of mediated communication (Armstrong & Rubin, 1989; Avery et al., 1978), conservatives within the United States have leveraged this media format to gain popularity and motivate political action (Matzko, 2020; Young, 2020). However, these tools apply to other audio data, particularly other forms of spoken language (e.g., podcasts, interviews, interpersonal communication, voicemails, and digital videos).

Talk radio combines elements of "shock jock" broadcast entertainment with the promotion of conservative viewpoints, usually hosted by an individual charismatic host. By forming talking points reinforced and amplified within conservative newspapers and on right-wing television, talk radio hosts effectively attack liberal ideas and defend conservatism to their audiences, ultimately driving the rise in outrage as a political media style (Berry & Sobieraj, 2013). Talk radio, as the name suggests, predominantly consists of spoken language by one or several speakers, but it will also include music without words, particularly as brief transitions between sections or as the background of advertising content and introductions. Many talk radio shows, including those studied here, are several hours long and are broadcast daily.

While there have been many popular conservative talk radio hosts, two epitomize the medium: Rush Limbaugh and Alex Jones. Limbaugh, dominant in the 1990s, influenced public opinion (Barker, 1998; Hall & Cappella, 2002; Jamieson et al., 1998; Lee & Cappella, 2001) and conservative rhetoric (Harris et al., 1996). He paved the way for contemporary right-wing media figures, including Glenn Beck and Sean Hannity, many of whom have evolved beyond talk radio into distributing their audio content through podcasting (e.g., Dowling et al., 2022) or YouTube (e.g., Wurst, 2022). Conversely, Jones, known for hosting The Alex Jones Show and Infowars (https://www.infowars.com) since 1999, shaped the conspiratorial wing of modern conservatism in the United States (Beauchamp, 2016), using his platform to propagate conspiracy theories, leading to significant legal and social repercussions (Slater, 2022). His "dangerous demagoguery" (Mercieca, 2019) fosters a unique bond with his audience, affecting trust and media interaction (Madison et al., 2019, 2020), thereby contributing to a fragmented reality among his listeners (Dunne-Howrie, 2019).

We use these two conservative talk show hosts as cases because of their social and political relevance and substantial amount of content: three hours daily. They provide vast and rich corpora of speech-language, making them ideal for comparing and validating audio-as-data tools. Another advantage of talk radio is that it is professionally produced, making it similar to traditional radio, podcasts, and broadcast television content. Audio data produced professionally tends to have greater clarity, as the speakers use professional equipment

and background noise may be minimized. In contrast, both reductive and additive processing may be more difficult in informal recordings with significant background noise. Thus, we expect these results to be most relevant to professionally produced audio media with limited music.

Based on the above literature, we propose the following two research questions:

> RQ1: When processing talk radio audio data into transcripts, to what extent will two different tools for automated transcription reproduce similar data?
>
> RQ2: When processing talk radio audio data to annotate audio features, to what extent will two different tools for audio feature annotation reproduce similar data?

3. Methods

3.1. Data Collection

For our analysis of audio data, we collected data from two right-wing media figures: Alex Jones and Rush Limbaugh. First, the primary data source for our audio recordings of The Alex Jones Show was the official Infowars website's RSS section, specifically titled "Infowars Audio/Video Resource Links." This digital archive, accessible to the public, offers a vast collection of episodes spanning multiple years. In this article, we downloaded all shows from January 1, 2016 to December 31, 2018. We select this time period as Jones received significant and negative attention for his talk shows during this time (in late 2018, Jones was sued by several parents of victims of the Sandy Hook shooting, leading to greater scrutiny of his show; see Williamson, 2022). Each episode within this period adheres to a distinct and consistent URL format, integrating the broadcast date and the day of the week. We employed a custom-built Python tool built on the urllib Python library to ensure a systematic and efficient data retrieval process. This software was designed to: (a) navigate sequentially through the dates from the start of 2016 to the end of 2018, (b) dynamically generate the appropriate URL for each episode based on the date, and (c) initiate the download of the associated audio file, ensuring data integrity and completeness. We sampled 100 episodes from this three-year period to use in the analysis.

Next, we collected audio data for The Rush Limbaugh Show. We used Python code and scraped the audio files hosted on the website (https://www.rushlimbaugh.com) in October 2022. Due to data availability, we could download all Rush Limbaugh shows between January 1, 2020, and June 30, 2021. We randomly sampled 50 shows from 2020 and 50 from 2021 for The Rush Limbaugh Show.

3.2. Reductive: Transforming Audio Data to Text Data

After downloading the audio transcripts, we performed automated speech-to-text using two tools: Google Cloud Platform (GCP) Speech-to-Text and OpenAI's Whisper speech-to-text processing. First, we used Google Speech-to-Text, a leading automatic speech recognition tool (Shakhovska et al., 2019), to generate transcripts of the audio recordings. After experimenting with various models, we settled on Google's "latest_long" with default parameters. To identify different voices within the recordings, we incorporated a diarization configuration, setting the speaker count to range between 2 and 10. The transcription process was automated using a Python script. This approach also significantly reduced the overall processing time.

Compared to GCP Speech-to-Text, OpenAI Whisper is newer. However, it has already been applied to study a variety of communications, including political discourse (Bianchini et al., 2023) and social media audio (e.g., Sihag et al., 2023). Whisper is an "automated speech recognition" tool for multiple languages (Radford et al., 2023). Like other OpenAI tools, Whisper leverages large neural networks—in this case, for conducting automated transcription. The Whisper package in Python contains five models varying in size (and, by extension, speediness and processing requirements). We use the base, English-only model as we anticipate that neither Limbaugh nor Jones would have much non-English in their shows. Open AI's Whisper transcriber is imperfect, but it has achieved significant milestones to make it more applicable to human subject research. We keep these scales at cost by minimizing machine translation for contextualization; however, we acknowledge the process is also subjective.

Owing to cost (while Open AI's Whisper is free, the cost of GCP's Speech-to-Text was about US$50 per transcript, with an overall transcription cost of US$4,106), we only conducted this comparison on Alex Jones for this analysis. In addition to these two specific tools, there are other popular speech-to-text processors that we did not consider due to costs, including Amazon Transcribe (which is known to suffer from long processing times), Microsoft Azure Cognitive Services (which is the costliest for very long recordings, such as talk shows), and IBM Watson Speech to Text (which is more sensitive to more noisy data). While our analysis is exclusively focused on the spoken English language, it is worth noting that of these tools, GCP Speech-to-Text can support a greater variety of languages.

3.2.1. Tool Comparison

To compare the speech-to-text results of GCP's Speech-to-Text with Open AI's Whisper, we compared the transcripts using word error rates (WERs; Klakow & Peters, 2002). WER is a metric that quantifies the difference between the two documents—in this case, the two transcriptions. It measures the percentage of words in the reference transcription incorrectly recognized or omitted in the system's output. Importantly, this is not a measure of understanding (Wang et al., 2003) but a measure of similarity; that is, to what extent are the data similar when the procedure is reproduced across these two tools? We calculated a WER for each episode using the {wersim} package (Proksch et al., 2019). A lower WER is typically better, whereas a higher WER suggests differences in the two transcripts. Previous work has considered WERs of 0.70 to be different (e.g., Jeanrenaud et al., 1995), and more recent studies have considered WERs of 0.40 to be "high error rates" (Morchid et al., 2016, p. 76). This analysis aims to compare the results of Whisper and GCP Speech-to-Text.

An important caveat here involves the data processing for Whisper: WhisperAI has a maximum size of 25 MB. As most of Alex Jones's shows are long (most MP3 recordings were about 55 MB), it was necessary to split the data into three files before proceeding with the analysis. Once these were transcribed, we then combined the three files.

3.3. Additive: Annotating Audio Features

We additionally compare and validate tools that enhance and annotate audio data with information about pitch and intensity, two important metrics in signal processing and auditory analysis (Samrose & Hoque, 2021). To make this comparison, we focus on two packages: parselmouth (Jadoul et al., 2018) and librosa (McFee et al., 2015).

Parselmouth is a Python interface for praat, one of phonology's most popular computer software tools (Kinoshita, 2015; Loakes & Keith, 2013). Praat—and by extension, parselmouth—detects various audio signals, including pitch, tonal intensity, loudness, formants, pacing, and timbre. This tool makes studying accents, prosody (the pacing of someone's speech), and other spoken language phenomena especially useful. For example, Lukito, Gursky, et al. (2023) used praat to study in-person rhetoric at a far-right QAnon event.

Librosa is a tool used for auditory signal processing. Whereas praat and parselmouth are common in linguistics, librosa appears to be more popular in computer engineering and signal processing research to detect sociolinguistic features such as emotion (Babu et al., 2021) and sarcasm detectors (e.g., Tomar et al., 2023).

While some other tools can conduct audio analysis, these two were selected because they specialize in analysis. Other relevant pages include Pysptk, which is also used for speech synthesizing (the artificial production of spoken language) and TorchAudio, which is built on top of PyTorch and, while useful, is more geared towards preparing audio data for machine learning rather than extracting pitch and intensity accurately.

We focus on intensity and pitch detection for this analysis because both are available in parselmouth and librosa, and because of their popularity in signal processing and auditory research. Intensity is most commonly measured using the root mean square amplitude of a sound wave (parselmouth's function is "intensity" and librosa's is "rms," which stands for root mean square). The larger the root mean square, the more intense (or louder) the sound.

Compared to intensity, pitch detection is more challenging. In signal processing and computational phonology, there is no definitive algorithm or way to calculate pitch (Verteletskaya et al., 2009). Correspondingly, these two packages take two different approaches: In parselmouth, pitch is derived from the approximate lowest frequency of the waveform (also known as the fundamental frequency), whereas in librosa, pitch is derived from the melspectogram, an approximation of the waveform's amplitude and frequency over time. In both cases, a higher score constitutes a higher pitch.

3.3.1. Tool Comparison

To compare parselmouth and librosa, we use both to detect the same auditory features (i.e., intensity and pitch) and compare the features extracted with each. For this analysis, we ran the Alex Jones and Rush Limbaugh samples through both. Our goal is to compare the results of parselmouth and librosa for the same transcripts. Notably, both parselmouth and librosa's outputs used different levels of time aggregation, resulting in different numbers of time points. More specifically, parselmouth's analysis was often more granular than librosa's, resulting in typically two to three more time points. Additionally, to make matters more complicated, parselmouth relied on different levels of aggregation for both its mel-spectrometer (for pitch detection) and its intensity detection.

Because these strategies resulted in different temporal aggregation levels, we apply dynamic time warping (DTW) to compare the librosa and parselmouth processing results. DTW is a common approach to compare two-time series that do not perfectly synchronize or vary in speed (Berndt & Clifford, 1994). DTW models leverage non-linear mapping to identify a minimized distance between two or more time series. In DTW,

distance is measured on a scale of 0 to 1, with 0 being *perfect alignment* (no distance, or difference, between the time series) and 1 being *no alignment* (i.e., substantive differences between the time series).

To prepare the data for DTW, we first normalized the time series data (Shao, 2015) to ensure appropriate comparisons of scale between the two-time series. We then conduct four DTW comparisons using the R package {IncDTW} (Leodolter et al., 2021), which provides a vector-based approach to DTW and greater computational efficiency. This step is important for longer-form content, such as podcasts, which create long and highly granular time series data (compared to short online videos and clips). The four DTW comparisons are: (a) a comparison of intensity in Alex Jones's Infowars recordings, (b) a comparison of pitch in Alex Jones's Infowars recordings, (c) a comparison of intensity in Rush Limbaugh's recordings, and (d) a comparison of pitch in Rush Limbaugh's recordings. Below, we present the results for the intensity comparisons and then the pitch comparisons, which are measured as the average distance between each of the point pairs in the two-time series.

4. Results

Our collection consists of 100-episode samples from two sources: Alex Jones's Infowars from 2016 to 2018 and The Rush Limbaugh Show from 2020 to 2021. Both shows are roughly three hours long, totaling 600 hours (about 300 for Jones and Limbaugh each). We begin with comparing the speech-to-text tools for Alex Jones (GCP Speech-to-Text and OpenAI's Whisper), using WER to assess differences in the transcripts. We then compare the computational phonology tools (librosa and parselmouth) using DTW to understand whether these tools produced different auditory features.

4.1. Transcription Results

To address RQ1, we compare the results of GCP's Speech-to-Text with OpenAI's Whisper using WERs. For Alex Jones, the average WER across the 100 videos was 0.76 (SD = 0.21), suggesting a substantial difference between the transcripts produced by GCP Speech-to-Text and Whisper (to compare, the WER for two different episodes was 0.95). These results are presented in Table 1.

A qualitative assessment of the recordings suggests several reasons for this. First, the data splitting ultimately impacted the transcription process despite no content removal. Sometimes, Whisper would not transcribe the first few words of each respective section, resulting in some discrepancies.

A second reason may be related to Whisper's default identification of syntax structure (in particular, punctuations and proper nouns). Whereas automated punctuation requires a secondary model (that was not used in this analysis), Whisper's automated transcription provided fairly accurate punctuations and identification of proper nouns, making it more likely to transcribe names of public figures and groups correctly. We illustrate with an example below (Table 2), from an episode in 2016.

Table 1. WER metrics for Alex Jones data.

Metric	*M*	*SD*	Min	Max
WER	0.762	0.214	0.54	0.89

Table 2. Comparison of GCP Speech-to-Text and OpenAI Whisper.

GCP Speech-to-Text	OpenAI Whisper
Leon McAdoo this put together a powerful six-minute report today that was going to be airing on the Nightly News tonight it's still will we're going to be premiering that here but also be on with more tonight than she's going to Flint Michigan and this evening or tomorrow morning soo	*Leanne* McAdoo has put together a powerful six minute report today that was going to be airing on the nightly news tonight. *And* it still will. We're going to be premiering that here, but *we'll* also be on with more tonight. And she's going to Flint, Michigan this evening or tomorrow morning.

It is important to note that while Whisper appeared to have more accurate punctuations and proper noun identification, both were relatively inconsistent when it came to individual words that are also in portmanteaus (e.g., "takeaway" the noun and "take away" the verb and adverb).

A third reason is that both tools transcribed filler words (e.g., "you know," "ok"), interjectory words (e.g., "wow," "yuck"), and repeated words (e.g., "let, let me start..") differently and inconsistently, which was particularly problematic for sections involving two speakers. GCP was more likely to include repeated and interjectory words, including those spoken softly and in the background. However, this was also inconsistent: There were places where Whisper had identified repeated words, but GCP had not.

4.2. Computational Phonology Results

To address RQ2, we now turn to our analysis of the auditory feature annotations from parselmouth and librosa. Given the differences in how these two tools process the audio data, and in alignment with our methods section, we use DTW to compare the output of these two tools.

Generally, we find that librosa and parselmouth have similar detections of intensity. For the Alex Jones data, the average distance between the two-time series, across our 100-podcast sample, was 0.30 (SD = 0.04), suggesting that librosa and parselmouth had similar identifications of intensity, though these were not identical. Similarly, the average distance between the librosa-derived and parselmouth-derived intensity time series for Rush Limbaugh was 0.27 (SD = 0.014), suggesting that these findings are consistent across podcasts. By comparison, the DTW for intensity of two Alex Jones shows (distance = 0.68) and two Rush Limbaugh shows (distance = 0.74) was very high.

The difference between librosa and parselmouth for pitch is similar. In the case of Alex Jones's data, we find that the average distance between librosa and parselmouth's normalized measure of pitch is 0.30 (SD = 0.04). For Rush Limbaugh, the average distance is 0.29 (SD = 0.016). For comparison, we calculated DTW for intensity of two different Alex Jones shows (distance = 0.77) and two different Rush Limbaugh shows (distance = 0.82). These results are presented in Table 3.

As expected for both pitch and intensity, we find that parselmouth's output is more granular than librosa. Figure 1 illustrates this with a 10-second normalized sample comparison of how parselmouth (red) and librosa (blue) operationalized auditory intensity (from the January 19, 2016 Infowars recording). In this figure, "index," the horizontal axis, refers to the length of the time series (parselmouth's output has more time points and will therefore have a longer index). The y-axis refers to the intensity value as measured by parselmouth or librosa.

Table 3. DTW metrics for Alex Jones and Rush Limbaugh data.

Radio host	Feature	*M*	*SD*	Min	Max
Jones	Pitch	0.306	0.04	0.251	0.322
Jones	Intensity	0.308	0.043	0.262	0.347
Limbaugh	Pitch	0.295	0.016	0.279	0.315
Limbaugh	Intensity	0.271	0.014	0.255	0.289

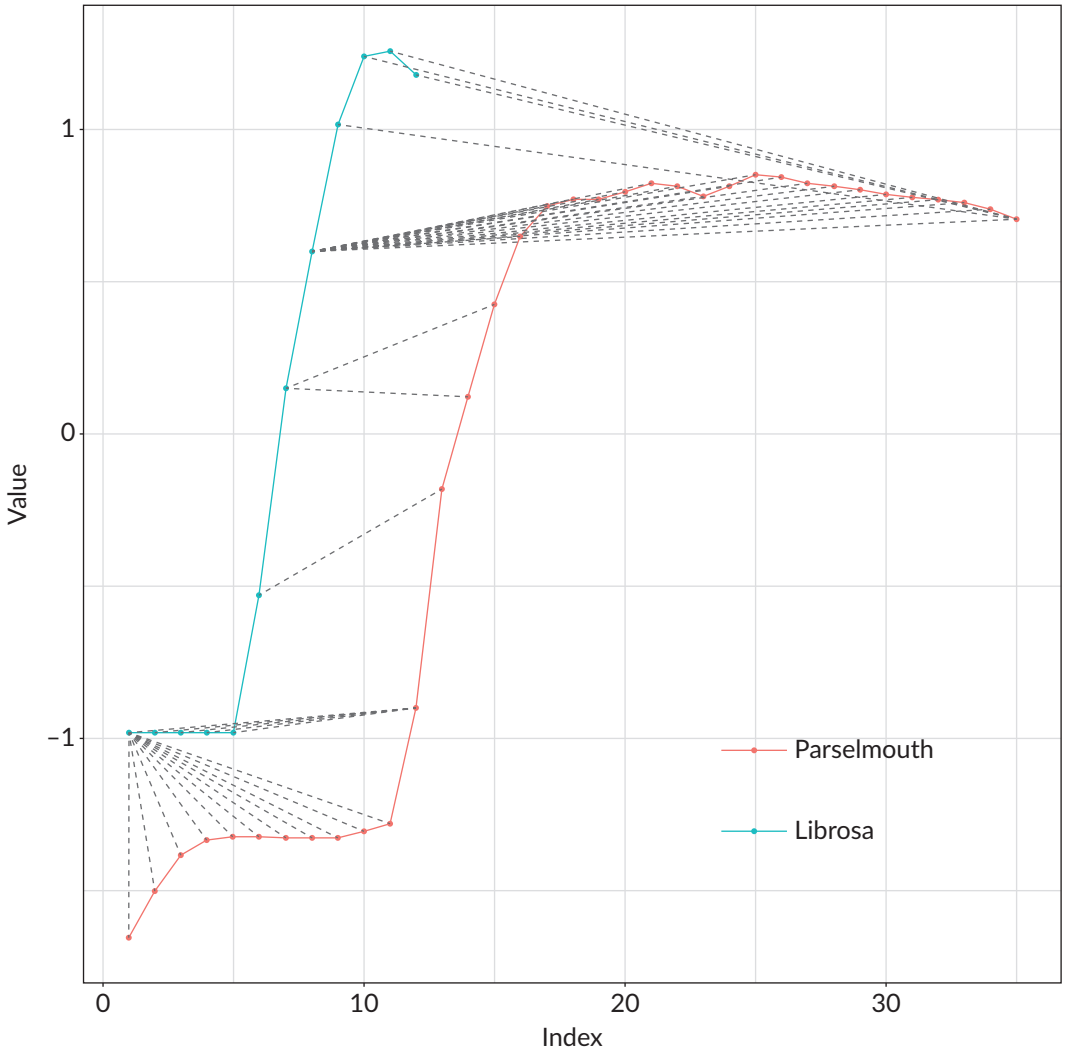

Figure 1. Ten-second time series of intensity as measured by parselmouth and librosa.

The dotted lines between the parselmouth and librosa lines refer to the optimal alignment. In the case of this data, one librosa point may be illustrated by many parselmouth points (the first librosa point on the bottom left has 11 parselmouth points, for example).

While this granularity may be useful when researchers are identifying a millisecond-level analysis (such as with fMRI or physiological phenomena; see Jahn et al., 2022; Lang et al., 2009), these differences may not be as substantive at higher levels of aggregation, such as at the daily or monthly level.

5. Discussion

Our comparison of the speech-to-text and computational phonology tools suggests several differences. Regarding automated transcription, the WER indicates that GCP Speech-to-Text and OpenAI's Whisper may differ. Notably, GCP appears to provide more direct transcriptions, including stutters and fillers, whereas Whisper's transcriptions may have these auditory features removed. Based on these findings, we suggest that researchers select a transcription tool that aligns best with their question or the type of spoken language they seek to study. For example, researchers studying anxiety and pausing in interpersonal conversation may want to identify filler words. However, studies of scripted broadcasts, which already lack many of these language features, may be better suited for processing by OpenAI's Whisper. We also encourage researchers to consider the cost difference between these two tools and whether the increased price for GCP justifies its use.

The comparison of librosa and parselmouth indicates that these two tools were more similar, at least regarding pitch and intensity. However, we also found that parselmouth provided more granular data. This difference has its benefits and disadvantages: Millisecond-level analyses may be necessary for some types of data (e.g., fMRI). However, the more granular output also results in larger and more computationally intensive datasets.

These results highlight the importance of reproduced processing for audio-as-data. While the results produced by the two automated transcription tools (the reductive processing approach) and by the audio feature extraction tools (the additive processing approach) were similar, our study also found some key differences that may motivate researchers to use one tool over the other. This is an essential consideration for reproduction and replication studies as results may not reproduce, not because of the data or context, but because of a difference in processing. Such findings also highlight the importance of methodological transparency and being specific about what packages or programs a researcher used to process their audio data.

In conducting this work, we contribute to the growing literature on data processing (e.g., Denny & Spirling, 2018; Tabassum & Patil, 2020)—and, specifically, its reproducibility—and expand it in the context of audio-as-data. By showcasing methods for reductive and additive processing, and by comparing several tools, we hope this work motivates other media and communication scholars to study audio data more.

Based on these findings, we make several important recommendations for improving the replicability and reproduction of research using audio-as-data. First, reproduction studies should use the same processing approaches, if possible. It is not only a matter of using the same methods, but being explicit about the specific tool, software, programming language, or package/library used to conduct the analysis. Because of the myriad of ways researchers can prepare their data—particularly in computational research and content analysis methods—the ability to replicate a finding is contingent on methodological clarity.

This result leads to our second recommendation: Researchers must also be transparent in their processing approach, including providing information about the tools they used, the version of that tool (if relevant), and the specific steps they conducted in data preparation. This recommendation aligns with open science practices (Dienlin et al., 2020), particularly regarding making research materials more open.

Finally, researchers should validate their results with reproducible processing (i.e., using different processing software to achieve the same task). This procedure ensures the robustness of one's results. Future studies can build on this work by analyzing data processing tools across different types of media (e.g., podcasts, social media content, broadcasts of speeches). For example, one area that would benefit from a greater assessment is the consideration of audio-as-data with music features. While some parts of talk radio contain music (e.g., advertisements), most talk radio audio is spoken language; as such, a limitation of our study is its focus on spoken language. Future studies should seek to reproduce these results with media content that contains music audio, including recordings of songs and movies. Another limitation of this study is our focus on English-language talk radio. As GCP Speech-to-Text claims to support 125 language variations, and OpenAI Whisper claims to support 99, future work can and should consider how these tools may expand non-English audio-as-data studies.

Acknowledgments

We would like to express our gratitude to the reviewers in *Media and Communication* who provided feedback on earlier drafts of the manuscript, as well as the support of the Media and Democracy Data Cooperative members. Ross Dalhke is supported by graduate fellowship awards from Knight-Hennessy Scholars and Stanford Data Science Scholars at Stanford University.

Funding

Funding for this work was provided by the John S. and James L. Knight Foundation.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online in the format provided by the author (unedited).

References

Alakrot, A., Murray, L., & Nikolov, N. S. (2018). Towards accurate detection of offensive language in online communication in Arabic. *Procedia Computer Science*, *142*, 315–320.

Armstrong, C. B., & Rubin, A. (1989). Talk radio as interpersonal communication. *Journal of Communication*, *39*(2), 84–94.

Avery, R., Ellis, D., & Glover, T. (1978). Patterns of communication on talk radio. *Journal of Broadcasting & Electronic Media*, *22*(1), 5–17.

Babu, P., Nagaraju, V., & Vallabhuni, R. (2021). Speech emotion recognition system with librosa. In G. S. Tomar & K. Sudhakar (Eds.), *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)* (pp. 421–424). IEEE.

Barker, D. (1998). Rush to action: Political talk radio and health care (un) reform. *Political Communication*, *15*(1), 83–97.

Beauchamp, Z. (2016, December 7). Alex Jones, Pizzagate booster and America's most famous conspiracy theorist, explained. *Vox*. http://www.vox.com/policy-and-politics/2016/10/28/13424848/alex-jones-infowars-prisonplanet

Benoit, W., & Holbert, R. (2008). Empirical intersections in communication research: Replication, multiple quantitative methods, and bridging the quantitative–qualitative divide. *Journal of Communication*, *58*(4), 615–628.

Berndt, D., & Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In U. M. Fayyad & R. Uthurusamy (Eds.), *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining* (pp. 359–370). AAAI Press.

Berry, J. M., & Sobieraj, S. (2013). *The outrage industry: Political opinion media and the new incivility*. Oxford University Press.

Bianchini, G., Zanotti, L., & Meléndez, C. (2023). *Using OpenAI models as a new tool for text analysis in political leaders' unstructured discourse*. Unpublished manuscript. https://osf.io/preprints/psyarxiv/kdngb/download

Bokhove, C., & Downey, C. (2018). Automated generation of "good enough" transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations*, *11*(2). https://doi.org/10.1177/2059799118790743

Chen, X., Yang, J., Gan, S., & Yang, Y. (2012). The contribution of sound intensity in vocal emotion perception: Behavioral and electrophysiological evidence. *PLoS One*, *7*(1), Article e30278.

