Humanities & Social Sciences Communications

COMMENT

https://doi.org/10.1057/s41599-025-04381-2 OPEN

Let's talk about language—and its role for replicability

Xenia Schmalz[1]✉, Johannes Breuer[2], Mario Haim[3], Andrea Hildebrandt[4], Philipp Knöpfle[3], Anna Yi Leung[1] & Timo Roettger[5]

Science strives towards a credible and comprehensive understanding of the world around us. Across disciplines within the social and behavioural sciences (and beyond), limitations in the implementation of the scientific approach have been identified in recent studies, showing low replicability of many results. This is an issue for knowledge accumulation, theory-building, and evidence-based decision and policy making. Researchers have proposed several solutions to address these issues, focusing mainly on improving statistical methods, data quality, and transparency. However, relatively little attention has been paid to another key aspect that affects replicability: language. Across fields, language plays a central role in all steps of the research cycle and is a critical communication tool among researchers. Neglecting its role may reduce replicability and limit our understanding of theoretically interesting differences and similarities across languages. After identifying these challenges, we provide some recommendations and an outlook on how replicability challenges related to language may be addressed.

[1] Department of Child and Adolescent Psychiatry, Psychosomatics, and Psychotherapy, University Hospital, Ludwig-Maximilians-University of Munich, Munich, Germany. [2] GESIS—Leibniz Institute for the Social Sciences, Cologne, Germany. [3] Department of Media and Communication, Faculty of Social Sciences, Ludwig-Maximilians-University of Munich, Munich, Germany. [4] Department of Psychology, School of Medicine and Health Sciences, Carl von Ossietzky Universität Oldenburg, Oldenburg, Germany. [5] Department of Linguistics and Scandinavian Studies, Faculty of Humanities, University of Oslo, Oslo, Norway. ✉email: xenia.schmalz@gmail.com

The overarching goal of research is to produce knowledge. This involves ensuring that the accumulated knowledge is applicable to various research contexts as well as real-world settings. Across academic disciplines, one key criterion for achieving this is maximising the replicability of research. The exact definition of replicability varies across contexts; here, we follow the FORRT Glossary of Open Science Terms and the Turing Way Community, and use a broad definition of replicability as reaching the same conclusions when repeating a study with the same methods but new data (Parsons et al. 2022; The Turing Way Community, 2022). Replicability is typically distinguished from reproducibility, or reaching the same results when repeating the analysis of a study with both the same methods and the same data. Crucially, while the exact definitions might differ slightly across disciplines, a lack of replicability, in its broad sense, has recently been identified for large sets of studies in psychology (Open Science Collaboration, 2015; Klein et al. 2022), medicine (Ioannidis, 2005), economics (Camerer et al. 2016), and the behavioural and social sciences more generally (Camerer et al. 2018). Efforts to increase replicability rates have recently been discussed at great length, with suggestions to increase transparency (e.g., Asendorpf et al. 2013), engage in preregistration (e.g., Nosek et al. 2022), and apply more rigorous statistical methods (e.g., Simmons et al. 2021). At the same time, the in-depth examination of replicability has also put the credibility of science as a whole to the test, calling for a "credibility revolution" (Angrist and Pischke, 2010; Korbmacher et al. 2023), as well as more for purposeful communication of scientific uncertainty to the public (e.g., Howell, 2020). Communication, both with the public and with fellow researchers, depends crucially on natural language, which is inherently ambiguous and multifaceted. We argue that the improper or negligent use of language can pose another major challenge to replicability. In the discussions on replicability, this challenge has not received much attention yet. Language plays a central role throughout the research process—from theory formulation, study design, and data collection, all the way to documentation and dissemination. As such, we call attention to its critical role and the ways in which it interacts with replicability.[1]

Language as a medium of research

In research, language is the primary medium that conveys meaning to its users. Thus, its presence across the research process is ubiquitous. It is used, for example, to search or summarise existing literature, to define technical jargon accurately, to formulate research questions and hypotheses, and to communicate results and interpretations. However, natural language can be imprecise, ambiguous, and context-dependent (Leung et al. 2024), which can pose challenges to replicability. For example, ambiguous formulations of research hypotheses can affect replicability in two distinct ways (Scheel, 2022). First, it can lead to different interpretations of verbal elements within the hypothesis, resulting in different conceptions of how the hypothesis should be tested and how data should be interpreted to evaluate it. Two researchers testing the supposedly same hypothesis, thus, might end up with different results (e.g., Auspurg and Brüderl, 2021). Second, a vague hypothesis allows researchers more degrees of freedom when analysing and interpreting their data. Multi-lab studies have shown that these researcher degrees of freedom can lead to multiple possible analytical strategies, which often yield categorically different results (Silberzahn et al. 2018), blur the line between confirmatory and exploratory research, and might drastically inflate false positive rates (e.g., Simmons et al. 2011).

The intrinsic imprecision of language is exacerbated in academic communication contexts with language users who differ in their linguistic and cultural backgrounds (Vander Beken et al. 2020). Regardless of the language employed for scientific discourse, readers from different backgrounds will have varying degrees of proficiency in that target language and exhibit different degrees of distance to cultural references and conventions, and their academic backgrounds will affect their interpretation of technical jargon. In light of these considerations, the intrinsically imprecise nature of language as a medium for scientific communication might constitute a crucial factor in the observed low replicability rates.

Language as a tool for research

Language is also a central part of the research toolkit. It is integral for designing a study, preparing materials, and collecting and analysing data. For example, in the social and behavioural sciences, language is necessary for designing survey items and interview questions, as a modality to present experimental stimuli, or to deliver instructions to participants and informants. Language is also used to provide instructions among researchers (or research assistants), such as for experimental procedures, protocols, or data processing and documentation. When replicating a given study in a different language, the researcher must ensure that the translated materials are clear to understand, while ascertaining that the translated texts still capture the meanings as intended in the original study's measures.

This is especially important for measurements in humans, where the translations are expected to measure the exact same constructs as in the original tests. Yet, this is often challenging in cross-linguistic studies, as the test materials might have been conscientiously translated but not tested for measurement invariance (Klein et al. 2018; Luong and Flake, 2023). The linguistic properties of the translated materials might differ from the originals, which might affect participants' understanding of the test items' or instructions' meanings. This leads to measurement non-invariance, which means that the psychometric properties of items or questions are not equivalent and therefore the results are not comparable. In closely related languages and cultures (e.g., Dutch - German - English), examining and comparing the measurement properties of the tools in quantitative analysis may be sufficient; however, additional considerations are needed when conducting research in dissimilar languages and cultures. This requires a qualitative examination of aspects of the tool that may be perceived differently in the context and language of interest, or that may not apply at all. For example, participants across different populations might have varied comprehension proficiency of spoken, written or signed messages across cultural, educational, or clinical backgrounds. Moreover, direct translation may have different implications across cultures: cultural expectations on what is taboo might limit some types of stimuli; for example, it is inappropriate to show participants alcohol-related words in Arabic-speaking countries. This example shows that test materials with potentially taboo contents cannot be directly translated and applied in studies across cultures, limiting the breadth of cross-cultural and cross-linguistic replications. This is especially relevant for research where language is an object of study (see next section), but it also affects other types of research.

These challenges might increase with the cultural and linguistic distance between a population of interest and WEIRD populations on which much of the social and behavioural sciences are based (Henrich et al. 2010; Blasi et al. 2022). This is due to the inherent linguistic and cultural differences in material translations, which might thus impede replication studies across languages from obtaining comparable results.

While we have considered natural languages so far, some considerations also need to be extended to programming

languages. In most scientific disciplines, the use of computer code is common for creating research software, and for collecting, processing, or analysing data. While there are several structural differences between the two regarding replication, many of the issues described for human languages also apply to the use of programming languages. To replicate a study, the replicator needs to be able to reproduce and, hence, understand each relevant decision made for the original study. If computer code is used to collect, process, or analyse data, other researchers must be able to comprehend what was done to use this information for their replication work. This requires ensuring that the code is accessible and keeping it well-documented for the use of other researchers. Investigations into the computational reproducibility of research have revealed that, oftentimes, even the requirements for reproducing results using the same code are not met because the code is either not shared or not properly documented (Perkel, 2020; Krähmer et al. 2023).

Notably, different researchers also use different tool stacks, including different programming languages. As with human languages, translations are possible, but they are often associated with some degree of conversion loss. A particular problem that is unique to the realm of programming languages is the use of proprietary solutions that not every researcher has access to, a limitation that disproportionately affects some researchers more than others. Overall, for the cases of both human, as well as programming languages, it is clear that language as a tool for research can, in several ways, introduce difficulties in replications due to challenges related to translating, conveying, and preserving meaning.