Christenson, P. G., & Lindlof, T. R. (1983). The role of audio media in the lives of children. *Popular Music & Society*, *9*(3), 25–40.

Denny, M. J., & Spirling, A. (2018). Text preprocessing for unsupervised learning: Why it matters, when it misleads, and what to do about it. *Political Analysis*, *26*(2), 168–189.

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. (2020). An agenda for open science in communication. *Journal of Communication*, *71*(1), 1–26. https://doi.org/10.1093/joc/jqz052

Dietrich, B., Hayes, M., & O'Brien, D. Z. (2019). Pitch perfect: Vocal pitch and the emotional intensity of congressional speech. *American Political Science Review*, *113*(4), 941–962.

Dowling, D., Johnson, P., & Ekdale, B. (2022). Hijacking journalism: Legitimacy and metajournalistic discourse in right-wing podcasts. *Media and Communication*, *10*(3), 17–27.

Dunne-Howrie, J. (2019). Crisis acting in the destroyed room. *Performance Research*, *24*(5), 65–73.

Gold, B., Morgan, N., & Ellis, D. (2011). *Speech and audio signal processing: Processing and perception of speech and music*. Wiley.

Hall, A., & Cappella, J. N. (2002). The impact of political talk radio exposure on attributions about the outcome of the 1996 US presidential election. *Journal of Communication*, *52*(2), 332–350.

Harris, C., Mayer, V., Saulino, C., & Schiller, D. (1996). The class politics of Rush Limbaugh. *The Communication Review*, *1*(4), 545–564. https://doi.org/10.1080/10714429609388278

Hegazi, M. O., Al-Dossari, Y., Al-Yahy, A., Al-Sumari, A., & Hilal, A. (2021). Preprocessing Arabic text on social media. *Heliyon*, *7*(2), Article e06191. https://doi.org/10.1016/j.heliyon.2021.e06191

Howell, E. (2020). Science communication in the context of reproducibility and replicability: How nonscientists navigate scientific uncertainty. *Harvard Data Science Review*, *2*(4). https://doi.org/10.1162/99608f92.f2823096

Indrayani, Asfiati, S., Riky, M. N., & Rajagukguk, J. (2020). Measurement and evaluation of sound intensity at the Medan Railway Station using a sound level meter. *Journal of Physics: Conference Series*, *1428*(1), Article 012063. https://doi.org/10.1088/1742-6596/1428/1/012063

Iosad, P. (2015). "Pitch accent" and prosodic structure in Scottish Gaelic: Reassessing the role of contact. In M. Hilpert, J.-O. Östman, C. Mertzlufft, M. Rießler, & J. Duke (Eds.), *New trends in Nordic and general linguistics* (pp. 28–54). De Gruyter.

Jadoul, Y., Thompson, B., & De Boer, B. (2018). Introducing parselmouth: A Python interface to praat. *Journal of Phonetics*, *71*, 1–15.

Jahn, N. T., Meshi, D., Bente, G., & Schmälzle, R. (2022). Media neuroscience on a shoestring. *Journal of Media Psychology*, *35*(2). https://doi.org/10.1027/1864-1105/a000348

Jamieson, K. H., Cappella, J. N., & Joseph, T. (1998). Limbaugh: The fusion of party leader and partisan mass medium. *Political Communication*, *15*(Suppl. 1), 1–27. https://doi.org/10.1080/10584609.1998.11672652

Jeanrenaud, P., Eide, E., Chaudhari, U., McDonough, J., Ng, K., Siu, M., & Gish, H. (1995, May 9-12). Reducing word error rate on conversational speech from the Switchboard corpus. In *1995 International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 53–56). IEEE.

Juneja, A., & Das, N. (2019). Big data quality framework: Pre-processing data in weather monitoring application. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)* (pp. 559–563). IEEE.

Kinoshita, N. (2015). Learner preference and the learning of Japanese rhythm. In J. Levis, R. Mohammed, M. Qian, & Z. Zhou (Eds.), *Proceedings of the 6th Pronunciation in Second Language Learning and Teaching Conference* (pp. 49–62). Iowa State University Press.

Klakow, D., & Peters, J. (2002). Testing the correlation of word error rate and perplexity. *Speech Communication*, *38*(1/2), 19–28.

Klofstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political Psychology*, *37*(5), 725–738.

Klofstad, C. A., Anderson, R., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), 2698–2704.

Knox, D., & Lucas, C. (2021). A dynamic model of speech for the social sciences. *American Political Science Review*, *115*(2), 649–666.

Krumhansl, C., & Iverson, P. (1992). Perceptual interactions between musical pitch and timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(3), 739–751.

Lang, A., Potter, R., & Bolls, P. (2009). Where psychophysiology meets the media: Taking the effects out of mass media research. In J. Bryant & M. B. Oliver (Eds.), *Media effects* (pp. 201–222). Routledge.

Larrouy-Maestri, P., Kegel, V., Schlotz, W., van Rijn, P., Menninghaus, W., & Poeppel, D. (2023). Ironic twists of sentence meaning can be signaled by forward move of prosodic stress. *Journal of Experimental Psychology: General*, *152*(9), 2438–2462. https://doi.org/10.1037/xge0001377

Larsen, E., & Aarts, R. (2005). *Audio bandwidth extension: Application of psychoacoustics, signal processing and loudspeaker design*. Wiley.

Lee, G., & Cappella, J. N. (2001). The effects of political talk radio on political attitude formation: Exposure versus knowledge. *Political Communication*, *18*(4), 369–394.

Leodolter, M., Plant, C., & Brändle, N. (2021). IncDTW: An R package for incremental calculation of dynamic time warping. *Journal of Statistical Software*, *99*(9), 1–23. https://doi.org/10.18637/jss.v099.i09

Loakes, D., & Keith, A. (2013). From IPA to praat and beyond. In K. Allan (Ed.), *The Oxford handbook of the history of linguistics* (pp. 123–140). Oxford University Press.

Lucas, C. (2022). *Package "communication": Feature extraction and model estimation for audio of human speech*. CRAN. https://cran.r-project.org/web/packages/communication/communication.pdf

Lukito, J. (2023). *Political language and the computational turn: Political communication report, 2023*. https://doi.org/10.17169/refubium-39046

Lukito, J., Brown, M. A., Dahlke, R., Suk, J., Yang, Y., Zhang, Y., Chen, B., Kim, S. J., & Soorholtz, K. (2023). *The state of digital media data research, 2023*. Media & Democracy Data Cooperative. https://doi.org/10.26153/tsw/46177

Lukito, J., Gursky, J., Foley, J., Yang, Y., Joseff, K., & Borah, P. (2023). "No reason[.] [i]t /should/ happen here": Analyzing Flynn's retroactive doublespeak during a QAnon event. *Political Communication*, *40*(5), 576–595. https://doi.org/10.1080/10584609.2023.2185332

Luz, S., Masoodian, M., Rogers, B., & Deering, C. (2008). Interface design strategies for computer-assisted speech transcription. In N. Bidwell (Ed.), *Proceedings of the 20th Australasian Conference on Computer–Human Interaction: Designing for Habitus and Habitat* (pp. 203–210). Association for Computing Machinery.

Madison, T. P., Covington, E. N., Wright, K., & Gaspard, T. (2019). Credibility and attributes of parasocial relationships with Alex Jones. *Southwestern Mass Communication Journal*, *34*(2). https://doi.org/10.58997/smc.v34i2.45

Madison, T. P., Wright, K., & Gaspard, T. (2020). "My superpower is being honest:" Perceived credibility and parasocial relationships with Alex Jones. *Southwestern Mass Communication Journal*, *36*(1), 50–64.

Matzko, P. (2020). *The radio right: How a band of broadcasters took on the federal government and built the modern conservative movement*. Oxford University Press.

McEwan, B., Carpenter, C., & Westerman, D. (2018). On replication in communication science. *Communication Studies*, *69*(3), 235–241.

McFee, B., Raffel, C., Liang, D., Ellis, D., McVicar, M., Battenberg, E., & Nieto, O. (2015). Librosa: Audio and music signal analysis in Python. In K. Huff & J. Bergstra (Eds.), *Proceedings of the 14th Python in Science Conference* (Vol. 8, pp. 18–25). SciPy.

Mercieca, J. (2019). Dangerous demagogues and weaponized communication. *Rhetoric Society Quarterly*, *49*(3), 264–279.

Morchid, M., Dufour, R., & Linarès, G. (2016). Impact of word error rate on theme identification task of highly imperfect human–human conversations. *Computer Speech & Language*, *38*, 68–85.

Naseem, U., Razzak, I., & Eklund, P. (2021). A survey of pre-processing techniques to improve short-text quality: A case study on hate speech detection on Twitter. *Multimedia Tools and Applications*, *80*, 35239–35266.

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. https://nap.nationalacademies.org/catalog/25303/reproducibility-and-replicability-in-science

Nosek, B. A., & Errington, T. (2020). What is replication? *PLoS Biology*, *18*(3), Article e3000691.

Piñeiro-Otero, T., & Pedrero-Esteban, L.-M. (2022). Audio communication in the face of the renaissance of digital audio. *El Profesional de la Información*, *31*(5). https://doi.org/10.3145/epi.2022.sep.07

Plesser, H. E. (2018). Reproducibility vs. replicability: A brief history of a confused terminology. *Frontiers in Neuroinformatics*, *11*(76). https://doi.org/10.3389/fninf.2017.00076

Proksch, S., Wratil, C., & Wäckerle, J. (2019). Testing the validity of automatic speech recognition for political text analysis. *Political Analysis*, *27*(3), 339–359.

Purwins, H., Li, B., Virtanen, T., Schlüter, J., Chang, S. Y., & Sainath, T. (2019). Deep learning for audio signal processing. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 206–219.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492–28518). PMLR.

Samrose, S., & Hoque, E. (2021). Quantifying the intensity of toxicity for discussions and speakers. In A. Leontyev, T. Yamauchi, & M. Razavi (Eds.), *2021 9th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)* (pp. 1–5). IEEE.

Shah, D. V., Sun, Z., Bucy, E. P., Kim, S. J., Sun, Y., Li, M., & Sethares, W. (2023). Building an ICCN multimodal classifier of aggressive political debate style: Towards a computational understanding of

candidate performance over time. *Communication Methods and Measures*, *18*(1), 30–47. https://doi.org/10.1080/19312458.2023.2227093

Shakhovska, N., Basystiuk, O., & Shakhovska, K. (2019). Development of the speech-to-text chatbot interface based on Google API. In M. Emmerich, V. Lytvyn, I. Yevseyeva, V. Basto-Fernandes, D. Dosyn, & V. Vysotska (Eds.), *MoMLeT&DS—2019: Modern Machine Learning Technologies and Data Science Workshop* (pp. 212–221). CEUR Workshop Proceedings. https://shorturl.at/elwBO

Shao, X. (2015). Self-normalization for time series: A review of recent developments. *Journal of the American Statistical Association*, *110*(512), 1797–1817.

Sihag, M., Li, Z. S., Dash, A., Arony, N. N., Devathasan, K., Ernst, N., Branzan Albu, A., & Damian, D. (2023). A data-driven approach for finding requirements relevant feedback from TikTok and YouTube. In K. Schneider, F. Dalpiaz, & J. Horkoff (Eds.), *2023 IEEE 31st International Requirements Engineering Conference (RE 2023)* (pp. 111–122). IEEE.

Slater, J. (2022, October 12). Connecticut jury orders Alex Jones to pay nearly $1 billion to Sandy Hook families. *The Texas Tribune*. https://www.texastribune.org/2022/10/12/alex-jones-sandy-hook-shooting

Spinelli, M., & Dann, L. (2019). *Podcasting: The audio media revolution*. Bloomsbury.

Tabassum, A., & Patil, R. (2020). A survey on text pre-processing & feature extraction techniques in natural language processing. *International Research Journal of Engineering and Technology (IRJET)*, *7*(6), 4864–4867.

Tomar, M., Tiwari, A., Saha, T., & Saha, S. (2023). Your tone speaks louder than your face! Modality order infused multi-modal sarcasm detection. In A. El Saddik, T. Mei, & R. Cucchiara (Eds.), *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 3926–3933). Association for Computing Machinery.

Van Atteveldt, W., Strycharz, J., Trilling, D., & Welbers, K. (2019). Toward open computational communication science: A practical road map for reusable data and code. *International Journal of Communication*, *13*, 3935–3954.

Verteletskaya, E., Sakhnov, K., & Simak, B. (2009). Pitch detection algorithms and voiced/unvoiced classification for noisy speech. In *2009 16th International Conference on Systems, Signals and Image Processing* (pp. 1–5). IEEE.

Wang, Y., Acero, A., & Chelba, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. In J. Bilmes & W. Byrne (Eds.), *2003 IEEE Workshop on Automatic Speech Recognition and Understanding* (pp. 577–582). IEEE.

Williamson, E. (2022). *Sandy Hook: An American tragedy and the battle for truth*. Penguin.

Wisniewski, G., Michaud, A., & Guillaume, S. (2020). Phonemic transcription of low-resource languages: To what extent can preprocessing be automated? In D. Beermann, L. Besacier, S. Sakti, & C. Soria (Eds.), *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)* (pp. 306–315). European Language Resources Association.

Wurst, C. (2022). Bread and plots: Conspiracy Theories and the rhetorical style of political influencer communities on YouTube. *Media and Communication*, *10*(4), 213–223.

Young, D. G. (2020). *Irony and outrage: The polarized landscape of rage, fear, and laughter in the United States*. Oxford University Press.

About the Authors

Josephine Lukito (PhD) is an assistant professor at the University of Texas at Austin's School of Journalism and Media and director of the Media and Democracy Data Cooperative. She is also a Senior Faculty Research Affiliate for the Center for Media Engagement.

Jason Greenfield is a research engineer at NYU's Center for Social Media and Politics. He is especially interested in studying instances of online extremism, the radicalization process, and the interplay between humor and hate.

Yunkang Yang (PhD) is an assistant professor at the Department of Communication and Journalism at Texas A&M University where he is also affiliated with the Data Justice Lab. He has a forthcoming book titled *Weapons of Mass Deception: How Right-Wing Media Wage Information Warfare and Undermine American Democracy*.

Ross Dahlke is a PhD candidate at Stanford University in the Department of Communication.

Megan A. Brown is a PhD student at the School of Information at the University of Michigan.

Rebecca Lewis is a Stanford Graduate Fellow and PhD candidate in Communication at Stanford University. She is an expert on disinformation and far-right digital media.

Bin Chen is a doctoral candidate in Journalism and Media at the University of Texas at Austin. His research focuses on political communication, multi-platform research, and computational social science.

Media and Communication
2024 • Volume 12 • Article 7836
https://doi.org/10.17645/mac.7836

ARTICLE

Open Access Journal

Standardized Sampling for Systematic Literature Reviews (STAMP Method): Ensuring Reproducibility and Replicability

Ayanda Rogge, Luise Anter, Deborah Kunze, Kristin Pomsel, and Gregor Willenbrock

Institute of Media and Communication, TU Dresden, Germany

Correspondence: Ayanda Rogge (ayanda.rogge@tu-dresden.de)

Submitted: 10 November 2023 **Accepted:** 5 February 2024 **Published:** 3 April 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Systematic literature reviews (SLRs) are an effective way of mapping a research field and synthesizing research evidence. However, especially in communication research, SLRs often include diverse theories and methods, which come with a considerable downside in terms of reproducibility and replicability. As a response to this problem, the present article introduces the method of standardized sampling for systematic literature reviews (STAMP). The method is a structured, four-stage approach that is centered around score-based screening decisions. Originating from principles of standardized content analysis, a method common in communication research, and supplementing established guidelines like Cochrane or PRISMA, the STAMP method contributes to more transparent, reproducible, and replicable SLR sampling processes. As we illustrate throughout the article, the method is adaptable to various SLR types. The article also discusses the method's limitations, such as potential coder effects and comparatively high resource intensity. To facilitate the application of STAMP, we provide a comprehensive guideline via the Open Science Framework that offers a succinct overview for quick reference and includes practical examples for different types of SLRs.

Keywords

content analysis; replicability; reproducibility; STAMP method; standardized sampling; systematic literature review

1. Introduction

Systematic literature reviews (SLRs) are an effective way of mapping a research field, synthesizing research evidence, and distinguishing "between real and assumed knowledge" (Petticrew & Roberts, 2006, p. 2). However, when applied in the social sciences, such as in communication research, scholars often face a

theoretically and methodologically diverse range of publications that do not permit quantitative summarization of results. For example, when summarizing evidence on the production or use of news, literature samples might well include both semi-structured interviews and ethnographies as well as standardized or automated content analyses and experiments (e.g., Anter, 2023; Melchior & Oliveira, 2022). Furthermore, the research interests of SLRs in communication research often extend beyond synthesizing knowledge on a specific question. Instead, these SLRs repeatedly aim to build theories, critically investigate a research field, and identify its weaknesses and research gaps (e.g., Engelke, 2019; Ratcliff et al., 2022; see also Paré et al., 2015). Against this background, traditional meta-analyses that aggregate quantitative data from similar studies to determine an overall effect estimate (Davis et al., 2014) are often unsuitable for SLRs in the field of communication. Instead, quantitative analysis is primarily employed for mapping the research field in terms of study characteristics and theories as well as applied methods (e.g., Kümpel et al., 2015). For synthesizing existing evidence, many SLRs also apply qualitative techniques such as textual analysis (e.g., Humayun & Ferrucci, 2022) or open coding of relevant text parts (e.g., Belair-Gagnon & Steinke, 2020). However, despite the necessity and utility of these mixed-methods or qualitative SLRs, these approaches come with a considerable downside in terms of reliability and reproducibility. In fact, it is hardly possible to calculate reliability scores for qualitative content analysis, nor is it feasible to reproduce the analysis precisely since qualitative techniques generally allow for more interpretive freedom. Against this backdrop, the screening and selection process of the publications becomes even more critical for the diverse range of SLR efforts in communication research, including (partly) qualitative, quantitative, computational, or mixed-methods SLRs. Accordingly, this article addresses the need for improved reproducibility and replicability in SLR samples given a heterogeneous research field, advocating for using standardized content analysis to systematically apply eligibility criteria to literature selection.

To date, to the best of our knowledge, this issue has not been thoroughly addressed in established practical guides for SLRs nor in existing SLRs. For instance, while Petticrew and Roberts (2006) offer detailed information about developing criteria for the inclusion or exclusion of publications and conducting the literature search, they appear to "skip" the critical step of applying the criteria to the publications resulting from the literature search. Similarly, the PRISMA protocol (Page et al., 2021, Item 8), which many SLRs in the field of communication already adhere to (e.g., Melchior & Oliveira, 2022; Pirkis et al., 2019), the SLR preregistration form recently provided by van den Akker et al. (2023), and the guidelines provided by the Methods of Synthesis and Integration Center (Pigott & Polanin, 2020) emphasize the importance of the screening stage and discuss different approaches (e.g., priority screening for only highly relevant records or single screening of all records). Still, they neither offer guidance on how to apply the eligibility criteria systematically to specific parts of the publications nor do they elaborate on efficiently documenting the screening and selection process—especially for the sake of reproducibility.

Correspondingly, many existing SLRs in the field of communication describe both their literature search procedure and the development of their criteria in a detailed manner (e.g., Belair-Gagnon & Steinke, 2020; Engelke, 2019). However, information about the inclusion/exclusion process often remains limited to whether and when researchers applied criteria to the abstract or the whole text. In addition, it is often unclear under what criteria researchers consulted the abstract or the full text for their inclusion decision (e.g., Hase et al., 2023), or full texts are not consulted at all (e.g., Joris et al., 2020). Therefore, even though the screening result (inclusion/exclusion) is sometimes provided, readers and researchers may find it challenging to replicate the sampling process or even reproduce the sampling decision for single

publications—especially regarding the content criteria, such as for assessing the thematic fit of a publication. This issue becomes even more significant in the case of SLRs that do not (e.g., Gambo & Özad, 2020) or only vaguely (e.g., Castells-Fos et al., 2023) provide their eligibility criteria.

In response to the issues regarding the replicability and reproducibility of SLR sampling processes outlined above, we argue that the screening and selection of literature (i.e., sampling) significantly benefit from standardized content analysis, a method that is genuine to and prevalent within communication research (Haim et al., 2023, p. 280). The method is described as a valid and replicable measurement of texts' meaning by assigning categories to content (Krippendorff, 2004, p. 18). This describes what standardized SLR sampling should involve: predefined categories (i.e., eligibility criteria) that are systematically applied to specific text parts in an academic publication (i.e., title, keywords, abstract, or full text). This coding provides the basis for inferring a publication's fit with the SLR's research interest. Understanding SLR sampling as standardized content analysis also means making use of the research techniques that have been developed to ensure reproducibility and replicability—methodological requirements that do not only apply to content analysis but have been dealt with prominently in the respective literature (e.g., Haim et al., 2023; Lombard et al., 2002). Most basically, applying these techniques instructs researchers not only to report the eligibility criteria and the inclusion results of their SLR sampling (as is already common for SLRs in our field) but also to plan, conduct, and document the whole sampling process systematically and transparently.

To guide this process, we propose a four-stage approach for systematically sampling publications applicable to the variety of SLR efforts in communication research. At its core, the procedure employs scores that quantify a paper's eligibility for the respective SLR to standardize and protocol the sampling process, which is why we propose our so-called STAMP method for STandardized sAMPling. Specifically, the STAMP method includes:

- Stage 1: The development of eligibility criteria for including publications and reflecting the review's objective and scope;
- Stage 2: The identification of potentially relevant literature through databases;
- Stage 3: Narrowing the sample by assigning an abstract-based screening (ABS) score;
- Stage 4: Determining the final sample by assigning each publication a full-text reading (FTR) score.

We particularly emphasize the importance of thorough documentation of every step in the review protocol, which functions as a codebook for the content analyses conducted in Stages 3 and 4, when the eligibility criteria are applied to the publications in order to deduce the score-based screening decision. This continuous documentation of each screening stage allows researchers to constantly reflect on the SLR's progress to conform to its goals. The procedure also increases both the replicability and the reproducibility of the SLR's findings: Using STAMP and its rigorous documentation enables both replicating the SLR with newly compiled literature samples and reproducing the coding process within the same literature sample. Another benefit of the proposed method is its customizability to different research questions and techniques: Our prior SLRs show that the procedure is convenient for both evidence-summarizing SLRs that investigate a clearly defined research field (Anter, 2023) and concept-building SLRs that have to iteratively approach a literature sample to provide a valid theory synopsis (Rogge, 2023). Moreover, our approach involves the strength of measuring sample reliability as the ABS and FTR stages evaluate the content fit of publications by applying the principles of standardized content analysis to the SLR sampling process.

Importantly, we do not consider STAMP as a substitute for established SLR procedures, such as the Cochrane guidelines for systematic reviews (Higgins et al., 2023) or the PRISMA 2020 statement (Page et al., 2021). Rather, we see it as a *supplement* that refines the sampling process in order to increase the reproducibility and replicability of SLRs. Both the Cochrane and the PRISMA guidelines originate from the context of health care research, whereas STAMP is based in the field of communication. Thus, STAMP uses the opportunity to connect field-specific practices of communication research with established SLR guidelines. Inherently, the STAMP method aligns with guidelines established by Cochrane and PRISMA, employing their key practices such as using a review protocol, defining eligibility criteria, and documenting the search strategy. Additionally, STAMP provides more concrete guidance for inclusion and exclusion decisions than Cochrane and PRISMA, which merely stress the significance of this process in general and provide basic suggestions regarding the sampling process, such as pre-testing the eligibility criteria, training the (independent) reviewers (Cochrane), and transparently documenting the number of reviewers per publication (PRISMA). Moreover, while Cochrane and PRISMA primarily address possible bias within individual publications as well as bias within the SLR's synthesis, STAMP focuses on the *reduction* of bias within the sampling process.

To present our methodological procedure and facilitate its adoption by researchers planning to conduct an SLR, we describe the standardized sampling method for SLRs in detail. In addition, we provide a guideline including practical examples where the method has been successfully tested in concept-building (inductive) and evidence-summarizing (deductive) SLRs via OSF. These SLRs cover diverse communication research areas, such as health communication, human–machine communication, and journalism studies, highlighting the adaptability of the STAMP method.

2. The Four Stages of STAMP

In this section, we present the method designed to guide communication researchers through a systematic and reproducible sampling process for SLRs. We introduce the four stages, as summarized in Figure 1, as well as their goals, procedures, and practical recommendations. There are different ways to apply the approach to SLRs, so a guideline including practical examples of the four-stage procedure is available on OSF (https://bit.ly/4atGsvN).

2.1. Stage 1: Eligibility Criteria

The primary goal of this stage is to transparently establish the scope and objectives of the SLR.