Language as an object of study

In several disciplines, language itself or the role of language in cognition, society, and culture are objects of scientific inquiry. When language is the object of study, its role in replicability takes on another, more theoretically relevant dimension compared to the issues in language as a communication or research tool, which we discuss above. Specifically, low replicability when language is an object of study may reflect a lack of generalisability across languages, rather than methodological artefacts.

Replication in cross-linguistic research. The role of language is ubiquitous in everyday life; thus, it is important to understand aspects such as language development and the use of language in documenting cultural knowledge. Language is a culturally evolved, complex adaptive system (Winter, 2014) that interacts with a large variety of human experiences. The structure of languages can differ substantially (Evans and Levinson, 2009), and these differences may affect other parts of cognition, such as working memory (Amici et al. 2019), attention (Wang, 2021), and perception (Kemmerer, 2023). In light of linguistic diversity and its complex interaction with cognition and behaviour, the question arises as to whether we should always expect findings to be replicable when a study is conducted in a different language or culture. Does a failed replication across languages suggest non-replicability, non-generalisability, or merely that the phenomenon in one language cannot be investigated with the same study design, measurement, sample selection, or materials in another language? For example, some research suggests that the processing and acquisition of nouns might differ from verbs (e.g., Cazden, 1968; Maratsos, 2013). Replicating such an asymmetry across different languages can be challenging or even impossible, because distributional, semantic, and morphological properties of categories such as nouns and verbs can drastically differ across languages, with some languages having been described as lacking such a distinction (Sasse, 2001). Covariates related to culture and social factors that are intricately connected to the language we speak may render a failed replication uninterpretable: it becomes unclear if a failed replication constitutes evidence against the original finding or a limitation of the context in which this finding can be obtained (Grieve, 2021; Roettger, 2021a). For example, showing that an intervention aiming to improve reading skills in children with developmental dyslexia in English does not work in German may be due to the non-replicability of the original English study, but it may be that the characteristics of the German language, such as the morphological complexity or the closer relationship between print and speech sounds, yield the intervention ineffective. Thus, without conducting further research, it is difficult to draw conclusions from such a failure to replicate.

Language as data. Challenges regarding replicability go beyond questions of translation when language is an object of study. These are relevant not only for replicability in the context of cross-linguistic research, but also when research is replicated or reproduced within a language. In areas such as communication sciences and linguistics, for example, audio or video recordings of language production, news articles, social media posts, or podcasts may be used as data sources. Depending on the source and type of data being used, research on language as an object of study often requires preprocessing steps, which can be complex and resource-intensive (in terms of time and/or required computing resources). Typical preprocessing steps include the transcription of audio material into text, manual or (semi-)automated coding/classification of content/text, or applying natural language processing (NLP) pipelines, such as for part-of-speech (POS) tagging or named entity recognition (NER). The pipelines are mostly developed for the most-studied languages; as such, resources for under-studied languages may not exist or be of lower quality, as there is less available data that can be used for their development (e.g., Chilson et al. 2024).

Similar to data analysis, preprocessing of linguistic data entails various researcher degrees of freedom, which can be particularly impactful for complex processing pipelines that are common for text- or even more so for audio- and video-as-data studies (Coretta et al. 2023; Lukito et al. 2024). Of course, the issue of translation and potential conversion loss, akin to the challenges faced when language is used as a tool in research, also warrants consideration in this context. Generally, if we rely on specific tools or tool chains in language-as-data settings, we need to properly document those. The methods used for processing and analysing data are hence especially important. Besides documentation, an important step that is often somewhat neglected in research with text as data is the validation of methods (Birkenmaier et al. 2023). For instance, during preprocessing procedures such as POS tagging, data validation may involve implementing cross-validation techniques, wherein the POS tags generated by the NLP pipeline are systematically compared to a manually annotated "gold standard" dataset to quantitatively assess accuracy. Another example for validation in a language-as-data setting could constitute a review by linguistic experts, who examine a representative sample of the text data to ensure that the automated tagging generated via NLP aligns with expert human annotations. When working with under-studied languages, researchers might need to validate the NLP pipeline by checking for biases or errors unique to that language. This could involve running an error analysis on the output to identify common misclassifications and refining the pipeline accordingly. The validation of computational text analysis methodologies has become increasingly critical with the proliferation of artificial intelligence (AI) tools, particularly large language models (LLMs).