The foundation of this stage lies in the specification of a review protocol, an established way to document SLRs (Nightingale, 2009; for an example, see also Page et al., 2021). First, a review protocol summarizes essential components such as the study's rationale, definitions of pivotal terms, and research questions/hypotheses. Second, the review protocol defines relevant keywords for constructing and validating the search string (Stage 2). To define the keywords for the SLR, researchers should choose keywords from pertinent literature as a starting point, such as background literature and related work (Mishra et al., 2009). The exploration of existing literature is crucial in order to identify synonyms for terms in other disciplines or earlier works. This initial selection of keywords should be enriched by using the snowball method or own keywords. Third, the review protocol contains the SLR's eligibility criteria, that is, criteria for deciding which publications remain in the literature review and which are excluded (Petticrew &

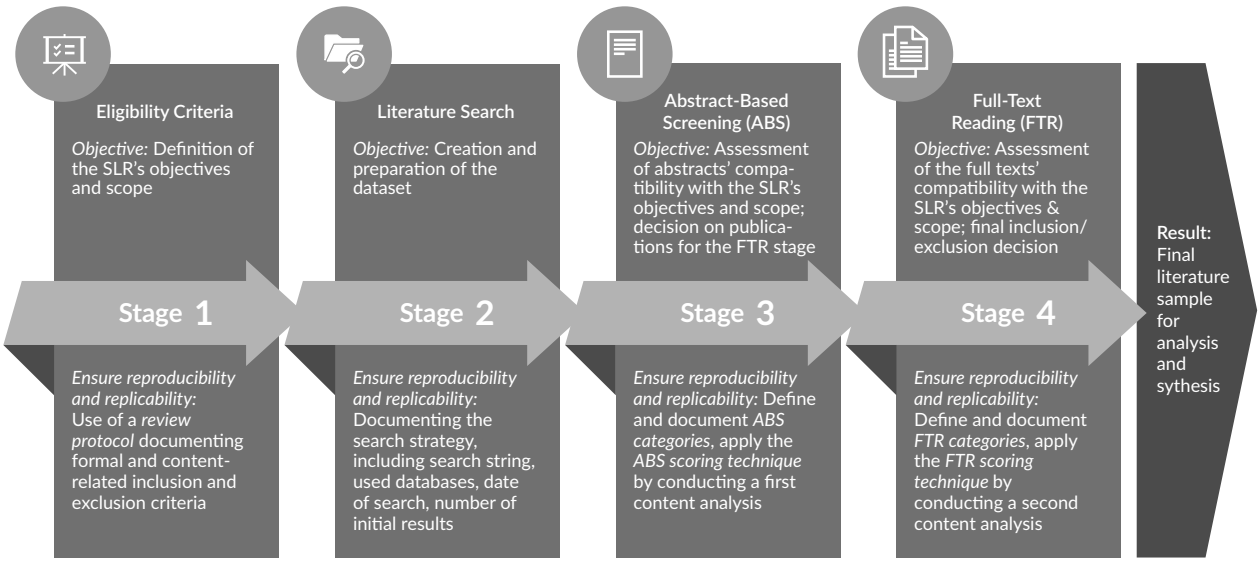

Figure 1. Illustration of the four stages of the STAMP method.

Roberts, 2006). These criteria comprise both formal (e.g., the language of the publications, publication years and types, countries of origin, and scientific disciplines) and content-related factors (e.g., topics or variables based on research questions). Defining and transparently documenting the inclusion and exclusion criteria is essential to enable a transparent research process. These eligibility criteria are also used for the ABS (Stage 3) and FTR (Stage 4). In a nutshell, the review protocol explicitly defines the observer-independent rules that govern the sampling process and are applied to all units of analysis in the abstract and full-text stages—a key requirement for replicability (Krippendorff, 2004, p. 19). Therefore, it is a central component of our method and it is conceived to take on the role of a codebook in Stages 3 and 4.

At its core, the first stage aims at setting up the review protocol and the eligibility criteria that ensure systematic alignment with the stipulated research objectives and scope. Thus, the first stage lies the foundation for an efficient and systematic review of the literature.

Is the SLR's scope feasible? Researchers need to consider available resources when planning and conducting an SLR. Primarily, these resources include time, personnel, and access (to databases, literature, etc.). The availability of these resources will impact the duration of the SLR project. We suggest that researchers prepare a project plan and consider the possibility of reducing the SLR scope if resources are limited. Also, regardless of the available resources, one should be open to adjusting the project plan and scope once researchers become more familiar with the nature of their analysis material.

2.2. Stage 2: Literature Search

The main objective of this stage is to conduct a literature search and collect the sampling units for the content analysis in Stages 3 and 4, i.e., a dataset including all possibly relevant publications.

For this purpose, researchers define the search strategy in the review protocol and document the search. This documentation includes the search string (Bartels, 2013; Mishra et al., 2009) as well as all employed bibliographic databases, search engines, and repositories. A well-curated search string equips the team to

discover a broad spectrum of related literature and is pivotal to discerning the core consistencies or discrepancies across the research landscape. The search strategy, and especially the search string, should be developed iteratively. Hence, exploring and documenting different search string combinations is essential. The quality of the search strategy can be assessed by the amount of possibly relevant literature that the search generates. For example, researchers can include only search strings that yield a defined number of results. Additionally, researchers should predefine pertinent publications, authors, or journals in the protocol whose appearance in the search results validates the search strategy. After defining and validating the search strategy, the actual literature search is executed and the results are transferred into a dataset, including the publications' abstracts (see Stage 3). For all databases, search engines and repositories, the same search strategy should be applied. Duplicates resulting from searching multiple databases must be cleared in the dataset and logged in the protocol to ensure reliability.

The second stage revolves around systematically sourcing and organizing relevant literature—again, an already established procedure (Page et al., 2021). As a result, the first dataset of publications for the SLR emerges from Stage 2.

How to decide on databases? The selection of databases (e.g., Web of Science), repositories (e.g., arXiv), and scholarly search engines (e.g., Google Scholar) depends on the SLR's underlying rationale. While field-specific databases (e.g., Communication & Mass Media Complete) are particularly relevant for SLRs focusing on a narrow and discipline-specific topic, others encompass a broader research landscape (e.g., Scopus; Falagas et al., 2008). If the SLR needs to include grey literature, researchers should also use "regular" search engines (Falagas et al., 2008). In any case, we recommend using multiple databases to combine the strengths of different providers.

How to decide on a narrow or broad search string? Depending on the SLR's objectives and scope, search strings can either be sensitive (i.e., include a broad range of keywords and their synonyms) with the aim of exhaustion or specific (i.e., include a small selection of pertinent keywords) with the aim of precision (Petticrew & Roberts, 2006). For example, concept-building SLRs could use a broader search string that employs multiple theoretical dimensions or incorporates interdisciplinary approaches. In comparison, SLRs aiming to summarize evidence on a specific question might use a narrower search string. However, depending on the characteristics of the research field, it might also be necessary to use a more sensitive search string. This means a search string does not only depend on the type of SLR but must also capture the scope of the research objective. Thus, an evidence-summarizing SLR can also use broad terms in the search string if the associated research question requires this search approach. Options to expand the search string include the generous use of synonyms (Anter, 2023; Rogge, 2023) or wildcards (Kunze, 2024). Importantly, a broad search string inherently results in a larger data set, which may require drawing a random sample in Stages 3 or 4 to render the SLR feasible.

How to create a search string? We recommend using advanced search techniques such as Boolean operators, truncations, and wildcards in search terms to amplify the search's reach (e.g., Bartels, 2013; Mishra et al., 2009). Here, different databases might have distinct requirements, such as word limits causing minor variations in the search string. To assure the search's reproducibility and replicability, the review protocol should contain every search string and, if applicable, additional filters and limitations used for the respective database or search engine.

What to do with additional literature? Researchers might identify additional relevant literature *after* the systematic search is carried out (e.g., through recommendations of colleagues or reviewers). If researchers consider ignoring this literature as detrimental, it is suitable to define an extended sample (Petticrew & Roberts, 2006), including all publications sampled because of the researchers' subjective decision. To ensure transparency, each publication's origin (systematic vs. extended sample) should be documented in the final sample.

2.3. Stage 3: ABS

The third stage focuses on an ABS of the publications collected in Stage 2. The primary purpose is to assess the compatibility of each abstract with the research objectives.

In Stage 3, researchers read all abstracts and assess their fit regarding the eligibility criteria. Therefore, inclusion and exclusion criteria (see Stage 1) are transferred into categories for a content analysis of the publications' abstracts (ABS categories), again defined in the review protocol, which now takes the function of a codebook. These categories are used to assess the fulfillment of the respective criteria through coding the abstracts. For each eligibility criterion, it is possible to use one or multiple ABS categories. While most formal criteria will only need one ABS category (e.g., language), using multiple categories for more complex content criteria (e.g., study findings) might be suitable. Moreover, as not every relevant section of a full article may be addressed in the abstract, it might be pragmatic to limit the ABS inclusion criteria to the presence of keywords in the abstract (e.g., a specific theory, topic, or method) and formal features of the publication (e.g., type). For instance, if an abstract mentions a relevant theory, concept, medium, method, or population, the respective category is coded with 1. As is apparent from this example, we recommend using binary coding, indicating whether an eligibility criterion (in the form of an ABS category) is *met* (1) or *not met* (0), but ordinal coding (e.g., 2 = *met*, 1 = *possibly met*, 0 = *not met*) is also reasonable for a more detailed evaluation. Additionally, researchers should adopt a "generous" coding attitude at this stage: In case of uncertainty about the fulfillment of a category, researchers can still include the respective publication and "postpone" the final decision to the following FTR stage. In order to ensure the reproducibility and replicability of coding, all coders should be provided with detailed coding instructions, precise definitions, and examples for each category (Krippendorff, 2004, p. 132). Additionally, as also highlighted in the Cochrane guidelines, we recommend coder training during which coders become familiar with the categories and code a small number of abstracts together with the head researcher, allowing the computation of intercoder reliability and quality assessments.

Based on the abstract coding, the ABS score is calculated as a sum index of all categories representing a publication's fit: the higher the ABS score, the better the fit. If multiple categories per eligibility criterion were used (e.g., several categories to assess the thematic fit), it is also possible to calculate separate scores per criterion. While an aggregated ABS score provides a holistic overview and simplifies the inclusion, scores per ABS category provide a more differentiated understanding of an abstract's alignment with various eligibility criteria. Usually, only publications that fulfill all ABS categories remain in the dataset. Alternatively, researchers can define a threshold value beforehand. The level of detail of the ABS assessment and the use of thresholds depends on the requirements of the literature review. For example, evidence-summarizing SLRs usually want to include only studies that deal with a specific question. Therefore, it is usually sufficient to directly transfer the eligibility criteria in an aggregate ABS score and exclude all publications that do not reach their maximum value (e.g., Anter, 2023). A concept-building SLR rather employs multiple ABS

categories reflecting different dimensions of the topic. In comparison to the evidence-summarizing SLR, the inclusion metric is less rigorous; abstracts must, for instance, address at least half of the ABS categories to remain for the FTR (e.g., Rogge, 2023).

Stage 3 comprises a truncated content analysis of the given abstracts. This screening process provides no analytical depth with regard to the SLR's research questions. Instead, its purpose is filtering, distinguishing relevant from less relevant literature. The benefit of the proposed scoring technique lies in a transparent, reasonable, and reproducible evaluation of the abstracts. As a result, the ABS score determines the publications that remain in the dataset for Stage 4.

How to deal with missing abstracts? This is particularly likely if an SLR also includes chapters in books or edited volumes, which often do not provide an abstract. In these cases, researchers can calculate the ABS score based on the publication's introduction and conclusion. Importantly, this "workaround" should be transparently documented in the review protocol.

Are the provided ABS decisions appropriate? Although it might be tempting, we do not recommend validating the ABS scores by delving deeper into the full texts. Not only could this cause self-referential loops, where researchers recurrently adjust the ABS categories and their coding decisions without moving to the next stage, but it also includes the risk of adjusting the ABS categories based on a potentially biased selection of full texts. To address uncertainty, we instead recommend using ordinal categories whose middle category indicates uncertainty. These publications can be re-assessed either during the next stage or at the end of the ABS phase. After reading a fair amount of both highly and hardly relevant abstracts, researchers usually gain sufficient experience to make clear coding decisions.

How to assure impartiality? An essential aspect of the ABS (and the FTR, Stage 4) is impartiality. For example, researchers might (unwillingly) be more generous with publications from scholars who are established and well-known in the field and oversee less prominent research that could be even more relevant. Therefore, researchers should ideally remain unaware of the authors behind each abstract/full text. A blind procedure is easily implemented in Stage 3 by anonymizing all bibliographic information in the data set (Stage 2) with an ID, which becomes the exclusive reference during Stages 3 and 4. Consequently, researchers can realize blind sampling at a low cost by using the export functions of today's databases with subsequent anonymization.

How to incorporate tools for automation? Automation tools can assist researchers in dealing with a large number of relevant publications, particularly when coding for relevant formal and content-related categories in ABS and FTR. When integrating automation tools into the STAMP method, it is essential to thoroughly document in the review protocol which tools (when and with which settings/outcomes) were used. One example of an SLR automation tool is the open-source software ASReview (van de Schoot et al., 2021). ASReview employs a machine-learning algorithm trained by the researchers to rank publications regarding their relevance to the SLR. This training process should be carefully documented, as there is no standardized guideline for when to end the training. While ASReview has the potential to reduce the amount of (irrelevant) publications that have to be screened during Stage 3, such automation tools limit the comprehensibility of the ABS coding and, in turn, the calculation of an ABS score. Consequently, while ASReview enhances efficiency, especially when a high number of publications have to be screened, it could hinder the detailed and transparent coding technique employed by the STAMP method.

2.4. Stage 4: FTR

In the final stage of the STAMP method, researchers perform an FTR of the publications remaining in the dataset and evaluate their fit with regard to the FTR categories by employing an FTR score. This stage aims to evaluate articles exclusively based on their contribution to the research interest and to create a final literature sample for the subsequent analysis and synthesis.

The FTR again utilizes content analytical principles to maintain consistency with the ABS stage. This includes creating an FTR category system with (binary/ordinal) categories, providing coding guidelines to ensure reliable assessment and comparability, as well as providing a coding scheme to document all coding decisions and the calculated FTR score(s). Therefore, similar to Stage 3, researchers create FTR categories based on the inclusion and exclusion criteria (eligibility criteria). As the FTR stage usually focuses on the publications' fit regarding *content*, the content criteria based on the SLR's research questions/hypotheses are transferred into concise categories—as before, documented in the review protocol serving as a codebook. For example, FTR categories could assess whether a publication investigates a specific variable relation or uses certain definitions and theories. Methodological aspects (e.g., study design) and results (such as effect sizes) might also be part of the FTR categories. Depending on the SLR's objective and scope, the FTR categories may be identical to the former ABS categories. In these cases, the researcher can confirm or revise a coding decision from Stage 3 based on the full text. This approach is convenient in evidence-summarizing SLRs with clearly defined research questions. On the other side, concept-building SLRs may require a more iterative procedure, which implies specifying the categories from the ABS to the FTR stage. Iteratively developing categories based on the material is necessary when the scope of a category system cannot be estimated initially. Although this procedure includes more effort, it brings the benefit of pretesting categories, reflecting on the purpose of a category, and inductively developing categories based on the given full texts. Moreover, especially for SLRs focusing on theoretical constructs, determining the fulfillment of inclusion criteria through standardized categories might be challenging. In these cases, it is important that the category system provides indicative phrases (e.g., "anchor examples" that include typical phrases of a definitional approach) for each category to facilitate the coding process.

After preparing the category system, researchers (roughly) read the full texts and code whether they contain relevant passages or arguments. As with the previous stage, an unbiased approach is crucial, so the evaluation should ideally be conducted blindly using the ID of an article. In Stage 4, blind sampling can be efficiently implemented by saving the full texts using the ID and blanking all bibliographic information. Again, there are different options to create an FTR score. Binary coding is appropriate if researchers only want to decide *if* a full text contains important information to answer a research question/hypothesis. For instance, the FTR category "topic" could indicate if a publication deals with a topic *within the scope of the SLR* (1) or *not within the scope* (0). Another option would be to evaluate the importance of a publication per eligibility criterion, e.g., through categories counting relevant arguments that a full text provides. For example, if the SLR aims to derive a definition, the associated categories capture how many definitional statements the full texts contain. As with the ABS, the FTR score can be calculated as a sum score per eligibility criterion or aggregated into a comprehensive FTR sum score. The FTR score then serves as an indicator of a publication's overall importance regarding the review's objectives.

Researchers transparently have to define the required value of the FTR score for a publication to remain in the final literature sample. Again, usually, only publications with the maximum score are included, as only these

fulfill all required eligibility criteria. However, defining a score threshold (cut-off criterion) might be suitable as well. For example, concept-building SLRs that aim at defining a concept might be confronted with the problem that almost all publications meet the categories because most publications include (short) sections that define pertinent concepts. For synthesis, however, it is more useful to include only publications that contribute substantially to the definition of a concept. In this case, a defined amount (e.g., the upper third) of publications with the highest FTR score (which indicated the number of, e.g., definitional statements) is retained for the analysis. This approach follows the premise that publications with a higher FTR score are particularly suitable to answer the research questions/hypotheses.

In the FTR stage, researchers conduct a second truncated content analysis based on the full texts, ensuring each included publication contributes to answering the SLR's research questions. Each publication is evaluated through FTR scoring based on clearly defined FTR categories. As a result, researchers end up with the final—transparent and reproducible—literature sample for the subsequent analysis and synthesis.

What if the dataset is too large to perform the FTR? Researchers' resources might be insufficient for reading all publications included during the ABS stage. In this case, we suggest drawing a random sample from the ABS publications to proceed with the FTR scoring technique. A randomized sample presents a cross-section of the research landscape under investigation. Since all relevant publications have an equal chance to be considered for the FTR sample (and therefore the later analysis), this does not violate the systematic approach of the SLR and, thus, presents a valid alternative to full suspension.

How can the subsequent analysis be prepared during the FTR stage? We recommend copying the text passages that were decisive for the coding into a data extraction form (e.g., a spreadsheet or an online survey; Petticrew & Roberts, 2006). Using additional software, such as survey tools, supports researchers to comfortably collect and synthesize data during the FTR stage, which reduces human labor costs and prepares the literature sample to be further processed during the phase of result synthesis by using automated content analysis. In addition, such software could not only be used to document the coding but also to automate the screening decisions (e.g., through IF/THEN conditions), which facilitates reproduction by future researchers.

3. Discussion

We introduced the STAMP method, a score-based sampling technique designed to enhance the reproducibility and replicability of SLRs by standardizing the literature selection procedure. Based on the principles of standardized content analysis, our proposed approach draws on the strengths of a genuine communication research method to offer advancements over preceding SLR methodologies, emphasizing aspects of scientific quality that are "seminal" for content analysis (Haim et al., 2023, p. 281) and add value to a diverse range of systematic research efforts.

3.1. Transparency and Reliability

Applying the STAMP method to SLRs improves *transparency* since each screening decision is systematically documented. In accordance with established approaches, our procedure also focuses on the review protocol (Nightingale, 2009), but deviates by integrating principles from standardized content analyses: The review

protocol not only documents all decisions made while designing the SLR. It also serves as a codebook that transfers eligibility criteria into a comprehensive category system with transparent coding guidelines. Consequently, the abstracts and full texts yielded by the systematic literature search are not merely used for heuristically deciding on a publication's inclusion—They are units of analysis, systematically coded with intersubjective categories. These categories, in turn, lead to scores that quantify the publications' relevance for the SLR. Through translating the selection process into two truncated content analyses during Stages 3 and 4, our approach also enforces standardized documentation, which enables assessing the reliability of the sampling. To be able to quantify the agreement of ratings from multiple coders is a fundamental prerequisite for replicability and reproducibility of content analysis (Krippendorff, 2004, p. 211). Hence, as for standardized content analysis, both intra-coder and inter-coder reliability scores can be computed for the ABS and the FTR stage, facilitating a reproducible and replicable sampling process.

To do so, each inter-coder *reliability* test should start with training, where all raters discuss the codebook. For example, an inter-coder reliability test could be conducted with two or three raters, which code around 10% of the dataset's publications (Petticrew & Roberts, 2006). After separate coding procedures, reliability coefficients can be calculated (for an example of inter-coder reliability assessment, see Joris et al., 2020). Both inter- and intra-coder reliability coefficients can be computed separately for each ABS/FTR category as well as for the overall ABS/FTR scores. We recommend calculating and reporting a variety of coefficients like Krippendorff's alpha, Cohen's kappa, and Holsti's coefficient (e.g., Kunze, 2024). For Krippendorff's alpha and Cohen's kappa coefficients, it should be considered that homogeneity (i.e., little variance) and dichotomy (binary coding decisions like *a criterion is met* [1]/*not met* [0]) of the data could negatively influence those coefficients. Besides these coefficient-specific annotations, reliability tests still identify potential coder effects and increase the clarity and intersubjectivity of the ABS/FTR categories. Employing coefficients brings the benefit of a comprehensible decision basis for including/excluding categories after pretesting for the complete coding procedure, for example, defining 0.8 as the acceptance threshold for content-related and 1.0 for formal categories (Lombard et al., 2002, p. 593).

Reproducibility and replicability require open reporting (Freiling et al., 2021). Accordingly, the method section should include the research question(s), eligibility criteria, sample sizes during Stages 2 to 4, and information on the coding process (how many coders were involved in the sampling, when and how coder training took place, and coefficients for intra- and inter-coder reliability). As suggested within the STAMP guideline template (https://bit.ly/4atGsvN), the SLR's appendix should additionally include the review protocol and the codebooks for Stages 3 and 4 (including definitions and examples for each category as well as the construction of the ABS and FTR scores). Ideally, ABS and FTR coding sheets with basic bibliographic information are also made accessible (for instance, via OSF).

3.2. Adaptability and Validity

In addition to the aforementioned advantages, one major methodological contribution of the STAMP method is its *adaptability*. STAMP is not only adaptable to different publication types—for the content analyses during Stages 3 and 4, it does not matter whether the studies are qualitative, quantitative, computational, or mixed-methods—but also to diverse types of SLRs (Paré et al., 2015) that are common for the interdisciplinary field of communication research. In our article, we repeatedly compared concept-building SLRs and evidence-summarizing SLRs. Our prior SLRs fit in that spectrum, demonstrating

the STAMP method's practicability. For instance, Rogge (2023), as a concept-building SLR, sought to unify a multifaceted term within an interdisciplinary research domain. This necessitated an inductive and iterative STAMP application: Every stage used standardized categories, but they were refined between Stages 3 and 4. In comparison, Anter (2023), within a well-defined research field and, therefore, at the evidence-summarizing end of the spectrum, utilized a rigidly deductive STAMP method with predefined criteria and fixed categories. Finally, our method remains fitting for intermediary SLRs such as the one by Kunze (2024), which combines both efforts. In this study, one research question aims to define a term (inductive, concept-building), while another investigates factors influencing the dependent variable (deductive, evidence-summarizing). Accordingly, the STAMP method was realized both with more fixed categories and categories based on an iterative reflection in relation to the available material.

In this sense, our approach also contributes to the *validity* of the literature sample: A thought-through and comprehensive codebook is the first step towards ensuring the validity of content analysis (Potter & Levine-Donnerstein, 1999, p. 266). Iteratively refining the eligibility criteria allows tailoring the respective categories to the unique material of analysis. At the same time, researchers are continuously forced to check the alignment of the sampled material with the SLR's objectives. Nevertheless, another necessary step to ensure validity is to assess the coding decisions against external standards (Potter & Levine-Donnerstein, 1999, p. 266). Examples of these standards include using relevant literature and expert consultations when designing and defining the SLR, as well as using pertinent literature for validating the search string.

3.3. Impartiality and Intersubjectivity

Especially for coding latent variables such as the content of a text, objectivity is "not a realistic expectation" (Potter & Levine-Donnerstein, 1999, p. 265). In these cases, it is all the more important to ensure the *intersubjectivity* and *impartiality* of coding decisions. Building on the standards of content analysis, STAMP includes multiple efforts to tackle potential biases in the sampling process. Among others, this is the persistent focus on eligibility criteria and their application by trained coders and on the basis of an established coding scheme (Stages 1–4), combining strengths of different databases in the search strategy (Stage 2), and employing scores instead of heuristic assessments for screening decisions (Stages 3 and 4). During conducting an SLR, biases can also arise through priming effects when prominent authors or popular journals are prioritized over lesser-known publications that could, however, equally or even more strongly contribute to answering the research questions. This is particularly problematic when such biases lead to systematic exclusion (and therefore, discrimination) of certain scientific publications—for example, by scholars from the Global South (Chakravartty et al., 2018). The STAMP method's hallmark is its blind assessment of the abstracts and full texts. We recommend disguising author names and—if compatible with the SLR's formal eligibility criteria—journal names in the reading phases (Stages 3 and 4). This enables blind sampling and prevents confirmatory tendencies or biased screening decisions. A paper is therefore judged based on its contribution to the research question, which means that every publication has the same chance of remaining in the sample as long as it fulfills the predefined and intersubjective assessment criteria. Accordingly, blind sampling increases impartiality and further reduces the risk of citation circles (continuously referring to the same authors within a particular topic), which would significantly limit the validity and intersubjectivity of the SLR's results.

4. Conclusion

SLRs are essential in synthesizing the abundance of available knowledge in the field of communication and related disciplines, building and further developing theoretical concepts, and consolidating empirical insights. Yet, the shortcomings of SLRs in our field often lie in their transparency, reproducibility, and replicability, challenging the foundation upon which syntheses are built. Recognizing gaps in existing methodologies, we proposed the STAMP method, a score-based sampling technique rooted in the principles of a method genuine to communication science: standardized content analysis. STAMP's methodological contribution in extension to existing SLR procedures is constituted in quantifying screening decisions and standardizing the sampling process. Further, several strengths characterize the method: transparency and reliability through full documentation and score-based screening decisions, adaptability to different SLR types, sustained validity evaluation, and measures towards increasing impartiality and intersubjectivity. The novelty of STAMP lies not in a sophisticated protocol or complex search strategies, but in a four-step model that guides and standardizes the sampling process for SLRs in a comprehensive and practical way. Moreover, since STAMP is rooted in a method genuine to communication science, it contains nothing alien to communication researchers, but brings together common methodological competencies with regard to systematic reviews and is, therefore, easy to adopt by communication scholars.

Many of the components of the STAMP method are already established. This applies not only to the definition of inclusion and exclusion criteria, as described above. Review protocols are already used in various SLRs in communication research as well (e.g., Melchior & Oliveira, 2022). Moreover, various SLRs in our discipline also include a precise description of the search string and the search results (e.g., Belair-Gagnon & Steinke, 2020) or distinguish between abstract and full-text screening (e.g., Ratcliff et al., 2022). However, due to a lack of publicly accessible documentation, replicability and reproducibility are often limited to single parts of the sampling process. Replicating and reproducing the whole sampling procedure is only guaranteed by a systematic and transparent approach as proposed by STAMP. Thus, applying STAMP for SLRs in our field would introduce an additional instance of control for these studies and their particular validity claim.

Of course, the STAMP method also comes with some limitations. Most obviously, while it ensures a reliable and valid sample construction, it only partially influences the reliability and validity of the synthesis itself: A proper sample is a necessary prerequisite for reliable and valid analysis but is not sufficient on its own. Additionally, as with every content analysis, the method is prone to coder effects. Consider, for example, learning effects: As coders become more experienced, their knowledge of the respective research field increases. Consequently, their coding decisions might become more elaborate, which would be detrimental to the validity as well as inter- and intra-coder reliability. Finally, the STAMP method is costly, as it includes detailed documentation, developing and pretesting codebooks and—if several coders are involved—training and supervising coders.