The reliability and validity of annotations or classifications generated by these technologies have already been demonstrated to present significant challenges (Kristensen-McLachlan et al. 2023; Pangakis et al. 2023; Reiss, 2023).

For research using language as data, however, issues related to replicability are not limited to data processing and analysis. Another domain that can produce replicability challenges is data access. Commonly used data sources, such as audio or video recordings, news texts, and social media content, are often proprietary and controlled by (commercial) third parties, such as media organisations and online platforms. A prevalent issue in work with textual but also image, audio, and video data in many fields across the social and behavioural sciences as well as the humanities is the change or closure of application programming interfaces (APIs) of platforms through which researchers can access data. Besides being possibly proprietary, video, audio, and text data are often also sensitive or involve data that is culturally inappropriate. These two key attributes introduce legal and ethical concerns regarding copyright and intellectual property and can impact research replicability, especially also when it comes to data sharing. For the specific case of data from social media, Davidson et al. (2023) have recently argued that "[…] platform-controlled social media APIs threaten open science […]" and studies by Küpfer (2024) and Knöpfle and Schatto-Eckrodt (2024) have demonstrated that the replicability of studies based on data from Twitter is strongly and negatively impacted by changes in the platform APIs and restrictions imposed on data sharing in their Terms of Service (ToS).

Recommendations and ways forward

Based on the considerations in the previous sections, we aim to put forth recommendations for addressing challenges for replicability related to language as a) a medium of research, b) a tool for research, and c) an object of study.

Community-driven refinement of term definitions for clearer conceptualisation. Considering language as a medium of research, jargon is unavoidably used for conveying complex ideas. Technical terms, theoretical descriptions, and research questions must be as precise as possible to communicate effectively. To improve replicability, the first step in this process is identifying terms that are inherently ambiguous or lack consensual definitions in the literature (Leung et al. 2024). Such terms tend to be more challenging to operationalise, which may lead to differences in measurement across studies and subsequently affect the replicability of the results. Reaching a broader consensus on the interpretation of technical terminology can support a more structured approach to theory formulation. This would require collective effort within scientific fields and communities to define and refine consensual scientific term definitions iteratively across time (Leising et al. 2024; see also Parsons et al. 2022 for a successful crowd-sourced glossary of term definitions).

Formalisation of research questions and hypotheses for effective communication. After examining, defining, and agreeing on the specific attributes of the concepts involved, researchers can create stronger connections between empirical evidence and theoretical predictions (e.g., Scheel, 2022). For example, using transparent and formalised formats to pose specific and machine-readable research questions and hypotheses could help increase the falsifiability of hypothesis tests (e.g., Lakens and DeBruine, 2021). Such hypothesis specifications do not only capture the conceptual descriptions of our predictions but also the operationalisation and the statistical predictions of the empirical tests. This can avoid using solely verbal descriptions to make hypotheses, thus reducing the degrees of freedom between the conceptual descriptions and the operationalisation or statistical predictions.

Increasing linguistic precision may, to a certain extent, rely on the use of statistical and mathematical expressions to capture a prediction. However, this may sacrifice the ease of communicating scientific results across disciplines, as well as to the public, when using technical expressions rather than the layman's language to disseminate scientific information (Bullock et al. 2019). Switching between a statistics-oriented scientific language system and a layman's language system to communicate research findings may impose difficulties in knowledge transfer and communication among researchers and the public. Both language systems are, however, equally important as the media of research and can co-exist to serve different audiences, be it the researchers of other fields or the general public. Hence, we call for the enhancement of proficiencies in both scientific language and science communication with the public in higher education. Researchers would then become more equipped with the skills to communicate science to peer researchers using formalised scientific language while disseminating information to the public in non-technical language.

Material and data sharing for comparable replications across communities. When language is a tool for research, to increase the comparability of replications across different languages, future work could focus on making high-quality resources available for under-studied languages. This involves developing and evaluating the quality, equivalence and applicability of research tools to different languages and generating language-specific instruments when a direct transfer to the different linguistic and cultural contexts is not possible. To achieve this, the scientific community should strive towards openness, not only by sharing already existing instruments but also by documenting and sharing the steps taken in their development. The measurement invariance of these tools across languages is the critical methodological issue to be addressed (e.g., Meredith, 1993). As a first step, researchers should consider if it is appropriate to apply the same tool in another language. This step requires close collaboration with researchers who are very familiar with this culture, ideally, who grew up in it and speak the language(s) fluently. Active exchange with the community will allow researchers to take cultural and linguistic differences into account appropriately.