Notwithstanding these shortcomings, the STAMP method helps researchers to further increase the transparency and, ultimately, the reproducibility and replicability of SLRs. Via OSF, we provide a detailed yet concise guideline that encapsulates the four stages of STAMP and includes practical examples for inductive, deductive, and mixed SLR approaches.

Acknowledgments

The authors wish to thank the three anonymous reviewers and the academic editors of this thematic issue for their valuable and constructive comments.

Funding

The article processing charges were funded by the joint publication funds of the TU Dresden, including Carl Gustav Carus Faculty of Medicine, and the SLUB Dresden as well as the Open Access Publication Funding of the DFG. No further funding was received for conducting the research presented in this article.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article is available online at https://bit.ly/4atGsvN

References

Anter, L. (2023). How news organizations coordinate, select, and edit content for social media platforms: A systematic literature review. *Journalism Studies*. Advance online publication. https://doi.org/10.1080/1461670X.2023.2235428

Bartels, E. M. (2013). How to perform a systematic search. *Best Practice & Research Clinical Rheumatology*, *27*(2), 295–306. https://doi.org/10.1016/j.berh.2013.02.001

Belair-Gagnon, V., & Steinke, A. J. (2020). Capturing digital news innovation research in organizations, 1990–2018. *Journalism Studies*, *21*(12), 1724–1743. https://doi.org/10.1080/1461670X.2020.1789496

Castells-Fos, L., Pont-Sorribes, C., & Codina, L. (2023). Decoding news media relevance and engagement through reputation, visibility and audience loyalty: A scoping review. *Journalism Practice*. Advance online publication. https://doi.org/10.1080/17512786.2023.2239201

Chakravartty, P., Kuo, R., Grubbs, V., & McIlwain, C. (2018). #CommunicationSoWhite. *Journal of Communication*, *68*(2), 254–266. https://doi.org/10.1093/joc/jqy003

Davis, J., Mengersen, K., Bennett, S., & Mazerolle, L. (2014). Viewing systematic reviews and meta-analysis in social research through different lenses. *SpringerPlus*, *3*(1), Article 511. https://doi.org/10.1186/2193-1801-3-511

Engelke, K. M. (2019). Online participatory journalism: A systematic literature review. *Media and Communication*, *7*(4), 31–44. https://doi.org/10.17645/mac.v7i4.2250

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *The FASEB Journal*, *22*(2), 338–342. https://doi.org/10.1096/fj.07-9492LSF

Freiling, I., Krause, N. M., Scheufele, D. A., & Chen, K. (2021). The science of open (communication) science: Toward an evidence-driven understanding of quality criteria in communication research. *Journal of Communication*, *71*(5), 686–714. https://doi.org/10.1093/joc/jqab032

Gambo, S., & Özad, B. O. (2020). The demographics of computer-mediated communication: A review of social media demographic trends among social networking site giants. *Computers in Human Behavior Reports*, *2*, Article 100016. https://doi.org/10.1016/j.chbr.2020.100016

Haim, M., Hase, V., Schindler, J., Bachl, M., & Domahidi, E. (2023). (Re)establishing quality criteria for content analysis: A critical perspective on the field's core method. *SCM Studies in Communication and Media*, *12*(4), 277–288. https://doi.org/10.5771/2192-4007-2023-4-277

Hase, V., Mahl, D., & Schäfer, M. S. (2023). The "computational turn": An "interdisciplinary turn"? A systematic review of text as data approaches in journalism studies. *Online Media and Global Communication*, *2*(1), 122–143. https://doi.org/10.1515/omgc-2023-0003

Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (Eds.). (2023). *Cochrane handbook for systematic reviews of interventions: Version 6.4*. Cochrane Training. https://training.cochrane.org/handbook

Humayun, M. F., & Ferrucci, P. (2022). Understanding social media in journalism practice: A typology. *Digital Journalism*, *10*(9), 1502–1525. https://doi.org/10.1080/21670811.2022.2086594

Joris, G., De Grove, F., Van Damme, K., & De Marez, L. (2020). News diversity reconsidered: A systematic literature review unraveling the diversity in conceptualizations. *Journalism Studies*, *21*(13), 1893–1912. https://doi.org/10.1080/1461670X.2020.1797527

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. SAGE.

Kümpel, A. S., Karnowski, V., & Keyling, T. (2015). News sharing in social media: A review of current research on news sharing users, content, and networks. *Social Media + Society*, *1*(2). https://doi.org/10.1177/2056305115610141

Kunze, D. (2024). *Systematizing destigmatization in the context of media and communication: A systematic literature review*. Manuscript in preparation.

Lombard, M., Snyder-Duch, J., & Bracken, C. C. (2002). Content analysis in mass communication: Assessment and reporting of intercoder reliability. *Human Communication Research*, *28*(4), 587–604. https://doi.org/10.1111/j.1468-2958.2002.tb00826.x

Melchior, C., & Oliveira, M. (2022). Health-related fake news on social media platforms: A systematic literature review. *New Media & Society*, *24*(6), 1500–1522. https://doi.org/10.1177/14614448211038762

Mishra, S., Satapathy, S. K., & Mishra, D. (2009). Improved search technique using wildcards or truncation. In R. Raghavan (Ed.), *2009 International Conference on Intelligent Agent & Multi-Agent Systems*. IEEE. https://doi.org/10.1109/IAMA.2009.5228080

Nightingale, A. (2009). A guide to systematic literature reviews. *Surgery*, *27*(9), 381–384. https://doi.org/10.1016/j.mpsur.2009.07.005

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *British Medical Journal*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183–199. https://doi.org/10.1016/j.im.2014.08.008

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences: A practical guide*. Wiley. https://doi.org/10.1002/9780470754887

Pigott, T. D., & Polanin, J. R. (2020). Methodological guidance paper: High-quality meta-analysis in a systematic review. *Review of Educational Research*, *90*(1), 24–46. https://doi.org/10.3102/0034654319877153

Pirkis, J., Rossetto, A., Nicholas, A., Ftanou, M., Robinson, J., & Reavley, N. (2019). Suicide prevention media campaigns: A systematic literature review. *Health Communication*, *34*(4), 402–414. https://doi.org/10.1080/10410236.2017.1405484

Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, *27*(3), 258–284. https://doi.org/10.1080/00909889909365539

Ratcliff, C. L., Wicke, R., & Harvill, B. (2022). Communicating uncertainty to the public during the Covid-19 pandemic: A scoping review of the literature. *Annals of the International Communication Association*, *46*(4), 260–289. https://doi.org/10.1080/23808985.2022.2085136

Rogge, A. (2023). Defining, designing and distinguishing artificial companions: A systematic literature review. *International Journal of Social Robotics*, *15*, 1557–1579. https://doi.org/10.1007/s12369-023-01031-y

van den Akker, O. R., Peters, G.-J. Y., Bakker, C. J., Carlsson, R., Coles, N. A., Corker, K. S., Feldman, G., Moreau, D., Nordström, T., Pickering, J. S., Riegelman, A., Topor, M. K., van Veggel, N., Yeung, S. K., Call, M., Mellor, D. T., & Pfeiffer, N. (2023). Increasing the transparency of systematic reviews: Presenting a generalized registration form. *Systematic Reviews*, *12*, Article 170. https://doi.org/10.1186/s13643-023-02281-7

van de Schoot, R., de Bruin, J., Schram, R., Zahedi, P., de Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., Harkema, A., Willemsen, J., Ma, Y., Fang, Q., Hindriks, S., Tummers, L., & Oberski, D. L. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature Machine Intelligence*, *3*, 125–133. https://doi.org/10.1038/s42256-020-00287-7

About the Authors

Ayanda Rogge (MA, TU Berlin) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her dissertation focuses on human–machine communication and social robotics. In her project, she explores the social design of artificial companions, with a particular interest in their communication and interaction characteristics allowing users to perceive a technical agent as an artificial companion.

Luise Anter (MA, LMU Munich) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. In her dissertation, she explores how the characteristics of social media platforms shape journalistic production processes. Her other research interests include the perception and use of news and information in social media environments, news bias, and the interaction between journalists and their sources.

Deborah Kunze (MA, Leipzig University) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her research interests focus on the fields of health communication (e.g., the promotion of cancer prevention) and media psychology (e.g., effects on stigmatizing attitudes). In her dissertation project, she explores how destigmatization can be fostered, with a focus on the potential of media and communication in this context.

Kristin Pomsel (MA, TU Dresden) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. Her research interest focuses on political communication processes, media literacy, and media trust. Her dissertation project deals with recipients' expectations and perceptions of news quality as well as the actual quality of alternative and established German news media.

Gregor Willenbrock (MA, HMTM Hanover) is a researcher and PhD student at the Institute of Media and Communication at TU Dresden. His research interests include exploring the dynamics of online communities and their collaborative endeavors in peer production, specifically within the context of the networked public sphere. He is also dedicated to investigating the use of computational methods in communication science research, extending their application beyond the scope of his current project.

Media and Communication
2024 • Volume 12 • Article 7971
https://doi.org/10.17645/mac.7971

cogitatio

ARTICLE

Open Access Journal

Direct Replication in Experimental Communication Science: A Conceptual and Practical Exploration

Ivar Vermeulen [1], Philipp K. Masur [1], Camiel J. Beukeboom [1], and Benjamin K. Johnson [2]

[1] Communication Science, Vrije Universiteit Amsterdam, The Netherlands
[2] Department of Advertising, University of Florida, USA

Correspondence: Ivar Vermeulen (i.e.vermeulen@vu.nl)

Submitted: 8 December 2023 **Accepted:** 27 March 2024 **Published:** 19 June 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Replication is generally considered a keystone of the scientific enterprise. Unfortunately, in communication science, there is a lack of clarity on what a replication actually entails, and to what extent replicators may deviate from original studies. In order to support researchers in conducting, evaluating, and justifying the setup of replications of communication science experiments, we provide a taxonomy of replication types. We argue that researchers almost always need to adapt some elements of an original communication study to meaningfully replicate it. The extent to which deviations—ranging from mere updates to deliberate deviations and additions—are permissible, however, depends on the motivation behind conducting a replication study. We distinguish three basic motivations: verification of an original study's findings, testing the generalizability of an original study (which we further differentiate into the generalizability of study outcomes vs. theoretical claims), and extending an original study beyond the original goals. We argue that these motivations dictate what types of deviations are permissible and thereby determine the type of replication (i.e., direct, modified, and conceptual). We end with concrete recommendations for replicators: to specify the motivation to conduct a replication study and clearly label and justify any deviations from the original study for all study elements.

Keywords

communication science; conceptual replication; direct replication; replication; stimulus development

1. Introduction

Replication is generally considered a keystone of the scientific enterprise (e.g., Dienlin et al., 2021; Keating & Totzkay, 2019; McEwan et al., 2018; Nosek et al., 2022). Popper (1959/2002, p. 64) famously noted that "non-replicable single occurrences are of no significance to science." In other words, the credibility of a scientific claim increases when a finding is repeatedly shown under sufficiently similar circumstances using sufficiently similar procedures and materials, that is, if it is replicated.

The process of replication is also what we allude to when we say that "science is self-correcting." Whenever the findings of invalid research designs, poor measurement, faulty statistical analyses, selective use of data, HARK-ing, p-hacking, or other misguided practices enter our body of literature (John et al., 2012), we expect, in principle, that someday a replicator will come and correct the score. That said, we also know that correction may take a long while. Relative to studies we call "original work," replication studies are still scarce (Keating & Totzkay, 2019). The reasons are well-known: Replication studies are hard to get published (McEwan et al., 2018), not well-cited (Hardwicke et al., 2021), and sometimes regarded as unduly critical (Peterson & Panofsky, 2021) and second-rate work (Spellman, 2015).

There might also be another reason why researchers shy away from conducting replication studies: Often, after close inspection of the original work, it becomes unclear which elements of a study should and can be replicated. Should the physical setting be replicated? What if the stimulus is outdated? Should one reproduce faulty designs, too? Unreliable measures? The non-blind experimenter? The ad hoc outlier handling? And would a reviewer—especially when this person is (affiliated with) the original author—agree that what a replicator might consider a small deviation or improvement still does justice to the original study? Clearly, without a common disciplinary understanding of what a good replication entails, it becomes challenging to conduct and publish one.

To illustrate this challenge, members of the current author team were once found ineligible for a social science replication grant because they proposed to update the original newspaper articles that Hovland and Weiss used in their 1951 study about source credibility. The authors' line of thought was that these articles (e.g., about how TV could come to replace movie theaters) would not elicit the same responses in a current audience as they did in 1951 and thus should be adapted. They were told by the funding agency that, unfortunately, the call was for "direct" replications only and that any deviations proposed to the original materials, design, and procedures would render applications ineligible.

In this article, we argue that, in fact, the opposite is true: For most social science research, direct replications *require* deviations from the original study. This holds especially true for experimental studies in the field of communication science, where (mediated) stimuli and response patterns are almost by definition subject to temporal and cultural changes. At times, the impossibility of using the exact same material is rather obvious: Adolescents of today would not respond similarly to video games from the 90s, even if required systems would still be available. Music from decades ago does not elicit the same affective responses today. And stimuli that were once regarded as scandalous may now be considered mild. Yet, this does not mean that it is impossible to reliably replicate studies that relied on such materials.

The aim of this article is to help researchers who aspire to conduct and publish replications of experimental communication science studies decide which original study elements can be reasonably altered while still performing a faithful and meaningful replication of the original study. We aim to provide such researchers with argumentation, as well as a discipline-specific taxonomy, to help them substantiate how and why any methodological changes still facilitate replication of the original study and its claims and to systematically categorize their replication in light of these changes.

In order to pursue this goal, in the following, we present a systematic conceptual analysis of replication studies in experimental communication science. Unlike other conceptual analyses of replication studies (e.g., Brandt et al., 2014; LeBel et al., 2018; Steiner et al., 2019), our analysis will question *why* researchers would want to replicate a particular study, as a point of departure. We will review different aims that could motivate a replication study and show how these motivations logically imply decisions on whether or not to update elements of original studies. We will look in detail at study elements that could be considered for updating and to what extent updates change the status of a replication.

Interestingly, prior analytical work on replications (e.g., Asendorpf et al., 2013; Schmidt, 2009) generally does not include an explicit definition of replication studies. However, based on this prior work, we are able to define a replication study as a study that adopts elements from a specific previous study, in order to reassess this previous study and/or the theoretical mechanism it tested. This means that replications can be methodologically very similar to a specific original study (e.g., operationalizations and methods are fully adopted), but also somewhat dissimilar (e.g., only the hypotheses are adopted). This also means that replications by definition challenge previous work: If a replication shows different results than the original study, our perspective on the original study may (and probably should) change. And, although in the Popperian paradigm all empirical studies can be considered challenges to theoretical claims, replication studies are a special case because they challenge specific *findings* of (other) researchers. This may be one of the reasons why replication work is sometimes considered adversarial, while it is rarely intended to be (Peterson & Panofsky, 2021). Note that in this article we discuss replications—which include the collection of new data—and not reproductions (e.g., Asendorpf et al., 2013) or robustness checks (e.g., Nosek et al., 2022), which verify claims by reanalyzing original data.

In our analysis, we focus on the discipline of communication science and experimental studies. We first explain this choice below, using foundational experiments as examples.

2. Replicating Experiments in Communication Science

The field of communication science is characterized by a large variety of employed research methods, including qualitative analysis, cross-sectional surveys, content analysis, and experiments. As chiefly experimental studies can provide evidence for causal relationships, they play a fundamental role in substantiating theory. For example, empirical evidence for a very central theory such as cultivation theory (Gerbner, 1969)—explaining how representations of reality by fictional and non-fictional media systematically distort perceptions of the social world over time—was almost exclusively obtained through content analyses and longitudinal surveys regarding people's media use and beliefs about crime and other social problems (Morgan & Shanahan, 2010). Because of a lack of experimental work, uncertainty persists about the precise causal mechanisms at play, and the explanatory power of the theory remains subject to debate (Potter, 2014).

Over the last two decades, communication science increasingly recognized the fundamental role of experiments and showed a stronger interest in experimental work (Rains et al., 2020). Some of the resulting studies are currently among the most cited in the field. However, few efforts are being made to replicate them (Keating & Totzkay, 2019; for a notable exception, see the special issue edited by McEwan et al., 2018). In 2018, the authors of the current article conducted a survey among 171 communication scholars, asking which experiments within the discipline of communication science were most in need of replication. The result was a list of 10 studies shown in Appendix I of the Supplementary Material. Most readers studying or teaching communication science will acknowledge these studies for their centrality to the field and will agree that attempts to replicate them would strengthen the discipline; yet few attempts have been made. We believe that if aspiring replicators have a better idea of how to explain their replication projects to readers, including precise reasons to conduct a replication and possible adaptations needed for fidelity, this will increase replication attempts in our field. Therefore, in the subsequent section, we review three typical motivations to replicate a study. We argue that the motivation for conducting a replication study directly guides decisions on how to approach potential alterations to the original study.

3. Motivations to Conduct Replication Studies: Verification, Generalizability, and Extension

Why would researchers want to conduct a replication study? First and foremost, researchers may simply regard replication as a fundamental aspect of the scientific process, a way to build on existing knowledge by reconfirming or refuting previous findings. More specifically, researchers may want to identify the boundary conditions under which particular effects occur—whether certain factors moderate or limit the generalizability of the findings, thereby providing a more nuanced understanding of the phenomenon. At other times, researchers may have doubts about the results of the original study, or they wonder whether particular research design choices may have unintentionally influenced the results.

In many replication studies, the motivation that drives replication remains unspecified or is expressed as a combination of all the reasons above. In this section, we argue that it is important to a priori specify the motivation behind conducting a replication, because the degree to which replicators can deviate from the original studies' setup is contingent on the motivation. We distinguish three basic motivations to conduct a replication study: *verification* of an original study, testing the *generalizability* of an original study (which we further differentiate into the generalizability of study outcomes vs. theoretical claim), and *extending* an original study beyond the original goals.

3.1. Verification

Perhaps the most basic motivation to conduct a replication study is to verify the reliability and validity of previously reported findings. This motivation may stem from aiming to repeatedly demonstrate an effect. Although such verifying replications should be standard practice and a sign of a healthy, progressing field, individual replication attempts often elicit defensive responses from original authors or their affiliates (Peterson & Panofsky, 2021), who may feel peers are casting doubt on the original study's correctness or validity. Surely, there can be reasons why someone could be legitimately doubtful of a study or of its results: counter-intuitive findings, unusually strong effects, small sample sizes, many barely significant findings, the use of non-standard or strongly shortened measurement instruments, ad hoc or excessive outlier removal,

seemingly unplanned comparisons between experimental conditions, a lack of descriptive data, or a lack of transparency overall, to name a few (cf. John et al., 2012). In case of such doubt, a primary, and probably healthy, response would indeed be to verify whether the results of such a study could be replicated. If the original results are based on a methodological error or a questionable analysis choice, they should likely not be replicated.

In light of this, it is not surprising that replication studies which aim to verify an original finding will often be heavily scrutinized for any design or operational deviation that might potentially explain a different result. For that reason alone, a replication spurred by a verifying motivation generally requires strict obedience to the original design and procedures. Nonetheless, as we will argue below, even a verifying replication study may need to introduce slight deviations in order to be valid.

3.2. Testing Generalizability

A second motivation to conduct a replication study could be to test whether study results can be generalized. First, the issue of generalization can pertain to the original study directly: the question of whether the observed results may have been contingent on particular choices that the original researchers made in terms of study design and methodology. Can results be replicated if a particular study element—e.g., the control condition, or the dependent measure—is deliberately modified? Second, the issue of generalizability can concern the theoretical claim that was studied. Can the theoretical claim be confirmed when it is tested in a different way, in different contexts, using different stimuli and different research populations? These two aspects are specified as follows:

Study generalizability has in common with the verification motivation that it reassesses the original study rather than the theoretical mechanism the study aims to test. The focus of such replications is on whether the original study's findings can be replicated with deliberately modified study elements, e.g., using slightly different stimuli, measures, or settings (in a different country, in a different language, in the field or a lab). However, such replication studies—which we will label "modified replications"—are not fueled by a motivation to exactly verify the original results but rather by an interest in testing the robustness of the original findings.

If such a modified replication study yields similar results as the original study, it provides evidence that the results are robust and generalize to the modified methodological conditions. If it yields different results, it is suggestive of potential boundary conditions of the original finding. Formulated more skeptically, modified replication attempts may help clarify whether original findings are explainable by idiosyncratic stimuli, measures, procedures, or contexts, instead of a generalizable theoretical model. From a technical perspective, testing whether a study hinges on specific study elements requires that replicators abide by original design and procedural choices closely and only deviate regarding these specific elements of interest (see also Steiner et al.'s, 2019, causal replication framework). Nonetheless, as was the case for verifying replications, we will argue that such close adherence to the original study may mean that, in practice, some study elements may have to be slightly updated.

Theory generalizability as a motivation entails that the focus is less on the original study's methodology but rather on the causal mechanism it tested. Thus, the focus is on retesting the hypotheses of the original study

and whether these can be reconfirmed with *different* designs and methodological choices. As a result of this shift in focus, any deviation from the original study would seem permissible as long as the tested hypotheses are still the same. In fact, in such studies—often dubbed "conceptual replications"—adherence to original design choices can be regarded as a limitation: If for some reason the original methodological choices, and not the focal theoretical mechanism, drive the effects found, studies staying too close to the original cannot corroborate the generalizability of a theoretical claim. Still, a genuine challenge in conceptual replications is to determine whether deviations from the original study (e.g., new stimuli or new measures) activate and measure the same theoretical concepts and mechanisms.

3.3. Extension

A third motivation of replicators could be to extend the original work. Such extensions are complementary to any of the three types of replications listed above. One can extend a verification-motivated replication by adding, for example, a quasi-experimental moderator, mediator, manipulation check, or dependent variable—as long as the addition is inconsequential for the verification purpose (i.e., the original procedures)—to find out more about the mechanisms driving the original study's effects. One can do the same for modified or conceptual replications aimed at testing the generalizability of a study or a theoretical claim. In the latter, one may introduce extra stimuli and measures in the procedure if methodologically viable. As such, the extension motivation by itself does not determine to what extent changes in original study elements are permissible—This is determined by the motivations of verification or testing generalizability. Instead, the extension motivation is decisive in whether to *add* elements to an original study.

All in all, in this section we argued that the motivation to conduct a replication study relates to different types of replication studies and to different degrees that replicators can afford to diverge from the original studies' design and procedures. In the following section, we will discuss whether modifying particular study elements should be considered a deliberate deviation or merely an update.

4. Deliberate Deviation or Mere Update?

In the previous section, we argued how replicators' motivations may warrant deviations from original methodological choices. But what exactly are deviations? What if a replicator translates measures and stimulus materials to make them accessible to a different population or to make them applicable in a new culture or country? What if measures and stimuli are only updated to make them fit current times? When is such a translation or update permissible, or perhaps even necessary? In order to specify this, we will, in the following, distinguish different types of adaptations to an original study: a mere update, an inconsequential deviation, and a deliberate deviation from the original study. The argument we make is this: Verification is still possible with updated materials and also with—carefully justified—inconsequential deviations. Deliberate deviations, however, are associated with a different motivation (i.e., testing generalizability) and replication types (i.e., modified and conceptual replications). We will discuss these types of deviations for seven elements of experimental (communication science) studies: sample, setting, procedure, design, measures, stimuli, and analyses.

4.1. Reproducing the Sample

An element of an original study that a replication study needs to alter, almost by definition, is the study sample. Usually, an experiment requires a sample of naive participants, which means that, even if it were practically possible, it is not sensible to re-recruit members of an original sample for a replication study. Next to no longer being blind to the study's purpose, the original sample is also no longer "original" in its sample characteristics as its average age or education level may have changed. Instead, replicators may aim to reproduce sample *features*, for example, gender, nationality, ethnicity, political, religious or cultural views, or social group membership.

In communication science experiments, such sample features may or may not be important in explaining participants' responses to experimental stimuli in a replication. In many studies, sample features are considered coincidental and of peripheral importance to the mechanism being tested. Nonetheless, verification-motivated replicators may wish to reproduce the features of the original study's sample. The goal would then be to recruit a similar group of participants to avoid the possibility that a different outcome emerges due to some unknown population characteristics. However, if sample features are considered theoretically irrelevant in the original study, employing a sample with a priori different features (e.g., resulting from a systematically different recruitment method) may even be permissible for replications with a verification motivation. We refer to such deviations as an *a priori inconsequential deviation*. For instance, McGuire and Papageorgis (1961) used students as participants in their famous inoculation experiment, but no theoretical reason was mentioned for this choice. In such cases, replicators may deem it appropriate to resort to a different sample. Also, if different sample features are observed a posteriori (e.g., observing a different average age or education level despite aiming to reproduce original sample features), this can generally be considered as an inconsequential deviation.

In other original studies, (sub)sample features are part of hypothesizing and study design. Iyengar and Hahn (2009), for example, studied differences in media selection in Democratic versus Republican voters. Replicators of such studies, driven by verification or study generalizability motivations, should use subsamples with similar features. However, in modified or conceptual replications, deliberate deviations from the original sample can be used to test whether findings generalize to different populations (e.g., recruit voters of other political parties in another country).

4.2. Reproducing the Research Setting

A replication research setting is also a study element that will generally differ from an original study. Research setting involves features such as the era or period in which a study is conducted, cultural characteristics, geographical location, and also the physical location (e.g., in a research lab). In particular, era-related characteristics of a research setting can be entirely non-reproducible. If it is not feasible to reproduce the research setting itself, replicators, instead, could focus on closely reproducing the features of the research setting.

A changed research setting (e.g., a different country, era, or physical location) may involve translation to make methods fit the new setting. If the research setting is considered to play a peripheral role, such translations are permissible even for replication with a verification aim—For example, Fransen et al. (2024) replicated McGuire and Papageorgis' (1961) inoculation effects in the Netherlands. If deviations in setting

are observed or planned a priori, replicators with a verification motivation should explain and justify whether they consider such deviations inconsequential. For example, Gorn's (1982) experiments on the effects of music in advertisements were conducted in a lecture room. As this setting was not considered to be theoretically relevant, replication with a verification aim could be conducted in an online setting (Vermeulen & Beukeboom, 2016).

Research setting can also be a central part of an experimental design or research question; for example, Roozenbeek et al. (2020) studied the effectiveness of pre-bunking strategies against misinformation across countries. In such cases, replicators with a verification motivation should aim at reproducing features of the original research setting as closely as possible (thus conducting the replication in the same countries). Any alteration could constitute a consequential deviation that limits the verification goal.

In contrast, deliberate deviations in the research setting can be used to test the generalizability of a study beyond its original setting. For example, the negation bias in stereotype communication was first tested in Dutch (Beukeboom et al., 2010) and later replicated in five other European languages (Beukeboom et al., 2020).

4.3. Reproducing the Procedure

Especially in experimental research, research procedures may considerably influence study outcomes. Consider procedural features such as informed consent, participant instructions, measurement order, or the presence of manipulation checks. In most experiments, procedures have been meticulously designed precisely because experimenters know they may significantly influence participant responses. Replicators with a verification motivation should therefore reproduce original procedures as closely as possible. That may also mean, in case these are unclear or incompletely reported, that original procedures have to be reconstructed based on secondary information or reasoning (e.g., Vermeulen et al., 2014). The responsibility of a replicator with a verification motivation is to—possibly despite having to work with frustratingly limited information—aim to faithfully reproduce original procedures as closely as possible. In such cases, contacting the original authors may be helpful. Note, however, that the value of replication lies also in independent verification of initial findings and that original authors may (even unknowingly) be biased towards their own findings. In any case, if there are known a priori deviations, replicators with a verification motivation should explain and justify why they consider such deviations necessary or inconsequential.

In contrast, replicators motivated to test generalizability may typically focus on the procedure to make deliberate deviations. Given that experimental procedures are influential, it may be a valid concern whether original results stem from an unintended feature of the research procedure. Using deliberate procedure deviations in modified replications, this can be purposefully tested. Note that adding new elements to a procedure *after* the original procedure is finished is inconsequential and thus should not be considered a deviation. Adding such new elements can serve extension purposes.

4.4. Reproducing the Design

Original experimental research designs (i.e., the number of experimental factors, their levels, and between or within-subject manipulation) are usually unproblematic to replicate. Commonly, experimental designs are

well-documented and experimental factors are clearly explained. Replicators with a verification motivation should aim to use the same design as the original study.

It may, however, occur that the focus of a replication attempt concerns only a subset of the full design, for example, a particular comparison between conditions. Replicators—even those with a verification motivation—could then choose to adapt the original design by only including the experimental conditions necessary to replicate the original comparison: The omission of the other conditions is an inconsequential deviation. In a similar vein, replicators with an extension motivation may opt to add between-participant conditions without consequentially deviating from the original.

Importantly, whether or not design adaptations are consequential depends on the type of design. Between-participant designs tolerate adaptations (added or omitted conditions) much better than within-participant designs. Adaptations to within-participant designs are very often consequential as they affect a study's procedure and, therefore, potentially also its results. Replicators with a verification motivation should avoid such consequential deviations. Those with a generalizability motivation may however use deliberate deviations in experimental design to test whether findings generalize beyond the original design.

4.5. Reproducing the Measurements

Communication science journals often allow for extensive method sections (Berger et al., 2010), so dependent measures should generally be well-described. It also becomes increasingly normative to share the complete methodology (Dienlin et al., 2021). This makes it, in principle, feasible to reproduce original measures in a replication study. However, particularly in older publications, often not all items of a scale are reported. Nevertheless, replicators with a verification motivation should aim to reproduce original measures as closely as possible.

Yet, it is important to recognize that measures, like stimuli (see below), are subject to interpretation and may be perceived differently across eras, cultures, and countries. For example, in his experiment on cultivation, Shrum et al. (2011) measured television viewing through items like "I have to admit, I watch a lot of television." In a 2023 diversified media setting, it is very likely that this question will be interpreted differently. In case a replicator is convinced that a measure or a scale item does not capture the same latent concept as it did in the original study, it becomes questionable whether it should nonetheless be included in a replication. A more valid approach would be to aim at constructing a modern counterpart of that measure or item: an update/translation to the current time and cultural context. In fact, in such cases, updating outdated measures or items would be a requirement for a verifying replication.

4.6. Reproducing the Stimuli

Reproducing stimuli from original studies is a central challenge for replicators. Especially for experimental studies in communication science, which rely heavily on media messages as stimuli, it is often impossible to re-use original stimuli as they have lost their original relevance or meaning. When the aim is to verify an original study, updating or translating is then required. As an example, eight out of the 10 "cornerstone" experiments of communication science presented in the Supplementary Material (Appendix I) use media messages as stimuli. Media messages are almost guaranteed to evoke different responses in participants from another era or culture

than they originally did. We argue that in such cases, and when one is motivated to verify an original study, it is bad practice to nevertheless use the exact same (outdated) stimuli.

A first-hand example is a three-study replication by some of the current authors of Gorn's (1982) experiment testing music effects in advertising. In the first two—supposedly exact—replications, we used the same music for a 2013 Dutch sample as Gorn used for his 1980's Canadian sample (Vermeulen et al., 2014). The music effects—obviously, with hindsight—were different in 2013, but that taught us little about the replicability of Gorn's original findings. A third study that employed updated musical stimuli in fact yielded results similar to those of Gorn.

The question of how to adapt measures and stimuli in such a way that is likely to produce the same responses in current participants as in the original study, or to translate a stimulus in such a way that it will produce the same responses in participants from other cultural backgrounds, is highly complex. The basic idea is that stimulus *features* should be reproduced in a current cultural context, instead of the stimuli themselves (D. M. Slater et al., 2015). This is done by conducting a content analysis of stimulus features and then systematically re-creating these features in geographical, contextual, and era-congruent stimuli (M. D. Slater, 1991). In Appendix II of the Supplementary Material, we present a practical guide on how to do this for stimuli and measures as commonly used in communication science experiments. To make the task more tangible, we discuss different stimulus features: source features, message features, and channel features. Additionally, we discuss recipient state features, which are stimulus features used to induce a particular psychological state in recipients (e.g., task involvement), which in turn may change a response to another stimulus. We also explicitly discuss that in updating stimuli from a previous communication science experiment, replicators should not only focus on features in which stimuli (experimentally) differ but also on features in which they are the same (e.g., if both stimuli were realistic then, they should be realistic now).

4.7. Reproducing the Analyses

A commonly used approach to determine whether a replication confirms the outcomes of an original study is testing whether the observed effect sizes are sufficiently similar, i.e., whether the original effect size's point estimate is included in the replication's 95% confidence intervals (cf. Asendorpf et al., 2013; LeBel et al., 2018). Due to the extensive method sections common in communication science (Berger et al., 2010), replicators can often clearly identify what type of statistical model was used and how the effect of interest was estimated. However, exact data preprocessing steps (e.g., what exclusion criteria were used and how missing values were treated) are often not reported. This lack of transparency makes it difficult to follow the original analysis procedure in every detail. If the aim is to verify the original study's findings, replicators should attempt to replicate the original analysis procedure as closely as possible, which may involve taking steps such as contacting the original authors, in search of missing information. That said, many data preprocessing steps are arbitrary, and different choices could be considered valid (e.g., Simonsohn et al., 2020; Steegen et al., 2016). Hence, one could argue that certain aspects of the analyses need not be perfectly aligned with the original study (e.g., thresholds for outlier removal). Again, it is important to a priori deem such (minor) deviations as inconsequential in order to still conduct a verifying replication.

It is important to mention here that extension-motivated replicators can explore alternative ways to analyze the data without sacrificing the ability to verify the original study's findings. As long as the original analysis is

reproduced, one can additionally report alternative analyses without endangering a replication's verification goal. As an example, Vermeulen et al. (2014), in their replication of Gorn (1982), needed to replicate Gorn's in part faulty data analysis method to compare effect sizes but also presented an improved analysis.

In this section, we reviewed how replication studies can include adaptations and deviations from the original study. We argued that updating and translating measures and stimuli is often required if one is motivated to verify an original study's findings. Replication studies with a verification goal may also include inconsequential deviations (e.g., in sample, setting, procedures, and analyses) if these are clearly justified. In contrast, deviations may also be deliberately introduced with a motivation to test generalizability. In the next section, we will bring this together and distinguish how different motivations relate to different types of replications, which allow for different types of deviations.

5. A Taxonomy for Replications In Experimental Communication Science

Several authors have previously presented taxonomies of replication studies (e.g., LeBel et al., 2018; Lykken, 1968). Notably, Kelly et al. (1979) developed a taxonomy of replication types specifically for experimental communication research. After applying their taxonomy on replication studies in the discipline, they found that studies that had made alterations in the stimulus materials were by far the most prevalent. This finding corroborates our observation that in order to replicate communication science experiments, it is often necessary to update stimuli.

LeBel et al. (2018) introduced a taxonomy that orders replications according to their similarity with the original study. On a broader level, they distinguish *direct* from *conceptual* replication, arguing that they serve different epistemological purposes. Only replication types subsumed under *direct replications* (exact, very close, and close replications that only differ in contextual and procedural aspects) are regarded as sufficiently similar to an original study to be considered evidence for the original study's claim. Because of their methodological similarity to the original study, they allow for the falsification of a hypothesis and thereby question the credibility of an effect (LeBel et al., 2018; Meehl, 1978). Interestingly, LeBel et al. also consider replications that use different stimuli as sufficiently close if hypothesis, constructs, operationalization, and population characteristics are the same—In their taxonomy, this would be a close replication, the furthest away from the original study that can still be seen as a direct replication.

Conceptual replications, in LeBel et al.'s (2018) framework, are characterized by deliberately introduced differences in study elements, such as different constructs, operationalizations, samples, or stimuli. Unsupportive evidence from such studies cannot question the original study's finding because it is unclear whether it is a falsification of the original hypothesis or simply highlights contingency on particular study elements that were changed. As such, conceptual replications can only provide insights into the generalizability of presumably replicable effects or hypotheses.

Our taxonomy (Figure 1) builds on previous work, particularly LeBel et al. (2018) and Steiner et al. (2019), but also differs in two substantial regards: First, we explicitly include researchers' motivations to replicate a study and argue that these motivations automatically imply particular replication types. Second, we explicate the difference between mere translations/updates, inconsequential deviations, and deliberate deviations. In frameworks such as LeBel et al. (2018), replication types are only differentiated by the amount

of study elements that differ. As a result, our taxonomy better facilitates replication researchers to evaluate and report the setup of their replication studies. Also, our taxonomy differs by integrating direct (aiming to verify by staying methodologically close to the original), modified (systematically exploring one or few deviations), and conceptual ([dis]similar studies testing the same hypotheses) replications in one framework.

Similar to prior taxonomies, we order replication types according to their similarity with the original study. Any individual replication study may be placed on this similarity continuum and may occasionally fall between those exemplary types that we highlight and discuss in the following. On the highest level, we differentiate direct, modified, and conceptual replications. We argue that these three types of replications serve different epistemological purposes, which can be expressed as researchers' motivations to conduct a replication (see our discussion of motivations above).

highly similar ⟷ highly dissimilar

	Direct replication			Modified replication	Conceptual replication		
	Exact replication*	Very close replication	Close replication	Proximate replication	Far replication	Very far replication	
Motivation	Verification			Generalizability			Extension
Goal	Verifying the original study's finding using the same or sufficiently similar methodology			Testing whether the original study's finding is contingent on particular study elements	Testing the overall generalizability of the theoretical claim		Extending the original study design and goals
Hypothesis	same	same	same	same	same	same	same and additional
Constructs *(theoretical constructs under investigation)*	same constructs	same constructs	same constructs	same constructs	same constructs	any valid constructs	
Experimental design *(e.g., between-vs. within-design, levels of manipulation, ...)*	same design	same design	same design	similar to a very close or close replication, except for focal, deliberate deviations in study element(s)	any valid design	any valid design	Depending on motivation/goal: Verification inconsequential additions to the original design Generalizability Any valid additions to the original design
Measures *(operationalization of instruments)*	same operationalization	same or updated/translated operationalization	same or updated/translated operationalization		any valid operationalization	any valid operationalization	
Stimuli *(e.g., message features, source/channel features, situational features, recipient features, ...)*	same stimuli	same or updated/translated stimuli	same or updated/translated stimuli		any valid stimuli	any valid stimuli	
Sample *(e.g., population features such as age, gender, distribution, education ...)*	similar sample (all characteristics are the same)	similar sample (relevant characteristics similar)	a priori inconsequential deviations		any valid sample	any valid sample	
Setting *(e.g., physical location, era/time, cultural characteristics, ...)*	same setting	similar setting (relevant aspects similar)	a priori inconsequential deviations		any valid setting	any valid setting	
Procedure *(e.g., instructions, order of elements, ...)*	same procedure	similar procedure (relevant aspects similar)	a priori inconsequential deviations		any valid procedure	any valid procedure	
Analysis *(e.g., type of data analysis approach, outlier removal, imputation ...)*	same analysis	same analysis	a priori inconsequential deviations		any valid analysis	any valid analysis	

Figure 1. A taxonomy for experimental replications in communication science. Notes: It is important to note that the classification does not provide any evaluation, all types of replications have their value in the general research endeavor; the taxonomy was inspired by, and builds on, the taxonomy proposed by LeBel et al. (2018); * in communication science, exact replications are almost always impossible to conduct, as media stimuli (and measures) require updates in order to be meaningful, and changes in sample, setting, and procedure are often inevitable.

If the motivation is verification, i.e., whether or not an original finding can be replicated with the same or at least sufficiently similar methodology, researchers need to make sure that their replication attempt indeed qualifies as a direct replication of the original study. Such replications generally require strict obedience to the original study elements or at least need to be sufficiently similar. Based on the nature of deviations, direct replications can be further distinguished into exact, very close, and close replications. Exact replications keep all study elements exactly the same. Although theoretically conceivable, exact replications are rarely, if at all, possible in (experimental) social science (Nosek & Errington, 2020) and particularly in communication science. As discussed earlier, it is almost always inevitable that a replication study deviates from one or more study elements. In the case of communication science experiments, particularly stimuli are contingent on time and context and easily outdated. In "very close" replications, such study elements are *updated* in order to enable a valid replication of the original study. Note that when we discuss dissimilarities, we purposefully differentiate between updates, (a priori) inconsequential deviations, and deliberate deviations (a difference compared to prior frameworks). If the aim of a replication is to verify the original findings (i.e., a direct replication), only necessary updates are allowed. As argued earlier, a stimulus from the 1950s may not elicit the same responses in participants today. Yet, it may be possible to create an updated stimulus that is context- and time-appropriate and elicits the same response in participants today as the "old" stimulus did in participants in the 1950s.

Another distinction within direct replications is expressed in the difference between a very close and a close replication. Whereas a close replication aims to adhere to the original design and methodology as much as possible and only introduces necessary updates to a few study elements, a close replication also allows deviations that are a priori deemed inconsequential to still allow for verification of the original study's claim. We purposefully limit these to the study elements sample, setting, procedure, and analysis, as those are likely to differ slightly in many verification-motivated replication attempts. Whether a deviation is a priori inconsequential may be hard to justify. A starting point may be to investigate the specificity of the original hypothesis and the original study's claim. For example, did the original hypotheses make auxiliary assumptions relating to these particular study elements explicit? Is the theoretical claim clearly limited to a particular sample, procedure, or context? Did the original study somehow explicate boundary conditions relating to the particular sample, setting, or procedure originally implemented? Overall, we argue that replications that only introduce justifiably inconsequential deviations, while keeping the same or merely updating all other core elements, can still be regarded as direct replications aimed at verifying the original finding.

Like prior authors (e.g., Hendrick, 1990; Steiner et al., 2019), we include a replication type that does not fall under direct replications nor under conceptual replications but sits somewhere in between. Such modified, but still proximate, replications—which are quite prevalent—are conducted when a replicator is no longer aiming at verifying an original study's claim but investigates the generalizability of an original study in light of particular methodological choices that were made. In other words, such replications test whether the original study's finding is contingent on particular methodological aspects. Such studies can be regarded as *modified replications*—deliberately varying one or more study elements, while keeping the same or merely updating all others. For example, a research team could be interested in testing whether the original findings still hold if a modified dependent measure or a modified stimulus is used (i.e., study generalizability). Clearly, such replications do not qualify as direct because they intentionally introduce deviations from the original study. Yet, they should be distinguished from conceptual replications, which are focused on merely testing

the same hypothesis using any valid, and potentially more different, methodology. For modified replications, the focus remains on the original study itself and on potential methodological questions that it raised. When modified replications yield the same result as the original study, they provide evidence that the findings generalize to the modified methodological element. In contrast, a failed modified replication provides a first indication of the boundary conditions of the original finding.

A third type of replication, subsumed under the label *conceptual replication*, aims at testing the generalizability of a theoretical claim (i.e., theory generalizability) while implementing dissimilar methodologies compared to the original study. The focus thus is on retesting the hypotheses of the original study (i.e., *theory generalizability*). In such studies, any deviation from the original study is permissible as long as they are theoretically valid and the tested hypotheses are still the same (LeBel et al., 2018). In fact, in order to prove the generalizability—not just the verifiability—of a theoretical claim, deviations compared to the original study are necessary. We distinguish *far* replications, where the theoretical constructs are still the same and the replication thus re-tests the original study's theoretical claims exactly, from *very far* replications, where theoretical constructs may be slightly varied to test the original study's claims in a broader sense.

Finally, our taxonomy also includes a motivation that can be combined with all three (direct, modified, conceptual) types of replications: extension. With an extension motivation, the goal is to gain additional insights compared to the original study. Motivated by extending and better understanding the original study's findings, replicators can add methodological aspects (e.g., additional measures, experimental groups) to their replication study. As already argued, however, the core motivation (i.e., verification, study generalizability, theoretical generalizability) may limit what type of additions are allowed.

Our taxonomy facilitates researchers (including replicators, reviewers, and peers) to conduct, evaluate, and justify the setup of replications of communication science. We urge replicators to (a) clearly identify their motivations to conduct a replication and, in line with their motivation, (b) justify any updates or deviations to the original study's design. We acknowledge that the types of replications that we emphasized as exemplary oversimplify the range of potential replications. That said, we strongly believe that they will help researchers in placing replication attempts on the similarity continuum and, thereby, evaluate in what ways these replications can say something about the replicability of the original study's findings or the generalizability of the theoretical claim.

6. Conclusion and Discussion

In order to facilitate researchers to conduct, evaluate, and report the setup of replication studies, we provided a taxonomy of replication types. We argued that researchers almost always need to update and/or translate some elements of an original communication study to meaningfully replicate it, and we provided guidelines and examples as to how to approach and justify such updates and translations. We also discussed the difference between inconsequential and deliberate deviations from an original study. We posited that the extent to which a deviation is permissible depends on the motivation to conduct a replication study. Here, we distinguish three basic motivations: verification of an original study's findings, testing the generalizability of an original study (which we further differentiate into the generalizability of study outcomes vs. theoretical claims), and (in combination with one of the other three motivations) extending an original study beyond the original goals.

Because these motivations dictate what types of deviations are permissible, they also determine the type of replication (i.e., direct, modified, and conceptual).

6.1. Limitations and Challenges

Although we believe that our taxonomy helps in categorizing replication attempts and facilitates researchers to consider and justify what types of deviations are permissible depending on the motivation they have, we acknowledge certain limitations. First, the distinction between direct and conceptual replications may not always be as clear-cut as we suggest here. By introducing a third type of replication (modified), we already propose a more fine-grained differentiation, yet still different researchers may interpret, for example, the status of particular deviations differently, leading to ambiguity in the proposed categorization and potentially diverging opinions on what constitutes a replicated effect. In such cases, we hope our categorization will facilitate the scientific debate.

Second, we acknowledge certain constraints in conducting replications that cannot be solved even with a granular taxonomy. For example, certain large-scale experiments may be difficult to replicate due to their size and resource-intensiveness. Similarly, replicating certain experiments may pose ethical challenges, especially if the original study involved controversial or sensitive procedures.

Another challenge relates to how to deal with replication outcomes. It is likely that replication studies (across all types discussed above) produce diverse outcomes, including partial replications, variations in effect sizes, or even contradictory results. Making sense of these outcomes is not as trivial as it seems. A non-replicated finding could mean that the original study was a false positive, but it could also be that the replication is a false negative. In fact, cumulative evidence for a study's results requires several (direct) replications, a time-consuming and slow process that, at least currently, does not seem to be valued sufficiently in our field (Dienlin et al., 2021).

6.2. Recommendations for Replicators

Following our discussion and taxonomy, we recommend the following guidelines for researchers interested in conducting and publishing a replication study. First, clearly specify the motivation behind conducting your replication study: Do you want to (a) verify the original study's findings, (b) test the study's contingency on particular methodological choices, or (c) test the generalizability of the theoretical claim? Additionally, do you aim to (d) extend the original study in any way?

Second, specify as concretely as possible any deviations from the original study for all study elements (see Figure 1) and explain whether these are updates, inconsequential deviations, or deliberate deviations. In case of verification aim (direct replication), justify why the planned deviations are deemed necessary and why you deem them sufficiently similar or *inconsequential* (e.g., update due to a different era, translation to a new context). In case of focus on study generalizability (modified replication), again, justify any planned deviations but additionally specify which *deliberate* deviation(s) are made and what the aim of this deviation is (e.g., to test whether findings generalize to a different measure). In the case of focus on theory generalizability (conceptual replications), any theoretically valid deviation is permissible, yet we nonetheless urge you to discuss (dis)similarities to the original study.

Third, in case of a complementary extension motivation, specify which extensions you make and which further insights to the original study you aim to gain. Also specify the convoying motivation (verification, study generalizability, or theory generalizability) and justify the deviations and extensions accordingly (e.g., new measures are added after the original procedure).

Fourth and finally, diligently report the protocol of your replication study and make it available to reviewers and peers (e.g., as supplementary material and/or in a public repository). This will help others to (a) scrutinize your design choices, (b) conduct subsequent or even collaborative ("many-labs") replication efforts, and (c) specify their own design deviations more easily.

Acknowledgments

We would like to thank Emma Bryan for her notes on this manuscript.

Conflict of Interests

The authors declare no conflict of interests.

Supplementary Material

Supplementary material for this article can also be found at the Open Science Framework (https://osf.io/wbmh6).

References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Berger, C. R., Roloff, M. E., & Roskos-Ewoldsen, D. R. (2010). What is communication science? In C. R. Berger, M. E. Roloff, & D. R. Ewoldsen (Eds.), *The handbook of communication science* (pp. 3–20). SAGE.

Beukeboom, C. J., Burgers, C., Szabó, Z. P., Cvejic, S., Lönnqvist, J. E. M., & Welbers, K. (2020). The negation bias in stereotype maintenance: A replication in five languages. *Journal of Language and Social Psychology*, *39*(2), 219–236. https://doi.org/10.1177/0261927X19869759

Beukeboom, C. J., Finkenauer, C., & Wigboldus, D. H. (2010). The negation bias: When negations signal stereotypic expectancies. *Journal of Personality and Social Psychology*, *99*(6), 978–992. https://doi.org/10.1037/a0020861

Brandt, M. J., IJzerman, H., Dijksterhuis, A., Farach, F. J., Geller, J., Giner-Sorolla, R., Grange, J. A., Perugini, M., Spies, J. R., & van 't Veer, A. (2014). The replication recipe: What makes for a convincing replication? *Journal of Experimental Social Psychology*, *50*, 217–224. https://doi.org/10.1016/j.jesp.2013.10.005

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

Fransen, M. L., Mollen, S., Rains, S. A., Das, E., & Vermeulen, I. (2024). Sixty years later: A replication study of McGuire's first inoculation experiment. *Journal of Media Psychology*, *36*(1). https://doi.org/10.1027/1864-1105/a000396

Gerbner, G. (1969). Toward "cultural indicators": The analysis of mass mediated public message systems. *AV Communication Review*, *17*(2), 137–148. https://doi.org/10.1007/BF02769102

Gorn, G. J. (1982). The effects of music in advertising on choice behavior: A classical conditioning approach. *Journal of Marketing*, *46*(1), 94–101. https://doi.org/10.1177/002224298204600109

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. (2021). Citation patterns following a strongly contradictory replication result: Four case studies from psychology. *Advances in Methods and Practices in Psychological Science*, *4*(3), Article 5. https://doi.org/10.1177/25152459211040837

Hendrick, C. (1990). Replications, strict replications, and conceptual replications: Are they important? *Journal of Social Behavior and Personality*, *5*(4), , 41–49.

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, *15*(4), 635–650. https://doi.org/10.1086/266350

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, *59*(1), 19–39. https://doi.org/10.1111/j.1460-2466.2008.01402.x

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. https://doi.org/10.1177/0956797611430953

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239. https://doi.org/10.1080/23808985.2019.1632218

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research: An analysis. *Human Communication Research*, *5*(4), 338–342. https://doi.org/10.1111/j.1468-2958.1979.tb00646.x

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, *70*(3), 151–159. https://doi.org/10.1037/h0026141

McEwan, B., Carpenter, C. J., & Westerman, D. (2018). On replication in communication science. *Communication Studies*, *69*(3), 235–241. https://doi.org/10.1080/10510974.2018.1464938

McGuire, W. J., & Papageorgis, D. (1961). The relative efficacy of various types of prior belief-defense in producing immunity against persuasion. *The Journal of Abnormal and Social Psychology*, *62*(2), 327–337. https://doi.org/10.1037/h0042026

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, *46*, 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Morgan, M., & Shanahan, J. (2010). The state of cultivation. *Journal of Broadcasting & Electronic Media*, *54*(2), 337–355. https://doi.org/10.1080/08838151003735018

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, *18*(3), Article e3000691. https://doi.org/10.1371/journal.pbio.3000691

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*(1), 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Peterson, D., & Panofsky, A. (2021). Self-correction in science: The diagnostic and integrative motives for replication. *Social Studies of Science*, *51*(4), 583–605. https://doi.org/10.1177/03063127211005551

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge. (Original work published 1959)

Potter, W. J. (2014). A critical analysis of cultivation theory. *Journal of Communication*, *64*(6), 1015–1036. https://doi.org/10.1111/jcom.12128

Rains, S. A., Keating, D. M., Banas, J. A., Richards, A., & Palomares, N. A. (2020). The state and evolution of communication research: A topic modeling analysis of 20,000 journal article abstracts from 1918–2015. *Computational Communication Research*, *2*(2), 203–234.