Development of invariant measurement for cross-linguistic replications. As a second step, researchers should conduct quantitative analyses to ensure measurement invariance. Such analyses are based on multi-group confirmatory factor analytic methods (see Hildebrandt et al. 2016, for details and extensions to nonlinear approaches), which, through parameter restrictions, allow for testing the equivalence of item difficulty, discriminative power, and item reliability within a measurement tool across languages. These qualitative and then quantitative analysis steps combined will allow researchers globally to create and adjust tools that can be used in their own languages and, thus, potentially contribute to reducing the WEIRD problem in research with human participants.

Promotion of reusable and interoperable use of programming languages. With regard to programming languages, we urge the adoption and promotion of practices that increase reusability and interoperability, such as proper documentation (e.g., via annotating research material and code and through version control of all software and research-related tools), as well as avoiding proprietary closed-source solutions. In addition, we emphasise the

recommendations of previous scholars to rely on free and open-source tools for scientific research (e.g., Asendorpf et al. 2013).

Final remarks: Big Team Science drives open science and large-scale replications. Language as an object of scientific inquiry warrants both strong quantitative and mechanistic theories on how language, behaviour, and cognition interact in general, and how language-specific traits moderate these interactions. Without such efforts, the field lacks a principled way of integrating empirical findings and, ultimately, advancing our understanding of human language and related areas in an effective manner (Roettger, 2021b). If scientists fail to (or cannot) specify the contexts where a given effect is replicable, and if they dismiss failed replications due to context sensitivity, scientific progress is seriously impeded (Simmons et al. 2011). Theory building and data collection form a closed loop; as such, large-scale replication efforts should be conducted involving researchers dispersed across geographic locations, languages, and cultures. For example, the recently launched ManyLanguages consortium (many-languages.com) aims to directly replicate experimental findings related to language sciences across many languages (Faytak et al. 2024).

This recommendation feeds into our next suggestion, which is relevant whenever language is used as data: as we discussed in the paragraph above, for language as a tool, we need to develop tools for processing and analysing language data (text and audio) for multiple languages. Notably, this cannot be done without a large-scale initiative to produce sufficient and accessible written materials in each language for the continuous development of these study resources across contexts in the first place. In addition, when processing and analysing language data, for reasons of transparency and accessibility, open-source tools should be given preference and all steps in the pipeline should be properly documented and explained. The importance of documentation and the use of open-source solutions also extends to the use of programming languages for research in the social and behavioural sciences in general.

We encourage researchers to attempt replications across different countries and languages, even when language is not the primary focus of the study. While linguistic and cultural variations introduce complexities, they should not obstruct cross-cultural replication efforts. Instead, we suggest that researchers aim to account for context-specific factors that may affect the generalisability of their findings and to provide clear, comprehensive documentation of methodologies and potentially relevant contextual variables. Collaborative efforts across diverse cultural and linguistic contexts are essential for enhancing the robustness of research and an important step towards improving the generalisability of scientific findings.

Received: 14 August 2024; Accepted: 13 January 2025;
Published online: 25 January 2025

Note

1 Notably, many of the aspects we discuss in this paper are also relevant for reproducibility. However, our focus here is on replicability, so we will not specifically discuss the role of language for reproducibility.

References

Amici F, Sánchez-Amaro A, Sebastián-Enesco C et al. (2019) The word order of languages predicts native speakers' working memory. Sci Rep 9(1):1124. https://doi.org/10.1038/s41598-018-37654-9

Angrist JD, Pischke J-S (2010) The credibility revolution in empirical economics: How better research design is taking the con out of econometrics. J Econ Perspect 24(2):3–30. https://doi.org/10.1257/jep.24.2.3

Asendorpf JB, Conner M, De Fruyt F et al. (2013) Recommendations for increasing replicability in psychology. Eur J Pers 27(2):108–119. https://doi.org/10.1002/per.1919

Auspurg K, Brüderl J (2021) Has the credibility of the social sciences been credibly destroyed? Reanalysing the "many analysts, one data set" project. Socius 7:23780231211024421. https://doi.org/10.1177/23780231211024421