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-008

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, *13*(2), 90–100. https://doi.org/10.1037/a0015108

Shrum, L. J., Lee, J., Burroughs, J. E., & Rindfleisch, A. (2011). An online process model of second-order cultivation effects: How television cultivates materialism and its consequences for life satisfaction. *Human Communication Research*, *37*(1), 34–57. https://doi.org/10.1111/j.1468-2958.2010.01392.x

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Slater, D. M., Peter, J., & Valkenburg, P. M. (2015). Message variability and heterogeneity: A core challenge for communication research. *Annals of the International Communication Association*, *39*(1), 3–31. https://doi.org/10.1080/23808985.2015.11679170

Slater, M. D. (1991). Use of message stimuli in mass communication experiments: A methodological assessment and discussion. *Journalism & Mass Communication Quarterly*, *68*(3), 412–421. https://doi.org/10.1177/107769909106800312

Spellman, B. A. (2015). A short (personal) future history of Revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science*, *11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Steiner, P. M., Wong, V. C., & Anglin, K. (2019). A causal replication framework for designing and assessing replication efforts. *Zeitschrift für Psychologie*, *227*(4), 280–292. https://doi.org/10.1027/2151-2604/a000385

Vermeulen, I., Batenburg, A., Beukeboom, C. J., & Smits, T. (2014). Breakthrough or one-hit wonder? Three attempts to replicate single-exposure musical conditioning effects on choice behavior. *Social Psychology*, *45*(3), 179–186. https://doi.org/10.1027/1864-9335/a000182

Vermeulen, I., & Beukeboom, C. J. (2016). Effects of music in advertising: Three experiments replicating single-exposure musical conditioning of consumer choice (Gorn 1982) in an individual setting. *Journal of Advertising*, *45*(1), 53–61. https://doi.org/10.1080/00913367.2015.1088809

About the Authors

Ivar Vermeulen is an associate professor at the Department of Communication Science at the Vrije Universiteit Amsterdam. He specializes in persuasion research. Most of his current research is experimental in nature and focuses on the effects of misinformation and the role of affective cues in persuasion processes. He published several replications of experimental communication science studies.

Philipp K. Masur is an assistant professor at the Department of Communication Science at the Vrije Universiteit Amsterdam. His research is concerned with applying socio-psychological and communication theories to study online communication. He focuses on privacy and self-disclosure processes in online environments, social influence and social norms on social media, and media literacy. He has conducted several replication studies.

Camiel J. Beukeboom is an associate professor at the Department of Communication Science at the Vrije Universiteit Amsterdam. His main research focus is on biases in language use and interpersonal communication. He studies how social category stereotypes and prejudice are expressed and communicated through language. In addition, he has studied social media use by individuals and organizations. He has conducted several replication studies.

Benjamin K. Johnson is an associate professor in advertising at the University of Florida. His research is focused on why and how people select and share messages in new media settings, especially as it relates to psychological processes such as impression management, social comparison, and self-regulation. He is the interim director of the University of Florida's STEM Translational Communication Center, an interdisciplinary research center focused on making complex science and health information as accessible and useful as possible to audiences.

Media and Communication
2024 • Volume 12 • Article 7666
https://doi.org/10.17645/mac.7666

cogitatio

ARTICLE

Open Access Journal

Attitudinal, Normative, and Resource Factors Affecting Communication Scholars' Data Sharing: A Replication Study

Jinghong Xu and Rukun Zhang

School of Journalism and Communication, Beijing Normal University, China

Correspondence: Rukun Zhang (202131021007@mail.bnu.edu.cn)

Submitted: 18 October 2023 **Accepted:** 22 February 2024 **Published:** 28 March 2024

Issue: This article is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

This study explores the factors affecting communication scholars' data-sharing intentions, a critical component of reproducibility and replicability in open science. We replicate Harper and Kim's (2018) study, which employs the theory of planned behavior to demonstrate the impacts of attitudinal, normative, and resource factors. Specifically, their original research examines data-sharing practices among psychologists, and our replication aims to reinforce their findings within the communication field. Data from a survey of Chinese communication scholars ($N = 351$) are analyzed using structural equation modeling. The findings indicate that perceived benefit and perceived risk significantly influence the attitudes of communication scholars towards sharing their data, positively and negatively, respectively. Additionally, attitudes, subjective norms, journal pressure, and the conditions facilitating data sharing have a significant positive impact on communication scholars' behavioral intentions. Perceived effort inversely affects attitudes toward data sharing but does not impact behavioral intentions. This study provides a theoretical framework for understanding data-sharing intentions and behaviors in the open science movement. The role of this research as a replication study serves as a compelling demonstration of scientific inquiry. Practical suggestions, such as fostering open dialog, institutional incentives, and cooperation between different actors to increase communication scholars' data-sharing intentions, and recommendations for carrying out replication and reproduction studies, are discussed. Finally, we judiciously reflect on the methodological limitations of our research and highlight directions for future research on open science.

Keywords

China; communication scholars; open science; replication study; structural equation modeling; theory of planned behavior

1. Introduction

The replication crisis, which suggested that many social science findings appeared to be unreliable, inspired the open science movement (Dienlin et al., 2021; Matthes et al., 2015). Like many disciplines, communication studies are also dealing with the promotion of openness (Bowman et al., 2021; Lewis, 2020). Following the publication of "An Agenda for Open Science in Communication" (Dienlin et al., 2021), many efforts have been proposed by various position articles to encourage and facilitate open communication science (OCS), such as the *Journal of Communication*'s special issue titled Open Communication Research (Shaw et al., 2021) and *Digital Journalism*'s special issue titled Analytical Advances Through Open Science (Haim & Puschmann, 2023).

Despite broad familiarity and support for OCS among International Communication Association (ICA) members, as evidenced by a recent survey ($N = 330$; Bowman et al., 2021), there remains a notable gap in actual engagement with OCS practices. While the potential legal and ethical challenges of OCS (Grand et al., 2012; Zhang et al., 2022) have been widely noted, we lack an understanding of these low levels of engagement with OCS practices. As one of the most salient aspects of open science, data sharing also faces the problem of low engagement (Tenopir et al., 2015; Vines et al., 2014; Zenk-Möltgen et al., 2018), which significantly hinders the reproducibility and replicability of communication research (Dienlin et al., 2021).

Hitherto, only limited research efforts have used empirical data to explore the factors that motivate communication scholars to share their data. To our knowledge, only Harper and Kim (2018) employed an empirical model based on the theory of planned behavior (TPB) to illustrate the elements that impact the willingness of American psychologists to embrace an open data badge. To address the research gaps in OCS practices within the field of communication, our study replicates the research design of Harper and Kim (2018). While their study focuses on the adoption of open data badges in psychology, exploring attitudinal, normative, and resource factors, our study extends this inquiry to the behavioral intentions of Chinese communication scholars in data sharing. This represents an empirical investigation of these factors through the lens of the TPB in a non-Western context, broadening the understanding of OCS practices globally.

2. Literature Review

2.1. Data Sharing and Its Role in Open Science

Although data sharing has been a component of scientific research for many years, recent technological advancements and the development of platforms such as the Open Science Framework have significantly enhanced the ease and scale at which data can be shared. This has led to an increased focus in the literature on data-sharing behaviors, particularly in the context of these modern platforms and the wider movement towards open science. Many such studies have identified the benefits of data sharing. For example, it has been claimed that data sharing can improve the reliability and robustness of communication scholarship (Dienlin et al., 2021) because the sharing of design protocols, measures, and analytic scripts can help improve the rigor of study designs and foster public trust (Banks et al., 2019). Researchers with interests that align can leverage existing scholarly work to either validate the results of prior studies or to explore new hypotheses, ultimately reducing costs, saving time with regard to data collection (Harper & Kim, 2018), and facilitating replication of studies and reproduction of analyses (Dienlin et al., 2021).

However, merely highlighting the benefits of sharing data does not provide a complete picture of healthy OCS. As de Oliveira et al. (2021) argued, OCS encompasses two distinct perspectives. The first perspective emphasizes principles such as acceleration, efficiency, and reproducibility, while the second perspective aims to foster participation, social justice, and the democratization of knowledge. The uneven development of OCS is what Dutta et al. (2021) called "hegemonic open science," which emerges from knowledge production systems in the Global North and primarily serves the economic interests of platform capitalism. This approach systematically excludes the voices of marginalized communities in the Global South. This concept raises important considerations. For instance, if open science practices and platforms are primarily designed with the Global North in mind, they may not be as effective or accessible for scholars in the Global South, including China. This could influence the attitudes and behaviors of Chinese communication scholars towards data sharing. Our research could therefore contribute to a more nuanced understanding of how open science can be more inclusive and equitable, ensuring that it serves the diverse needs of the global research community.

2.2. Data Sharing in the Global South

The Global South also plays a significant role in the OCS movement and faces unique challenges. In fact, studies have shown that OCS practices and scientific production in the Global South, such as Malaysia (Zhang et al., 2022), Thailand (Cheah et al., 2015), and Latin America (de Oliveira et al., 2021), are highly nuanced. As noted by Cheah et al. (2015), the Global South often exhibits little or no capacity for data sharing because data management is an expensive business, and skills in data collection, data validation, standardization of variables, and tabulation are rare. In addition, data analysis software/tools such as Statistical Package for Social Sciences (SPSS), Stata, and NVivo are costly, and only a few academic groups can afford them. In such cases, the difficulties of data sharing and researchers' willingness to engage in such practices might be different from the situation in Western developed countries. Therefore, investigating communication scholars' data-sharing intentions is crucial in the context of the Global South, and such an investigation could provide us with a holistic understanding of open science.

Notably, China is among the countries in which communication research is becoming more international due to the rapid development of 5G technology and the swift advancement of social media, which provides new avenues and subjects for research, thus contributing to the field's evolution and international relevance. Some Chinese scholars are actively involved in studying and advocating open science (Xu & Zhang, 2020, 2022). Research has shown that OCS in China faces challenges with regard to weak awareness of open sharing, the fuzzy boundaries of data sharing, the lack of norms of data management, and insufficient incentives (Xu & Zhang, 2020). Chinese scholars, specifically those who are interested in qualitative research, are especially concerned with difficulties pertaining to data verification, replication, and reuse in regard to data sharing (Xu & Zhang, 2022). Considering the large number of articles published by Chinese scholars and the feasibility of sampling, we developed a questionnaire to investigate the factors that impact Chinese communication scholars' data-sharing intentions.

3. Theoretical Framework

3.1. The Theory of Planned Behavior

The TPB is a model used to understand and predict human behaviors across various disciplines (Ajzen, 1991). This model was proposed to improve the theory of reasoned action (Fishbein & Ajzen, 1977) by adding "perceived behavior control." The theoretical structure of the TPB is shown in Figure 1. The TPB suggests that behavioral intentions affect actual behavior and that behavioral intentions are affected by attitudes, subjective norms, and perceived behavior control.

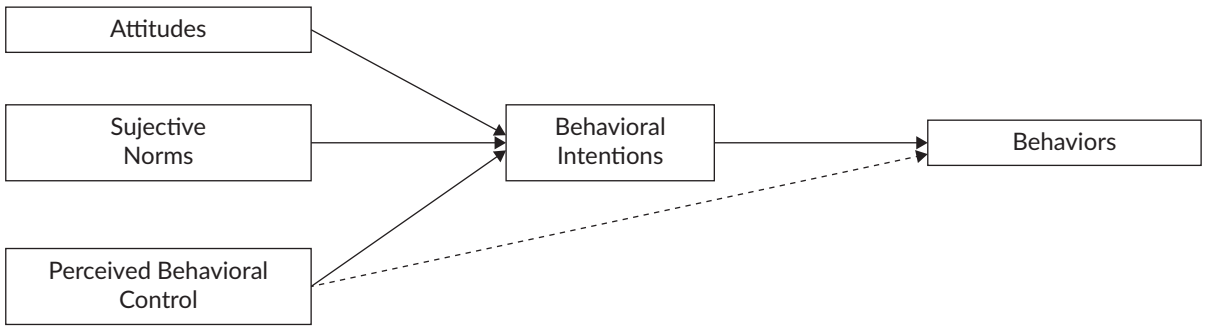

Figure 1. Conceptual model of the TPB.

The TPB has many mature applications in the social sciences, in which context researchers have explored the predictability of behaviors by explaining the formation of behaviors and the meaning of relevant influencing factors. Many studies have verified the ability of the TPB to predict behavioral intentions (Knowles et al., 2012; Kovac et al., 2016; St Quinton, 2022). Notably, the TPB is a suitable model for data policy research. For example, Sommestad et al. (2015) used the TPB to explain policy compliance intentions in information security and showed that the TPB made good predictions. Akdeniz et al. (2023) used the TPB and identified critical factors influencing intentions to share social media data, including past experiences with data sharing, attitudes towards sharing, perceived norms, and perceived behavioral control. As mentioned, Harper and Kim (2018) also employed the TPB to investigate the factors affecting American psychologists' intentions to adopt an open data badge. Such flexibility demonstrates the TPB's broad applicability across different domains, highlighting its value as a versatile framework for understanding the intricacies of behavioral intentions within various research settings.

Evidence for the TPB in non-Western cultures shows that it is a robust model for predicting behaviors across different cultural contexts. Studies have found that the core components of TPB—attitudes, subjective norms, and perceived behavioral control—are relevant and influential in shaping intentions and behaviors in various non-Western settings (Alzubaidi et al., 2021; White Baker et al., 2007). For example, Liu et al. (2019) examine factors influencing scientific data retrieval behaviors within the framework of TPB, offering insights into how attitudes, subjective norms, and perceived behavioral control can shape information-seeking actions in a Chinese academic setting. The findings underscore the theory's relevance and adaptability in understanding and predicting behavior in diverse cultural contexts. However, the specific impact and interplay of these components can differ based on cultural norms, values, and societal structures, necessitating adaptations or expansions of the model to fully capture the nuances of behavior in these contexts.

The TPB's adaptability in predicting behaviors across varied cultural contexts paves the way for an in-depth exploration of how disciplinary distinctions influence adherence to open science practices. This research delves deeper into the differences between psychology and communication studies in terms of their engagement with open science practices. Our study focuses on Chinese communication scholars, while Harper and Kim's (2018) targeted psychologists. Psychology, traditionally, has a strong emphasis on empirical, often quantitative, research that generates large datasets, making data sharing a significant aspect of open science. In contrast, communication studies encompass a broader range of methodologies, including qualitative and theoretical research, where data sharing might not always be straightforward or applicable. The discipline's diverse nature means that open science practices may vary widely, from sharing data sets to open peer review and publishing processes. The reliance on OCS in communication studies is thus a more multifaceted and nuanced issue, reflecting the diversity of its research methods and outputs.

3.2. Research Model and Hypotheses

Grounded in the theoretical framework of the TPB, we focused our study on replicating the research design of Harper and Kim (2018) and evaluating their hypotheses in the context of Chinese communication scholars. As shown in Figure 2, we incorporated attitudinal (the perceived benefit, risk, and effort of data sharing), normative (subjective norms for data sharing and data sharing pressure from journals), and resource factors (facilitating conditions for data sharing) into the model. Attitude can be defined as the overall evaluation or appraisal by communication scholars of the practice of data sharing, shaped by their personal beliefs regarding its potential benefits, consequences, and the associated effort. Subjective norms are defined in terms of communication scholars' perceptions of how others (research institutes, sponsors, etc.) view data sharing. When there is widespread support and encouragement among various groups for scholars to perform or consider a particular behavior, it increases the likelihood that these scholars will intend to adopt that behavior. In addition to individuals, academic journals also generate data-sharing pressure, since these journals usually ask authors about their data availability when submitting their manuscripts. Finally, perceived behavioral control (i.e., facilitating conditions) refers to communication scholars' perceptions of favorable resources for data sharing. Based on Harper and Kim (2018), we propose the following eight hypotheses:

H1: Perceived benefit positively affects communication scholars' attitudes toward data sharing.

H2: Perceived risk negatively affects communication scholars' attitudes toward data sharing.

H3: Perceived effort negatively affects communication scholars' attitudes toward data sharing.

H4: Perceived effort negatively affects communication scholars' behavioral intentions with regard to data sharing.

H5: Communication scholars' attitudes toward data sharing positively affect their behavioral intentions with regard to data sharing.

H6: Subjective norms of data sharing positively affect communication scholars' behavioral intentions with regard to data sharing.

H7: Data sharing pressure from journals positively affects communication scholars' behavioral intentions with regard to data sharing.

H8: Facilitating conditions for data sharing positively affect communication scholars' behavioral intentions with regard to data sharing.

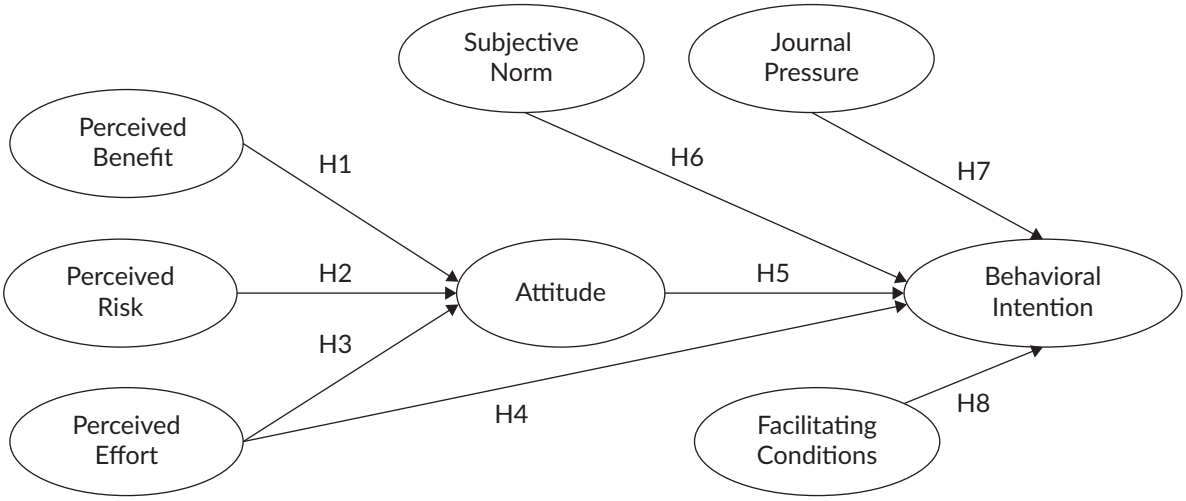

Figure 2. Research model and hypothesis. Source: Adapted from Harper and Kim (2018).

4. Research Method

4.1. Population and Sampling

This study's focus is primarily on scholars in the field of communication within China. We utilized the Chinese Association for History of Journalism and Communication (CAHJC) member list for its sampling frame. The CAHJC was established in 1989 and is the largest communication association in China. Instead of using email lists for communication, the CAHJC employs official WeChat groups and shares its latest events on that platform. We used a combination of purposive and snowball sampling methods, and we initially circulated our survey in the CAHJC's WeChat groups ($N = 871$). To reach as many Chinese communication scholars as possible, we then encouraged our participants to distribute the questionnaire to former or present coworkers or PhD students. Two screening questions were included to ensure that the respondents were conducting research in the field of communication and either had PhD degrees or were PhD candidates. Participants in the study were informed that their involvement was entirely optional, and they were assured that all responses would be treated with confidentiality and anonymity.

4.2. Measurement of Constructs

In this study, we used 28 survey items to assess eight research constructs. The majority of these measurement items were adapted from Harper and Kim (2018) and modified to suit communication scholars' data-sharing adoption contexts. Most items measuring communication scholars' diverse perceptions of data sharing were scored on a 7-point Likert scale ranging from *strongly disagree* to *strongly agree*. Multiple items were employed to gauge each research construct in this study. Detailed information on the measurement

items utilized for these constructs is provided in Table 1. The final survey commenced with the presentation of a consent form, followed by a comprehensive introduction to the concept of open science. This introduction included a detailed definition of open science and a discussion of its historical evolution and developmental context.

4.3. Data Collection Procedure and Results

The survey was distributed through WJX, an online survey management software from China (https://www.wjx.cn), on May 17, 2022, and remained available for four weeks. We collected a total of 412 valid responses from our survey participants. The average time (in seconds) to complete the survey was $M = 918.8$ ($SD = 736$), with a median of 918. As suggested by Bowman et al. (2021), we excluded responses that were either quicker than the bottom 5% (≤ 361.3, $n = 20$) or longer than the top 90% (≥ 1431, $n = 41$). This filtering resulted in the removal of 61 responses, leading to a final sample size of 351. The adjusted average completion time for the survey was 757 seconds ($SD = 248.7$). The median response time recorded was 729 seconds, with the shortest frequent response time being 363 seconds.

4.4. Demographics of the Respondents

The demographic information of the survey respondents encompassed gender, academic title, and research areas of the communication scholars, categorized according to the ICA classification system. Among the 351 participants, 192 (54.7%) identified as "women." Ninety-two respondents (26.2%) reported having "senior faculty" status, followed by 113 (36.2%) who reported having "mid-career faculty" status, 52 (14.8%) who reported having "junior faculty" status, and 94 (25.4%) who reported being "postdoctoral or doctoral students." In terms of research interest, all 25 ICA divisions and interest groups were represented, with each individual being categorized into only the one division that was closest to their research interest. At least 10% of the sample represented two subfields: popular media and culture ($n = 66$, 18.8%) and communication and technology ($n = 36$, 10.3%). Other significant categories included journalism studies ($n = 33$, 9.4%), political communication ($n = 31$, 8.8%), and mass communication ($n = 28$, 8%). This distribution closely mirrors that found by surveys of ICA members (Bowman et al., 2021).

5. Data Analysis and Results

5.1. Reliability and Validity Test

After the questionnaire was developed, a pilot test was conducted with the participation of 30 communication scholars. These scholars could complete the questionnaire within an average time of about 10 minutes. Feedback from the respondents indicated that the instructions and question wording were clear and effectively measured the intended constructs, affirming both the face and content validity of the questionnaire (Churchill, 1979). As depicted in Table 1, the average scores for the eight variables examined in the study varied from 3.314 to 5.256, with standard deviations ranging between 1.116 and 1.405. The Cronbach's α of all variables was above .7, thus suggesting good reliability.

The study utilized IBM SPSS AMOS 24 for confirmatory factor analysis to validate the measurement model's reliability and validity. This process included checking each variable's skewness and kurtosis to confirm data

Table 1. Measurement items and descriptive analysis for research constructs.

Construct	Items	*M*	*SD*	*α*
Perceived Benefit (Harper & Kim, 2018)	I can earn academic credit, such as more citations, by sharing data.	5.256	1.116	.894
	Data sharing would enhance my academic recognition.			
	Data sharing would improve my status in the research community.			
Perceived Risk (Harper & Kim, 2018)	There is a high probability of losing publication opportunities if I share data.	5.094	1.302	.835
	Data sharing may cause my research ideas to be stolen by other researchers.			
	My shared data may be misused or misinterpreted by other researchers.			
Perceived Effort (Harper & Kim, 2018)	Sharing data involves too much time for me (e.g., to organize/annotate).	4.747	1.272	.834
	I need to make a significant effort to share data.			
	I would find data sharing difficult to do.			
Attitudes (Harper & Kim, 2018; Venkatesh et al., 2003)	Data sharing is a good idea.	5.031	1.262	.945
	Data sharing is valuable for replication.			
	Data sharing is valuable for reproducibility.			
	Data sharing makes research more transparent.			
Subjective Norms (Harper & Kim, 2018; Taylor & Todd, 1995)	My university or research institutes think that I should share data.	3.314	1.405	.952
	My editors or reviewers think that I should share data.			
	The institutions funding my research think that I should share data.			
	My sponsors/consultants/bosses think that I should share data.			
Journal Pressure (Bowman et al., 2021; Harper & Kim, 2018)	The open science journal expects me to share the datasets for my manuscripts.	4.992	1.284	.903
	The open science journal expects me to share the research materials for my manuscripts.			
	The open science journal expects me to apply for open science badges when submitting manuscripts.			
Facilitating Conditions (Harper & Kim, 2018; Venkatesh et al., 2003)	I have the resources necessary to share data.	3.390	1.374	.942
	I have the knowledge necessary to share data.			
	I have the data repositories necessary to share data.			
	A specific person (or group) is available for assistance with data-sharing difficulties.			
Behavioral Intentions (Harper & Kim, 2018; Venkatesh et al., 2003)	I intend to share data in my future research.	4.613	1.296	.960
	I predict that I will share data in my future research.			
	I plan to share data in my future research.			
	I will likely share my research data in the future.			

normality. The analysis involved calculating composite reliabilities (CRs) and average variance extracted (AVE) values through standardized loadings, which were instrumental in evaluating the constructs' convergent and discriminant validity. Results presented in Table 2, including CRs and AVE values, indicated a high level of reliability (CRs between .902 and .971) and sufficient convergent validity (AVE values over .5). The study also examined discriminant validity by comparing the AVE values' square roots with the correlations among different constructs. The square roots of the AVE values exceeded all the interconstruct correlations, supporting discriminant validity (Fornell & Larcker, 1981). All bivariate correlations between these variables were below the critical .75 threshold, which implied that they did not strongly overlap. Meanwhile, the collinearity diagnostics indicated that the Variance Inflation Factors for all variables were below the threshold of 5, suggesting no concerns about multicollinearity in the model.

Table 2. CRs, AVE values, and correlations among the constructs.

Variable	CRs	AVE	1	2	3	4	5	6	7	8
Perceived Benefit	.936	.829	.910							
Perceived Risk	.936	.829	–.035	.910						
Perceived Effort	.902	.755	–.106*	.568**	.869					
Attitudes	.961	.859	.475**	–.210**	–.291**	.927				
Subjective Norms	.966	.876	.200**	.047	–.073	.342**	.936			
Journal Pressure	.940	.839	.339**	–.143**	–.190**	.634**	.240**	.916		
Facilitating Conditions	.959	.853	.095	–.066	–.210**	.253**	.486**	.165**	.924	
Behavioral Intentions	.971	.893	.408**	–.154**	-.265**	.687**	.418**	.546**	.331**	.945

Notes: * $p < .05$, ** $p < .01$; The square roots of the AVE values are diagonal.

5.2. Structural Model and Hypothesis Testing

Structural equation modeling with SPSS AMOS 24 was employed to test the hypothesis. This involved defining each latent variable by its observed variables and using six indices to evaluate the model's fit to the data. The criteria for these indices included χ^2/df below 3, Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and Incremental Fit Index (IFI) above .90, and RMSEA and SRMR below .08 (Hair et al., 2019). The analysis yielded fit indices of $\chi^2/df = 2.17$ ($\chi^2 = 1552.63$; $df = 737$; $p < .001$), TLI = .93, CFI = .93, IFI = .94, RMSEA = .06, and SRMR = .06, indicating a good fit of the structural model to the data. Figure 3 shows the results of the structural model. All paths were significant with the exception of the path between perceived effort and behavioral intentions toward data sharing ($\beta = –.022$). Therefore, H4 was rejected, while the other hypotheses were supported.

6. Findings and Discussion

6.1. Attitudinal Factors

Similar to Harper and Kim (2018), we found that perceived benefit ($\beta = .544$, $p < .001$) and perceived risk ($\beta = –.116$, $p < .05$) had significant positive and negative impacts on communication scholars' attitudes toward

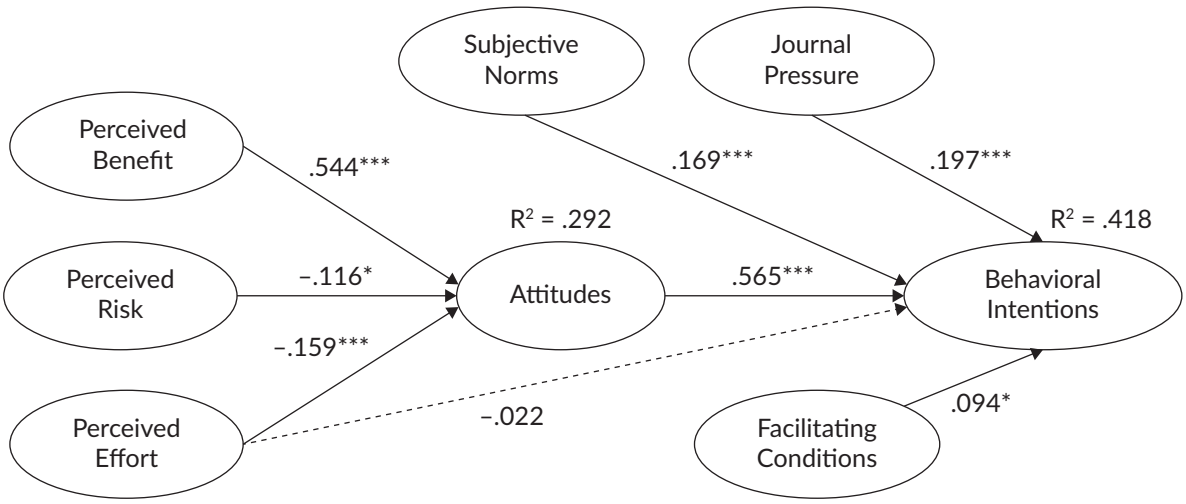

Figure 3. The results regarding the standardized path coefficients for the hypothesized paths. Notes: * $p < .05$, *** $p < .001$; the significant paths are shown as solid lines, while the nonsignificant paths are shown as dotted lines.

data sharing, respectively, and that perceived benefit was the strongest predictor among all attitudinal factors. Our study is also in line with Akdeniz et al. (2023), who found that while altruistic motives like contributing to open science and transparency were pivotal in sharing decisions, practical concerns like legal restrictions and ethical considerations also played a significant role. Such insights are crucial in understanding the dynamics of data-sharing behavior. To promote more positive attitudes toward data sharing, the perceived benefits ought to be emphasized, and the perceived risks and costs ought to be reduced. As data sharing is seldom acknowledged in key academic progression documents like tenure and promotion guidelines (Pontika et al., 2022), we urge policymakers and research organizations to reevaluate and enhance the incentives for data publication in the field of communication research. It is essential to align the benefits of data sharing, such as increased citations, academic recognition, and enhanced research opportunities, with institutional recognition and career advancement metrics (Bock et al., 2005). Second, perceived risk can be reduced by reframing data sharing as an opportunity to build on the extant literature and to promote potential future collaboration; furthermore, by increasing support at the level of academia, communication scholars can be provided with an institutional safety net for data sharing.

However, our study found different results regarding perceived effort. Harper and Kim (2018) claimed that perceived effort negatively affected individuals' behavioral intentions but did not influence their attitudes. In contrast, our results suggested that perceived effort negatively affected individuals' attitudes ($\beta = -.159$, $p < .001$) but had no significant impact on behavioral intentions ($\beta = -.022$, $p > .05$). This discrepancy is probably due to the nuanced differences between the different measurements used. The subject studied by Harper and Kim (2018) was open data badge adoption, which differs from our emphasis on data sharing, as data sharing requires more multilayered efforts than open data badge adoption, which refers simply to the initial attempt made by the journal to promote open data sharing. Therefore, in this particular situation, the impact of perceived effort may not be sufficiently strong to change individuals' behavioral intentions. Given the negative impact of perceived effort on attitudes, we urge research institutions to provide data management protocols or easy-to-follow recommendations. Consequently, sharing data can become less burdensome for individuals, thus helping improve their attitudes toward data sharing.

6.2. Normative and Resource Factors

In line with Harper and Kim (2018), our study indicated that communication scholars were positively influenced by the norms of data sharing ($\beta = .169$, $p < .001$), indicating that the more researchers believe that their research institutes, funding institutions, or sponsors expect them to share data, the more likely they are to perform that behavior. To foster a positive environment for data sharing, research institutions and academic bodies, like ICA, could actively promote data-sharing norms. This could involve distributing educational content about various data-sharing methods and incorporating data-sharing practices as a criterion in the distribution of research funding and resources.

Additionally, both data sharing pressure from journals ($\beta = .197$, $p > .001$) and facilitating conditions ($\beta = .094$, $p < .05$) positively impacted communication scholars' intention toward data sharing, a finding which differs from those reported by Harper and Kim (2018), who found no significant impacts on pressure from open science journals or the availability of data repositories. According to Harper and Kim (2018), the nonsignificant effects might be due to a lack of data repositories, and the pressure for researchers to share data with a publication generated by journals was optional rather than mandatory in 2018. However, at the time we conducted our research, journal requirements for data sharing had become more pervasive (e.g., through standardized preregistration, open-sharing protocols, and open data statements), and free data repositories such as Open Science Framework had become more easily accessible. Hence, the impacts of these two factors have become more significant. Given the significant impacts of journal pressure and open data badge adoption, we suggest that more measures should be taken to facilitate data sharing, including knowledge sharing on the part of experts and data scientists, lectures and skill training in data sharing, and customized librarian services and consultations. These efforts can encourage communication scholars with the necessary expertise and resources to engage in data sharing and enable them to receive help in a timely manner when they encounter difficulties.

7. Implications

7.1. Theoretical Implications

Our study focuses on examining the factors that influence data sharing within a new context, aiming to evaluate the actual impact and magnitude of variables based on the TPB. Our study aims to establish the generalizability of the predicted effects in a non-Western context. The innovative double nature of the study lies in the fact that it is both (a) an empirical investigation of the factors that impact communication scholars' data-sharing intentions and are relevant for replication work and (b) a replication study in its own right that aims to obtain insights from other disciplines in the social sciences (namely, in this case, psychology).

First, this research offers valuable insights into the cognitive decision-making process of communication scholars regarding open science. While it's notable for pioneering this area of inquiry, the significance of the study extends beyond its novelty. It delves into the intricate interplay of attitudinal, normative, and resource factors shaping scholars' intentions to share data. Drawing from Harper and Kim (2018), we construct a comprehensive theoretical framework, enhancing our understanding of data-sharing behaviors within the open science context. This approach not only fills a gap in existing literature but also enriches our

comprehension of scholarly behaviors in relation to the evolving open science movement, underscoring its broader implications and relevance in the field of communication.

Second, our focus on the non-Western context underlines the diversity and complexity in data-sharing practices globally. By exploring the factors influencing data sharing among Chinese communication scholars, this research not only provides empirical evidence from a non-Western perspective but also highlights the need for a more inclusive understanding of open science practices. It challenges the predominantly Western-centric view of data-sharing norms and opens up avenues for further research in diverse cultural and academic settings. Our findings suggest that Chinese researchers' attitudes towards data sharing do differ from their counterparts in other regions. These differences could stem from various factors, such as cultural attitudes towards intellectual property and collaboration, the influence of local academic and research policies, infrastructural and resource availability specific to China, and differing levels of emphasis on open science practices in the academic community. Understanding these nuances is crucial for developing effective strategies to promote data sharing across diverse academic cultures.

Third, replication represents a significant theoretical contribution (DeAndrea & Holbert, 2017) and is a fundamental component of scientific inquiry. If an effect is genuinely robust and valid, any competent researcher should be able to observe it when using the same procedures with sufficient statistical power (Simons, 2014). Our study confirms the effectiveness of the TPB in this context and further indicates that the adapted structural equation model leads to better interpretations of communication scholars' data-sharing intentions. As indicated in Figure 3, approximately 29.2% of the variance in attitudes (R^2 = .292) is explained by perceived benefit, perceived risk, and perceived effort, and approximately 41.8% of the variance in behavioral intentions (R^2 = .418) can be explained by perceived effort, attitudes, subjective norms, journal pressure, and facilitating conditions, both of which are higher than the R^2 values reported by Harper and Kim (2018). These two numbers indicate a good fit of the theories with regard to analyzing communication scholars' data-sharing intentions.

7.2. Practical Implications

This research has practical implications for both the academic field of communication studies and the broader community of researchers in the social sciences. Although this study focuses on data sharing, the model it proposes can be utilized to understand broad OCS-related practices. By employing the framework presented in this study, universities and research institutes can be better prepared to motivate communication scholars to share their research data, among many other OCS behaviors. More specifically, this study has the following practical implications.

The findings of this study emphasize the fact that attitudinal factors, including perceived benefits, risks, and benefits, have the most significant impacts on the willingness of communication scholars to share their research data. By encouraging open dialog within the research community to address concerns with perceived risks and enhance the perceived benefits of data sharing, communication scholars can collaboratively take steps toward promoting greater transparency within their profession. Universities and research institutions should consider implementing supportive institutional incentives for OCS. However, it is important to note, as revealed in the study by Bowman et al. (2021), that there is currently no consensus among scholars on mandating the publication of research data or giving hiring and promotion preference to

applicants who share their data. Therefore, while encouraging OCS practices, institutions should also be mindful of the diverse perspectives within the academic community and aim to foster an environment that balances open science initiatives with respect for scholarly discretion and diversity of opinions. In this case, communication scholars can perceive data sharing as less intimidating and more rewarding. If the benefits of data sharing are expanded, the risks are minimized, and the costs in terms of time and effort are decreased, communication scholars can become more willing to adopt positive attitudes toward data sharing.

Furthermore, normative and resource factors, such as the subjective norms of data sharing, data sharing pressure from journals, and facilitating conditions, positively influence communication scholars' behavioral intentions. Therefore, research institutions, academic associations such as the ICA, and publishers in the field should unite to encourage and facilitate data sharing. For example, these actors can establish friendly systems or produce helpful tools to facilitate data sharing, offer OCS-related training to increase communication scholars' skills with regard to data sharing, verification, and reuse (Zhang et al., 2022), and promote a shift toward the standardization of open data practices within their research community. Such efforts can help establish a better research environment, and greater engagement in data sharing and other OCS practices can thus be promoted within the field of communication.

8. Conclusion

8.1. Limitations

While we conduct this empirical investigation with great care, it is important to acknowledge certain limitations in the methodologies we use. First, our assessment primarily focuses on scholars' behavioral intentions to share data rather than their actual behaviors. This approach is chosen because data sharing has not yet become a common practice in China, and the majority of Chinese communication scholars lack experience in this area. It is worth noting that, consistent with Harper and Kim (2018), we do not include control variables such as personal traits and social factors. However, these variables could potentially influence an individual's behavioral intentions and might introduce confounding variables into our results.

Furthermore, our sampling methodology exhibits certain imperfections. Identifying a sampling frame that fully encompasses all Chinese communication scholars is a challenging task. While the CAHJC is the largest communication association in China, it is important to recognize that not all Chinese communication scholars participate in its WeChat contact groups.

Finally, our study exhibits a relatively low response rate, receiving only 412 responses from 871 individuals surveyed, and 351 of these responses are ultimately included in our final analysis. It is possible that individuals who chose to respond to our survey may have been more inclined to support OCS than those who did not participate.

8.2. Directions for Future Research

In conclusion, while our study provides valuable insights, it is essential to interpret our results in light of these methodological limitations, and future research in this area should aim to address these challenges to obtain

a more comprehensive understanding of the subject matter. Specifically, future research can be expanded in the following ways.

First, prior studies have indicated the presence of differences in the ways in which open science-related practices are embraced and adopted among individuals from diverse demographic and professional backgrounds (Bowman et al., 2021; Markowitz et al., 2021). It is advisable for future investigations to explore the impacts of personal characteristics, such as gender, career advancement, methodological preferences, and research interests, on communication scholars' perceptions of OCS. Furthermore, factors such as age, familiarity with OCS, publication history in academic journals, and experience in sharing research data could be subject to more extensive analysis. It will also be meaningful to investigate the correlation between expressed intentions and actual involvement. Such investigations would provide a comprehensive understanding of how these characteristics influence aspects such as the individual's self-efficacy with regard to adopting OCS and readiness for change.

Second, our study's focus on Chinese communication scholars might have limited our ability to capture other significant attributes. We acknowledge that the nuances between various academic systems worldwide could significantly impact open science practices. Different countries may have unique characteristics in terms of their academic culture, policies regarding intellectual property and data ownership, levels of funding support for open science initiatives, and the maturity of digital infrastructures for managing and sharing research outputs. These differences can lead to variations in how scholars perceive, adopt, and engage with OCS practices, including but not limited to data sharing. Future research conducted on a broader scale should take into careful consideration the nuances of different academic systems and the overall progress of OCS in different countries.

Finally, our focus on data sharing may have overshadowed other crucial dimensions of open science practices. Given the theoretical underpinnings of this model, which are centered around individual beliefs, social pressures, and available resources, it is plausible to hypothesize that similar mechanisms would operate when applied to other aspects of OCS, such as preregistration, replication, and reproducibility (Bowman et al., 2021). However, while there might be similarities in how these factors impact different OCS aspects, we acknowledge that unique characteristics associated with each dimension could lead to differential effects. For example, the specific challenges related to data sharing (e.g., privacy concerns, technical difficulties) may not be directly transferable to the context of preregistration, where issues like predicting outcomes and committing to methods before data collection become central. Similarly, the incentives and barriers surrounding replication might differ from those pertaining to data sharing due to the complexities involved in reproducing entire studies versus simply making data accessible. The study by Krähmer et al. (2023) sheds light on this aspect, examining the factors that influence researchers' willingness to share their analysis code. Their findings reveal that the framing of code-sharing requests, particularly those that underscore the replication crisis, significantly impacts researchers' sharing behavior. Therefore, future research should explore the application of our model across other dimensions of OCS, which may exhibit differences from the findings we obtained regarding data-sharing intentions.

Acknowledgments

We extend our deepest thanks to Johannes Breuer and Mario Haim for their expert guidance, and to Raquel Silva for her editorial excellence.

Funding

This article is supported by Anhui Province Philosophy and Social Science Planning Project (AHSKQ2020D53).

Conflict of Interests

The author declares no conflict of interests.

Data Availability

Our data are available at https://osf.io/g8qwd, reference number https://doi.org/10.17605/OSF.IO/G8QWD.

References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior Human Decision Processes*, *50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Akdeniz, E., Borschewski, K. E., Breuer, J., & Voronin, Y. (2023). Sharing social media data: The role of past experiences, attitudes, norms, and perceived behavioral control. *Frontiers in Big Data*, *5*, Article 971974. https://doi.org/10.3389/fdata.2022.971974

Alzubaidi, H., Slade, E. L., & Dwivedi, Y. K. (2021). Examining antecedents of consumers' pro-environmental behaviours: TPB extended with materialism and innovativeness. *Journal of Business Research*, *122*, 685–699. https://doi.org/10.1016/j.jbusres.2020.01.017

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

Bock, G.-W., Zmud, R. W., Kim, Y.-G., & Lee, J.-N. (2005). Behavioral intention formation in knowledge sharing: Examining the roles of extrinsic motivators, social-psychological forces, and organizational climate. *MIS Quarterly*, *29*(1), 87–111. https://doi.org/10.2307/25148669

Bowman, N., Rinke, E. M., Lee, E.-J., Nabi, R., & de Vreese, C. (2021). How communication scholars see open scholarship: A survey of international communication association scholars. *Annals of the International Communication Association*, *46*(3), 205–230. https://doi.org/10.1080/23808985.2022.2108880

Cheah, P. Y., Tangseefa, D., Somsaman, A., Chunsuttiwat, T., Nosten, F., Day, N. P., Bull, S., & Parker, M. (2015). Perceived benefits, harms, and views about how to share data responsibly: A qualitative study of experiences with and attitudes toward data sharing among research staff and community representatives in Thailand. *Journal of Empirical Research on Human Research Ethics*, *10*(3), 278–289. https://doi.org/10.1177/1556264615592388

Churchill, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research*, *16*(1), 12–27. https://doi.org/10.1177/002224377901600110

DeAndrea, D. C., & Holbert, R. L. (2017). Increasing clarity where it is needed most: Articulating and evaluating theoretical contributions. *Annals of the International Communication Association*, *41*(2), 168–180. https://doi.org/10.1080/23808985.2

de Oliveira, T. M., Marques, F. P. J., Veloso Leão, A., de Albuquerque, A., Prado, J. L. A., Grohmann, R., Clinio, A., Cogo, D., & Guazina, L. S. (2021). Towards an inclusive agenda of open science for communication research: A Latin American approach. *Journal of Communication*, *71*(5), 785–802. https://doi.org/10.1093/joc/jqab025

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

Dutta, M., Ramasubramanian, S., Barrett, M., Elers, C., Sarwatay, D., Raghunath, P., Kaur, S., Dutta, D., Jayan, P., Rahman, M., Tallam, E., Roy, S., Falnikar, A., Johnson, G. M., Mandal, I., Dutta, U., Basnyat, I., Soriano, C., Pavarala, V., . . . Zapata, D. (2021). Decolonizing open science: Southern interventions. *Journal of Communication*, *71*(5), 803–826. https://doi.org/10.1093/joc/jqab027

Fishbein, M., & Ajzen, I. (1977). *Belief, attitude, intention, and behavior: An introduction to theory and research*. Addison-Wesley.

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Grand, A., Wilkinson, C., Bultitude, K., & Winfield, A. F. T. (2012). Open science: A new "trust technology"? *Science Communication*, *34*(5), 679–689. https://doi.org/10.1177/1075547012443021

Haim, M., & Puschmann, C. (2023). Opening up data, tools, and practices: Collaborating with the future. *Digital Journalism*, *11*(2), 247–254. https://doi.org/10.1080/21670811.2023.2174894

Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. (2019). *Multivariate data analysis* (8th ed.). Pearson Education.

Harper, L. M., & Kim, Y. (2018). Attitudinal, normative, and resource factors affecting psychologists' intentions to adopt an open data badge: An empirical analysis. *International Journal of Information Management*, *41*(8), 23–32. https://doi.org/10.1016/j.ijinfomgt.2018.03.001

Knowles, S. R., Hyde, M. K., & White, K. M. (2012). Predictors of young people's charitable intentions to donate money: An extended theory of planned behavior perspective. *Journal of Applied Social Psychology*, *42*(9), 2096–2110. https://doi.org/10.1111/j.1559-1816.2012.00932.x

Kovac, V. B., Cameron, D. L., & Hoigaard, R. (2016). The extended theory of planned behaviour and college grades: The role of cognition and past behaviour in the prediction of students' academic intentions and achievements. *Educational Psychology*, *36*(4), 792–811. https://doi.org/10.1080/01443410.2014.923557

Krähmer, D., Schächtele, L., & Schneck, A. (2023). Care to share? Experimental evidence on code sharing behavior in the social sciences. *PLoS ONE*, *18*(8), Article e0289380. https://doi.org/10.1371/journal.pone.0289380

Liu, J., Wang, J., Zhou, G., Zhang, G., Cui, Y., & Liu, J. (2019). User's scientific data retrieval behavior study based on the model of TPB. *Proceedings of the 3rd International Conference on Computer Science and Application Engineering*, *2019*, Article 71. https://doi.org/10.1145/3331453.3360951

Lewis, N. A. (2020). Open communication science: A primer on why and some recommendations for how. *Communication Methods and Measures*, *14*(2), 71–82. https://doi.org/10.1080/19312458.2019.1685660

Markowitz, D. M., Song, H., & Taylor, S. H. (2021). Tracing the adoption and effects of open science in communication research. *Journal of Communication*, *71*(5), 739–763. https://doi.org/10.1093/joc/jqab030

Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable research practices in experimental communication research: A systematic analysis from 1980 to 2013. *Communication Methods and Measures*, *9*(4), 193–207. https://doi.org/10.1080/19312458.2015.1096334

Pontika, N., Klebel, T., Correia, A., Metzler, H., Knoth, P., & Ross-Hellauer, T. (2022). Indicators of research quality, quantity, openness, and responsibility in institutional review, promotion, and tenure policies across seven countries. *Quantitative Science Studies*, *3*(4), 888–911. https://doi.org/10.1162/qss_a_00224

Shaw, A., Scharkow, M., & Wang, Z. J. (2021). Opening a conversation on open communication research. *Journal of Communication*, *71*(5), 677–685. https://doi.org/10.1093/joc/jqab033

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/174569161351475

Sommestad, T., Karlzén, H., & Hallberg, J. (2015). The sufficiency of the theory of planned behavior for explaining information security policy compliance. *Information Computer Security*, *23*(2), 200–217. https://doi.org/10.1108/ICS-04-2014-0025

St Quinton, T. (2022). Student participation in gambling: The role of social cognition, past behaviour, and habit. *Psychology Health & Medicine*, *27*(8), 1774–1781. https://doi.org/10.1080/13548506.2021.1944657

Taylor, S., & Todd, P. A. (1995). Understanding information technology usage: A test of competing models. *Information Systems Research*, *6*(2), 144–176. https://doi.org/10.1287/isre.6.2.144

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., & Dorsett, K. (2015). Changes in data sharing and data reuse practices and perceptions among scientists worldwide. *PLoS ONE*, *10*(8), Article e0134826. https://doi.org/10.1371/journal.pone.0134826

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, *27*(3), 425–478. https://doi.org/10.2307/30036540

Vines, T. H., Albert, A. Y. K., Andrew, R. L., Débarre, F., Bock, D. G., Franklin, M. T., Gilbert, K. J., Moore, J.-S., Renaut, S., & Rennison, D. J. (2014). The availability of research data declines rapidly with article age. *Current Biology*, *24*(1), 94–97. https://doi.org/10.1016/j.cub.2013.11.014

White Baker, E., Al-Gahtani, S. S., & Hubona, G. S. (2007). The effects of gender and age on new technology implementation in a developing country: Testing the theory of planned behavior (TPB). *Information Technology & People*, *20*(4), 352–375. https://doi.org/10.1108/09593840710839798

Xu, J., & Zhang, R. (2020). Communication stepping towards open science: Opportunities, challenges, and future. *Editorial Friend*, *41*(12), 76–84.

Xu, J., & Zhang, R. (2022). "The future has come," open science and qualitative research: In-depth interviews with 30 Chinese communication scholars. *Journal of Communication University of China*, *44*(4), 11–18.

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

Zhang, R., Gong, J., Hou, W., Firdaus, A., & Xu, J. (2022). Is open communication scholarship a promise or peril? Preliminary interviews with qualitative communication scholars. *International Journal of Communication*, *16*, 5318–5337. https://ijoc.org/index.php/ijoc/article/view/18304

About the Authors

Jinghong Xu is a professor at the School of Journalism and Communication at Beijing Normal University. He is also a researcher at the Key Laboratory of Big Data Analysis and Application in Publishing Industry of the National Press and Publication Administration. His research interests include new media and internet governance, health communication, intercultural communication, film and television studies, and game research.

Rukun Zhang is a doctoral candidate at the School of Journalism and Communication at Beijing Normal University. Her research interests encompass new media and technology, science communication, and media effects studies.

Media and Communication
2024 • Volume 12 • Commentary 7852
https://doi.org/10.17645/mac.7852

COMMENTARY

Open Access Journal

Remembering Reasons for Reform: A More Replicable and Reproducible Communication Literature Without the Rancor

James D. Ivory

Department of English, Virginia Tech, USA

Correspondence: James D. Ivory (jivory@vt.edu)

Submitted: 16 November 2023 **Accepted:** 6 February 2024 **Published:** 12 March 2024

Issue: This commentary is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Increasing awareness of the "replication crisis" has prompted discussion about replicability and reproducibility in social and behavioral science research, including in communication. As with other fields, communication has seen discussion about concerns with the interpretation of existing research. One response has been the piecemeal adoption of "open science" practices in communication to reduce selectivity in analysis, reporting, and publication of research. Calls for further adoption of such practices have, in turn, been met with criticisms and concerns about the negative consequences of their adoption. Amidst disparate perspectives regarding solutions to replicability and reproducibility issues in communication science, difficulties building consensus and caution about negative outcomes are understandable, but they also present the risk of a status quo bias that could stall the improvement of the replicability and reproducibility of communication research. The urgency of the replication crisis for communication and the cost of inaction are presented here along three exemplifying dimensions perhaps of particular importance in communication research: (a) responsibility to the public, (b) stewardship of resources, and (c) membership in a community of scholars. While debate over solutions will continue, we would do well to keep in mind that problems with replicability and reproducibility in communication research are indeed a crisis needing immediate attention.