Birkenmaier L, Lechner CM, Wagner C (2023) The search for solid ground in text as data: A systematic review of validation practices and practical recommendations for validation. Commun Methods Meas 0(0):1–29. https://doi.org/10.1080/19312458.2023.2285765

Blasi DE, Henrich J, Adamou E et al. (2022) Over-reliance on English hinders cognitive science. Trends Cogn Sci 26(12):1153–1170. https://doi.org/10.1016/j.tics.2022.09.015

Bullock OM, Colón Amill D, Shulman HC et al. (2019) Jargon as a barrier to effective science communication: evidence from metacognition. Public Underst Sci 28(7):845–853. https://doi.org/10.1177/0963662519865687

Camerer CF, Dreber A, Forsell E et al. (2016) Evaluating replicability of laboratory experiments in economics. Science 351(6280):1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer CF, Dreber A, Holzmeister F et al. (2018) Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. Nat Hum Behav 2(9):637–644. https://doi.org/10.1038/s41562-018-0399-z

Cazden CB (1968) The acquisition of noun and verb inflections. Child Dev 39(2):433–448. https://doi.org/10.2307/1126956

Chilson S, Sineva E, Schmalz X (2024) FILMS: a multilingual word frequency corpus based on film subtitles with IPA transcriptions. OSF. Preprint at https://doi.org/10.31219/osf.io/zy5qf

Coretta S, Casillas JV, Roessig S et al. (2023) Multidimensional signals and analytic flexibility: estimating degrees of freedom in human-speech analyses. Adv Meth Pract Psychol Sci 6(3):25152459231162567. https://doi.org/10.1177/25152459231162567

Davidson BI, Wischerath D, Racek D et al. (2023) Platform-controlled social media apis threaten open science. Nat Hum Behav 7(12):2054–2057. https://doi.org/10.1038/s41562-023-01750-2

Evans N, Levinson SC (2009) The myth of language universals: Language diversity and its importance for cognitive science. Behav Brain Sci 32(5):429–448. https://doi.org/10.1017/S0140525X0999094X

Faytak M, Kadavá S, Xu C et al. (2024) Big team science for language science: Opportunities and challenges. OSF. Preprint at http://osf.io/c9w5b

Grieve J (2021) Observation, experimentation, and replication in linguistics. Linguistics 59(5):1343–1356. https://doi.org/10.1515/ling-2021-0094

Henrich J, Heine SJ, Norenzayan A (2010) The weirdest people in the world? Behav Brain Sci 33(2–3):61–83. https://doi.org/10.1017/S0140525X0999152X

Hildebrandt A, Lüdtke O, Robitzsch A et al. (2016) Exploring factor model parameters across continuous variables with local structural equation models. Multivariate behavioral research 51(2–3):257–258. https://doi.org/10.1080/00273171.2016.1142856

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

Ioannidis JPA (2005) Why most published research findings are false. PLOS Med 2(8):e124. https://doi.org/10.1371/journal.pmed.0020124

Kemmerer D (2023) Grounded cognition entails linguistic relativity: A neglected implication of a major semantic theory. Top Cogn Sci 15(4):615–647. https://doi.org/10.1111/tops.12628

Klein RA, Cook CL, Ebersole CR et al. (2022) Many labs 4: Failure to replicate mortality salience effect with and without original author involvement. Collabra: Psychol 8(1):35271. https://doi.org/10.1525/collabra.35271

Klein RA, Vianello M, Hasselman F et al. (2018) Many labs 2: Investigating variation in replicability across samples and settings. Adv Meth Pract Psychol Sci 1(4):443–490. https://doi.org/10.1177/2515245918810225

Knöpfle P, Schatto-Eckrodt T (2024) The challenges of replicating volatile platform-data studies: Replicating Schatto-Eckrodt et al. (2020). MaC 12:7789. https://doi.org/10.17645/mac.7789

Korbmacher M, Azevedo F, Pennington CR et al. (2023) The replication crisis has led to positive structural, procedural, and community changes. Commun Psychol 1(1):1–13. https://doi.org/10.1038/s44271-023-00003-2

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

Kristensen-McLachlan RD, Canavan M, Kardos M et al. (2023) Chatbots are not reliable text annotators. arXiv. Preprint at https://doi.org/10.48550/arXiv.2311.05769