Keywords

communication; open science; replicability; reproducibility; science reform

1. Introduction

This commentary discusses the relevance of the "replication crisis" to the field of communication and the urgent need to address issues with replicability and reproducibility in communication scholarship. The following brief sections overview scholarship and discussion about concerns with replicability and reproducibility across fields and in communication specifically, emphasize a few examples of implications for which the stakes are particularly high for the field of communication, and call for cooperative and collaborative conversation in the field to ensure a valid and credible research record in communication scholarship. While there are unresolved questions about how to confront these issues in communication, this commentary emphasizes that they must indeed be confronted.

2. The Replication Crisis in Communication and Debates About the Implementation of Open Science Practices

Replicability and reproducibility have both been described as cornerstones of science (Moonesinghe et al., 2007; Simons, 2014). As typically used in social and behavioral science, the concepts are distinct, but both closely related to the validity of findings. Reproducibility concerns the extent to which repeating the same analyses with existing study data produces the same result originally found and reported in a study, i.e., whether the original analyses' findings can be reproduced with the data. Replicability deals with whether conducting a study again will generate similar findings to the original study, i.e., whether the original study's findings are generalizable enough that they can be observed again in other research. (Nosek et al., 2022). A third relevant concept, robustness, refers to whether analyzing a study's data with minor variations in analyses will produce consistent results, i.e., whether the study's general findings can be observed consistently regardless of analysis strategy or whether a study's results are delicate enough to only be found with analyses employing a very particular combination of variables and cases. If a study has been conducted and reported correctly and the data curated accurately, then, its findings should in principle always be reproducible; replicability of a research finding, meanwhile, may vary from the finding being broadly observed across a vast range of settings to a finding that can only be replicated in a specific context to a finding that does not appear to be replicable even in a close attempt at a "direct replication" of a study's design.

It has been no exercise in hyperbole, then, for many in the field to use "replication crisis" to describe the burgeoning awareness over the past decade or two that vast swaths of influential empirical findings in a range of scientific fields have proven difficult to replicate and reproduce. Meta-scientific research in psychology, for example, suggests that reproducibility rates are some distance below the theoretically expected 100% (Artner et al., 2021) and that replicability rates are even lower (Open Science Collaboration, 2015). Communication research, particularly in the quantitative communication science tradition, has been among the disciplines touched by the replication crisis, and in ways that extend far beyond our collective confidence in study data or the conceptual models they support. The replication crisis has not only been a crisis of confidence in our data; in many ways, it has fomented a crisis of confidence in our community of researchers.

The most prominent studies that served as harbingers of the replication crisis (e.g., Ioannidis, 2005; Klein et al., 2014; Open Science Collaboration, 2015) were focused on the shortcomings of study designs and resultant

findings vis-à-vis replicability and reproducibility. Inevitably, though, discussions of biases in designs, analyses, and processes (e.g., Ioannidis, 2005) have led to necessary, but unpleasant conversations about human biases, motivated reasoning (Simonsohn et al., 2015), and flawed behaviors by researchers behind a flawed research record. Here, conversations about replicability and reproducibility issues may understandably put researchers on the defensive as terms like "*p*-hacking" and "questionable research practices" (John et al., 2012) enter the parlance as grim reminders that a primary culprit for the replication crisis is at best selective flexibility in analysis, reporting, and publication of data and, at worst, deliberate misrepresentation of findings (Fanelli, 2009; Simmons et al., 2011). The same is true in communication research; some trends in findings, such as distribution of significant test results, seem unlikely to have occurred without biased action by at least some of the researchers producing them (Vermeulen et al., 2015). Discussions about replicability and reproducibility can therefore feel accusatory; even if no one study or author is under scrutiny, the implication is that someone has a thumb on the proverbial scale.

This tension is exacerbated by disagreements over whether and what reform is needed in scientific practice to ensure the integrity of the replicability and reproducibility "cornerstones." A range of "open science" practices have been mooted to address this concern, from tools used to identify potential concerns with existing bodies of literature to procedures intended to increase the transparency of research practices at various points in the research and publication process. Some such practices have been adopted in communication, albeit in a somewhat piecemeal fashion across researchers and publication venues, and calls for more adoption of such practices in communication research have been made (e.g., Dienlin et al., 2021). Concerns about such reform efforts have been raised, though, including claims that enthusiasm around open science practices may be exclusionary and divisive, as well as a threat to the privacy of research participants (Fox et al., 2021). Further, the implementation of popular open science practices without adequate supervision and structure may enable "openwashing" behaviors that present the appearance of more replicable and reproducible research without actual reform (Markowitz et al., 2021).

3. A Common Crisis: The Urgent Human Cost of Replicability and Reproducibility Issues

As the debate about open science practices indicates, it will be difficult to reach a consensus on ideal solutions, especially considering that some practices introduce different concerns in certain specific communication research domains (e.g., health communication research involving disclosure of health information). Focusing solely on disagreements about the implementation of science reform practices in these conversations, though, may be conducive to status quo bias (Samuelson & Zeckhauser, 1988). This is a real danger, as the field's history of largely ignoring calls for a more replicable body of research (e.g., Kelly et al., 1979) should tell us that a change in process is sorely needed. Thus, the debate about how to solve replicability and reproducibility issues in communication research must remain mindful not only of the pros and cons of solutions, but of the urgent and very real cost of the issues for which solutions are needed. Here are three examples of areas perhaps particularly germane to communication as reminders that replicability and reproducibility issues in the field are both urgent and destructive.

3.1. Responsibility to the Public

Much communication research is of particular interest to the public. Audiences look to communication researchers for answers about such topics as what happens to their children when they play video games

and use social media, methods to resolve conflict with romantic partners and family members, the prevalence and effectiveness of political campaign tactics, how the news media portray us and our world's issues, and the promise of communication technologies for health and education. These findings reach the public, perhaps disproportionately compared to sometimes-arcane academic research in general, via press releases, news coverage, and popular media interviews. This public communication of research findings lacks sufficient detail for audiences to evaluate the details of the research, so trust is placed in us to ensure its validity. The public is listening to us. Thus, it is urgent that we ensure we are communicating the most accurate body of research we can.

3.2. Stewardship of Resources

As demand in some academic fields plateaus and contracts, communication remains a robust academic discipline even in the increasingly uncertain economies of higher education. Graduate and undergraduate enrollment trends in communication remain relatively verdant, at least in the United States, where the number of bachelor's degrees awarded in communication, journalism, and related fields has grown during this century, especially relative to other fields in social sciences and liberal arts. The number of higher education position announcements in communication in the United States has similarly increased over the same period, even accounting for a decline since the onset of the Covid-19 pandemic, keeping pace with doctoral degree awards (National Communication Association, 2023). Even as some valuable units in universities lose faculty lines or face administrative elimination, many departments, schools, and colleges of communication gain faculty or at least maintain their size. With many research faculty contracts allocating anywhere from a considerable minority to a majority of their work efforts to research in relation to teaching and service, the result is more faculty research. With communication units across the higher-education landscape often still in growth, serviced by more faculty conducting more faculty research, communication has a proportional duty to ensure that this snowballing knowledge production is leading to a body of scholarship in which we can be confident. The field's ranks and its production of research, are still growing. Thus, it is urgent that we ensure this growing body of research is as valid as it can be.

3.3. Membership in a Community of Scholars

Again, the relatively healthy demand for communication faculty in academia supports a growing number of doctoral students and encourages the presence of early-career faculty across campuses. These early-career faculty are under disproportionate pressure to have productive programs of research to obtain terminal degrees and related achievements such as habilitation, then reach career-securing benchmarks such as long-term contracts and tenure. While protections such as tenure are in part designed to allow senior scholars to explore uncertain intellectual seas in their scholarship, it is often the early-career scholars who are most likely to be working in new empirical and conceptual areas, enlightened by fresh perspectives and armed with up-to-date interdisciplinary literature from intensive graduate study at top doctoral-granting programs. Even with direct replications underutilized in communication and elsewhere, early-career researchers are likely to be working on scholarship that at least builds on the published work of others with some degree of replication effort. Therefore, it may be these "junior" scholars who suffer most in the field when prominent studies and areas of research are plagued with replicability and reproducibility issues; already toiling to generate quality research under time pressure, their efforts are stymied when they unknowingly build their careers on the shaky foundation of previous research that resists efforts at

replication. Our best new scholars build on the scholarship of others. Thus, it is urgent that we ensure they have replicable research on which to build.

4. Conclusion

While we debate solutions to issues with replicability and reproducibility in communication, we must not lose sight of the urgency of the problem. Voices have been raising concerns about replicability in communication research since before some of our leading scholars were born (e.g., Kelly et al., 1979), and these concerns remain. We may disagree on how best to address problems with replicability and reproducibility in communication research, but these cornerstones of science are essential to the continued credibility of our field among scholars and the public. Systemic problems with our research record lead to a misinformed public, a growing body of flawed studies, and mounting pressure on our most innovative and vulnerable young scholars. These are severe potential consequences for communication research and the scholars engaged in it. Thoughtful and productive debate over solutions is healthy for our field, but failing to act to ensure the consistent integrity of our research standards and published body scholarship is not an option. The path ahead may be uncertain, but the very real cost of inaction is clear.

Acknowledgments

The author would like to thank the editors of this thematic issue for contributing to the conversation about working together to produce a more replicable and reproducible body of literature in the field.

Conflict of Interests

The author declares no competing interests.

References

Artner, R., Verliefde, T., Steegen, S., Gomes, S., Traets, F., Tuerlinckx, F., & Vanpaemel, W. (2021). The reproducibility of statistical results in psychological research: An investigation using unpublished raw data. *Psychological Methods*, *26*(5), 527–546. https://doi.org/10.1037/met0000365

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. S., Lukito, J., Bier, L. M., Zhang, R., Johnson, B., 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. http://doi.org/10.1093/joc/jqz052

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLoS One*, *4*(5), Article e5738. https://doi.org/10.1371/journal.pone.0005738

Fox, J., Pearce, K. E., Massanari, A. L., Riles, J. W., Szulc, Ł., Ranjit, Y. S., Trevisan, F., Soriano, C. R. R., Vitak, J., Arora, P., Ahn, S. J., Alper, M., Gambino, A., Gonzalez, C., Lynch, T., Williamson, L. D., & Gonzales, A. L. (2021). Open science, closed doors? Countering marginalization through an agenda for ethical, inclusive research in communication. *Journal of Communication*, *71*(5), 764–784. https://doi.org/10.1093/joc/jqab029

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), Article e124. https://doi.org/10.1371/journal.pmed.0020124

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. https://doi.org/10.1177/0956797611430953

Kelly, C. W., Chase, L. J., & Tucker, R. K. (1979). Replication in experimental communication research:

An analysis. *Human Communication Research*, *5*(4), 338–342. https://doi.org/10.1111/j.1468-2958.1979.tb00646.x

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahnik, S., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "Many Labs" replication project. *Social Psychology*, *45*, 142–152. https://doi.org/10.1027/1864-9335/a000178

Markowitz, D. M., Song, H., & Taylor, S. H. (2021). Tracing the adoption and effects of open science in communication research. *Journal of Communication*, *71*(5), 739–763. https://doi.org/10.1093/joc/jqab030

Moonesinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published findings are false—But a little replication goes a long way. *PLoS Medicine*, *4*(2), Article e28. https://doi.org/10.1371/journal.pmed.0040028

National Communication Association. (2023). *2021–2022 academic job listings in communication report*. https://www.natcom.org/sites/default/files/NCA%20Jobs%20Report%202021-2022.pdf

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology*, *73*, 719–748. https://doi.org/10.1146/annurev-psych-020821-114157

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science 349*(6251), Article aac4716. https://doi.org/10.1126/science.aac4716

Samuelson, W., & Zeckhauser, R. (1988). Status quo bias in decision making. *Journal of Risk and Uncertainty*, *1*, 7–59. https://doi.org/10.1007/BF00055564

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Better p-curves: Making p-curve analysis more robust to errors, fraud, and ambitious p-hacking, a reply to Ulrich and Miller (2015). *Journal of Experimental Psychology: General*, *144*(6), 1146–1152. https://doi.org/10.1037/xge0000104

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., van de Velde, B., & Oegema, D. (2015). Blinded by the light: How a focus on statistical "significance" may cause p-value misreporting and an excess of p-values just below .05 in communication science. *Communication Methods and Measures*, *9*(4), 253–279. http://doi.org/10.1080/19312458.2015.1096333

About the Author

James D. Ivory is a professor of media studies in the Department of English at Virginia Tech. His scholarly interests deal with social and behavioral dimensions of media, particularly interactive digital media such as video games, simulations, and virtual environments, and research methods and practices in social and behavioral science.

Media and Communication
2024 • Volume 12 • Commentary 7935
https://doi.org/10.17645/mac.7935

COMMENTARY

Open Access Journal

On the Continued Need for Replication in Media and Communication Research

Nicholas David Bowman

S. I. Newhouse School of Communication, Syracuse University, USA

Correspondence: Nicholas David Bowman (nbowman@syr.edu)

Submitted: 1 December 2023 **Accepted:** 3 February 2024 **Published:** 25 March 2024

Issue: This commentary is part of the issue "Reproducibility and Replicability in Communication Research" edited by Johannes Breuer (GESIS—Leibniz Institute for the Social Sciences / Center for Advanced Internet Studies) and Mario Haim (LMU Munich), fully open access at https://doi.org/10.17645/mac.i429

Abstract

Common models of the scientific method articulate the processes by which we hypothesize about the correlation between variables and then test those predictions to make incremental conclusions about the world around us. Implied in this process is the replication and extension of that knowledge to various contexts. As with other social sciences, published analyses have demonstrated that media and communication scholarship suffers from a lack of replication studies, often due to presumptions about the lack of reward or incentive for conducting this work—such as perceived difficulties securing support for and/or publishing these studies. This commentary will reflect on and reinforce arguments for the intentional and important role of replication studies in media and communication scholarship. The essay reflects on replication as a key to post-positivist approaches, and then highlights recent developments that center replication work as key to scientific progression.

Keywords

open science; post-positivism; replication; research integrity; social sciences

1. Introduction

Among the many ways we understand the social sciences, we can look at two enduring maxims. The first is the Popperian perspective on falsification, that "every genuine test of a theory is an attempt to falsify it (or refute it)" (Popper, 1963, p. 36). The second is the Newtonian perspective on iterative science, that "if I have seen further, it is by standing on the shoulders of giants" (Newton, 1675). The former reminds us that theories and findings are tentative and accepted in the absence of more robust explanations, and the latter reminds us that social sciences build on each other—simply put, we need prior evidence and theory to advance on, but we must also consider the robustness of those prior advances, lest we advance on less than sturdy ground.

These twin perspectives are central to the notion of replication. The former reminds us that empirical claims are tentative and should be reexamined in the face of new evidence, and the latter reminds us that scientific progress is additive and iterative. Using these two maxims as points of departure, in this brief commentary I review and revive arguments for the importance of replication studies, framed as answers to questions we should be asking ourselves as we scrutinize and stabilize our science. Notably, this commentary focuses more directly on replications of prior research, rather than more direct reproduction efforts—the latter focused on using published data and code to verify published results (see LeBel et al., 2018). Notably, that same reference provides a more nuanced discussion of the different types of replication studies.

2. Replication as Post-Positivism?

Among many different approaches to social science, much of the empirical work in media and communication scholarship using quantitative data is rooted in a post-positivist perspective. The positivist approach promoted the use of the scientific method when engaging media and communication phenomena, using empirical observation and measurement to generate what was presumed to be verifiable, factual, and objective knowledge. Elegant on its face, the approach hardly represents the reality of scientific progress (see Kuhn, 1962) and as noted by Naveed (2014), "people are not beakers of water"—the units of analysis in media and communication research often involved variables that are measured imprecisely and that social reality is more volatile than physical reality, which might require some subjectivity in interpretation. As such, post-positivism allows for a measure of intersubjectivity. For example, post-positivist perspectives might interpolate between the findings of several scholars, rather than relying on a definitive claim from any one scholar (see Schutt, 2022); such claims also align with Merton's (1942) norms of science, including a reliance on communities of scholars with a skeptical view.

This intersubjectivity is key to notions of falsification and indeed highly relevant to the practice of replications. For Popper (1963), the hypothesis that "all swans are white" is disproven by the observance of a single black swan (or really, a swan of any color other than white). Tests of this hypothesis require more than a single dataset, however, just as it is unlikely that any one observation of swans would account for all swans in existence. Put another way, the inherent falsifiability of a hypothesis requires constant and persistent testing—observing swans wherever swans might (or might not) be known to exist. What we are essentially describing here is the process of replication, as replications are necessary for generating empirical data for which to continually test our hypotheses. When viewed this way, replication is not a novel concept but rather *a bedrock of the post-positivism*. Others have taken a similar position. For example, Iso-Ahola (2020) argues that replication studies are essentially "just another empirical investigation in an ongoing effort to establish scientific truth."

3. Replications as a Publication Paradox?

In recent years, one attempt to encourage replications has been through focused calls for papers, such as special issues in established journals—for example, a recent issue of *Communication Studies* "Special Issue for Replications" (volume 69, issue 3). Such efforts are invaluable tools for directly encouraging and spotlighting replication attempts and setting a clear standard for replication scholarship (see McEwan et al., 2018).

As important as special issues focused on replication are, I worry about the paradox that they might create—that replications are only feasible as publications if they are given extraordinary consideration and incentive. As suggested by Keating and Totzkay (2019) in their analysis of major communication journals from 2007–2016, less than 2% of all published research could be classified as a direct replication; in looking at psychology research since 1900, Makel et al. (2012) drop this number to about 1.6%, although these numbers have likely grown in recent years. Thus, on the one hand, special issues clearly encourage the submission and eventual publication of replication studies (especially if we presume that there were other manuscripts submitted to but ultimately not published with the special issue). On the other hand, special issues might encourage "one-off" replications wherein the replication efforts are taken up as an interesting but isolated challenge, rather than integrated into a programmatic line of research. Special issues could fetishize replications in ways that make them exotic rather than essential, and this is compounded given that so few journals *explicitly* call for replications in their calls for papers (as few as 3% of psychology journals; see Martin & Clarke, 2017). Special issues play a critical role in driving interest in replication research, but my hope is that they move replication research towards mainstream practice.

4. Do Replications Reap Rewards?

An enduring critique of replication studies is that their perceived lack of novelty makes them generally uninteresting for readers (and by extension, journals). That said, we can again look at the manuscripts in *Communication Studies* for evidence to the contrary. As of this writing (early 2024) and excluding the editorial on replications itself, the nine replication studies published in that issue have been collectively cited more than 160 times (according to Google Scholar). Moreover, successful replications draw needed and necessary attention to scholarship from broader audiences. One example of this is the "many labs" approach in which replication is a central aspect of the scientific process (Protzko et al., 2023)—when published, this article was in the 99th percentile of more than 245,000 articles tracked in terms of online engagement.

Replication studies are also quite useful for training early career scholars. As outlined by Janz (2016), replication studies are especially useful for graduate student training with respect to research methods and data analysis, as well as establishing replication as a routine approach to their scientific program. Such efforts, in tandem with arguments earlier in this essay, can also facilitate early career publications (for example, see Yoshimura et al., 2022) relevant to helping early career scholars establish themselves within a given area of research.

5. Conclusion

Replications reflect a core logos of post-positivism and as such, should be conducted and encouraged precisely because of what they are—introspective and reflective tests of reported findings to directly protect against the logical weaknesses of positivism (Tsang & Kwan, 1999). Scholars need not wait for an invitation to submit replication studies, either by way of special issues or stated shifts to editorial practices. Drawing from local experience (as editor-in-chief of the *Journal of Media Psychology*), the explicit inclusion of replications in our general call for papers has netted only three submissions in the last three years (one each in 2021, 2022, and 2023). One of those has been accepted for publication (Wright, 2022). As scholars, we have the agency to move towards a replication culture. Regarding "making the case" for a replication study, understand that the requirements of arguing for the novelty or necessity of a replication study are not so unique from the

arguments for arguing for the novelty or necessity of *any* study. Ironically, so-called novel studies (i.e., studies with "interesting" findings) do tend to be published and cited more often, but tend not to replicate and, thus, might not stand to extended scrutiny (Serra-Garcia & Gneezy, 2021). Regarding the labor of replication studies, emerging open science practices greatly facilitate the ability for scholars to replicate each other's research (see Dienlin et al., 2021) in ways that align with the philosophical norms of sciences (see Bowman & Keene, 2016). Drawing back to Yoshimura et al. (2022), that replication would not have been possible without having access to shared study materials and data from Eden et al. (2017).

In the late 20th century, an upstart advertising agency launched what would become one of the most iconic global branding campaigns in three simple words: "Just Do It" (López Restrepo, 2022). The slogan far outgrew the athletic apparel brand Nike, and we submit that it can and should be co-opted to address the consistently (over)stated need for replications in media and communication, and in social sciences broadly.

Replications? Just do them.

Acknowledgments

The author wishes to thank the editors for providing valuable insights in shaping the final presented argument.

Conflict of Interests

The author declares no conflicts of interest.

References

Bowman, N. D., & Keene, J. R. (2016). A layered framework for considering open science practices. *Communication Research Reports*, *35*(4), 363–372. https://doi.org/10.1080/08824096.2018.1513273

Dienlin, T., Johannes, N., Bowman, N. D., Masur, P. K., Engesser, S., Kümpel, A. N., 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

Eden, A., Daalmans, S., & Johnson, B. K. (2017). Morality predicts enjoyment but not appreciation of morally ambiguous characters. *Media Psychology*, *20*(3), 349–373. https://doi.org/10.1080/15213269.2016.1182030

Iso-Ahola, S. E. (2020). Replication and the establishment of scientific truth. *Frontiers in Psychology*, *11*, Article 2183. https://doi.org/10.3389/fpsyg.2020.02183

Janz, N. (2016). Bringing the gold standard into the classroom: Replication in university teaching. *International Studies Perspective*, *17*(4), 392–407. https://doi.org/10.1111/insp. 12104

Keating, D. M., & Totzkay, D. (2019). We do publish (conceptual) replications (sometimes): Publication trends in communication science, 2007–2016. *Annals of the International Communication Association*, *43*(3), 225–239. https://doi.org/10.1080/23808985.2019.1632218

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

LeBel, E. P., McCarthy, R. J., Earp, B. D., Elson, M., & Vanpaemel, W. (2018). A unified framework to quantify the credibility of scientific findings. *Advances in Methods and Practices in Psychological Science*, *1*(3), 389–402. https://doi.org/10.1177/2515245918787489

López Restrepo, M. (2022, October 6). Just do it: How the iconic Nike tagline built a career for the late Dan Wieden. *National Public Radio (NPR)*. https://www.npr.org/2022/10/06/1127032721/nike-just-do-

it-slogan-success-dan-wieden-kennedy-dies
Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688
Martin, G. N., & Clarke, R. M. (2017). Are psychology journals anti-replication? A snapshot of editorial practices. *Frontiers in Psychology*, *8*, Article 523. https://doi.org/10.3389/fpsyg.2017.00523
McEwan, B., Carpenter, C. J., & Westerman, D. (2018). On replication in communication science. *Communication Studies*, *69*(3), 235–241. https://doi.org/10.1080/10510974.2018.1464938
Merton, R. (1942). *The sociology of science: Theoretical and empirical investigations*. University of Chicago Press.
Naveed, F. (2014, May 29). Defining theory: Post-positivism-hermeneutic-critical-normative. *Mass Communication Talk*. https://www.masscommunicationtalk.com/defining-theory-post-positivism-hermeneutic-critical-normative.html
Newton, I. (1675). *Letter to Robert Hooke*. https://digitallibrary.hsp.org/index.php/Detail/objects/9792
Popper, K. (1963). *Science as falsification*. Routledge & Keagan Paul.
Protzko, J., Krosnick, J., Nelson, L., Nosek, B. A., Axt, J., Berent, M., Buttrick, N., DeBell, M., Ebersole, C. R., Lundmark, S., MacInnis, B., O'Donnell, M., Perfecto, H., Pustejovsky, J. E., Roeder, S. S., Walleczek, J., & Schooler, J. W. (2023). High replicability of newly discovered social-behavioural findings is achievable. *Nature Human Behavior*, *8*, 311–319. https://doi.org/10.1038/s41562-023-01749-9
Schutt, R. K. (2022). *Investigating the social world* (10th ed.). SAGE.
Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Sciences Advances*, *7*(21), Article eabd1705. https://doi.org/10.1126/sciadv.abd1705
Tsang, E. W. K., & Kwan, K.-M. (1999). Replication and theory development in organizational science: A critical realist perspective. *The Academy of Management Review*, *24*(4), 759–780. https://doi.org/10.2307/259353
Wright, P. J. (2022). A longitudinal analysis of political ideology, pornography consumption, and attitude change: Replication and extension. *Journal of Media Psychology*, *35*(6), 346–354. https://doi.org/10.1027/1864-1105/a000370
Yoshimura, K., Bowman, N. D., Cohen, E. L., & Banks, J. (2022). Character morality, enjoyment, and appreciation: A replication of Eden, Daalmans, and Johnson (2017). *Media Psychology*, *25*(2), 181–207. https://doi.org/10.1080/15213269.2021.1884096

About the Author

Nicholas David Bowman (PhD, Michigan State University) is an associate professor in the S. I. Newhouse School of Public Communications at Syracuse University, USA. His research considers and examines the cognitive, emotional, physical, and social demands of interactive media such as video games and extended reality technologies. He has published more than 150 peer-reviewed manuscripts and is the current editor of *Journal of Media Psychology*. His most recent book is the edited volume, *Entertainment Media and Communication* handbook (De Gruyter Mouton).

MEDIA AND COMMUNICATION
ISSN: 2183-2439

Media and Communication is an international, peer-reviewed open access journal dedicated to a wide variety of basic and applied research in communication and its related fields. It aims at providing a research forum on the social and cultural relevance of media and communication processes.

The journal is concerned with the social development and contemporary transformation of media and communication and critically reflects on their interdependence with global, individual, media, digital, economic and visual processes of change and innovation.

cogitatio

www.cogitatiopress.com/mediaandcommunication