Küpfer A (2024) NonRandom tweet mortality and data access restrictions: Compromising the replication of sensitive twitter studies. Polit. Anal.:1–14. https://doi.org/10.1017/pan.2024.7

Lakens D, DeBruine LM (2021) Improving transparency, falsifiability, and rigor by making hypothesis tests machine-readable. Adv Meth Pract Psychol Sci 4(2):2515245920970949. https://doi.org/10.1177/2515245920970949

Leising D, Liesefeld H, Buecker S et al. (2024) A tentative roadmap for consensus building processes. Pers Sci 5:1–5. https://doi.org/10.1177/27000710241298610
Leung AY, Melev I, Schmalz X (2024) Quantifying concept definition heterogeneity in academic texts: Insights into variability in conceptualisation. OSF. Preprint at https://doi.org/10.31219/osf.io/gu7b5
Lukito J, Greenfield J, Yang Y et al. (2024) Audio-as-data tools: Replicating computational data processing. MaC. 12(0). https://doi.org/10.17645/mac.7851
Luong R, Flake JK (2023) Measurement invariance testing using confirmatory factor analysis and alignment optimisation: A tutorial for transparent analysis planning and reporting. Psychol Methods 28(4):905–924. https://doi.org/10.1037/met0000441
Maratsos MP (2013) How the acquisition of nouns may be different from that of verbs. In: Krasnegor NA et al. (eds) Biological and behavioral determinants of language development, Psychology Press, pp. 77–98
Meredith W (1993) Measurement invariance, factor analysis and factorial invariance. Psychometrika 58(4):525–543. https://doi.org/10.1007/BF02294825
Nosek BA, Hardwicke TE, Moshontz H et al. (2022) Replicability, robustness, and reproducibility in psychological science. Annu Rev Psychol 73(1):719–748. https://doi.org/10.1146/annurev-psych-020821-114157
Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349(6251):aac4716. https://doi.org/10.1126/science.aac4716
Pangakis N, Wolken S, Fasching N (2023) Automated annotation with generative AI requires validation. arXiv. Preprint at https://doi.org/10.48550/arXiv.2306.00176
Parsons S, Azevedo F, Elsherif MM et al. (2022) A community-sourced glossary of open scholarship terms. Nat Hum Behav.:1–7. https://doi.org/10.1038/s41562-021-01269-4
Perkel JM (2020) Challenge to scientists: Does your ten-year-old code still run? Nature 584(7822):656–658. https://doi.org/10.1038/d41586-020-02462-7
Reiss MV (2023) Testing the reliability of ChatGPT for text annotation and classification: A cautionary remark. arXiv. Preprint at https://doi.org/10.48550/arXiv.2304.11085
Roettger TB (2021a) Pre-registration in experimental linguistics: Applications, challenges, and limitations. Linguistics 59(5):1227–1249. https://doi.org/10.1515/ling-2019-0048
Roettger TB (2021b) Context sensitivity and failed replications in linguistics – a reply to grieve. Linguistics 59(5):1357–1358. https://doi.org/10.1515/ling-2020-0239
Sasse H-J (2001) Scales between nouniness and verbiness. Lang Typo Lang Univ 1:495–509
Scheel AM (2022) Why most psychological research findings are not even wrong. Infant Child Dev 31(1):e2295. https://doi.org/10.1002/icd.2295
Silberzahn R, Uhlmann EL, Martin DP et al. (2018) Many analysts, one data set: Making transparent how variations in analytic choices affect results. Adv Meth Pract Psychol Sci 1(3):337–356. https://doi.org/10.1177/2515245917747646
Simmons JP, Nelson LD, Simonsohn U (2011) False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 22(11):1359–1366. https://doi.org/10.1177/0956797611417632
Simmons JP, Nelson LD, Simonsohn U (2021) Pre-registration: Why and how. J Consum Psychol 31(1):151–162. https://doi.org/10.1002/jcpy.1208
The Turing Way Community (2022) The Turing Way: A handbook for reproducible data science. Zenodo. https://doi.org/10.5281/zenodo.3233853
Vander Beken H, De Bruyne E, Brysbaert M (2020) Studying texts in a non-native language: A further investigation of factors involved in the L2 recall cost. Q J Exp Psychol 73(6):891–907. https://doi.org/10.1177/1747021820910694
Wang Q (2021) The cultural foundation of human memory. Annu Rev Psychol 72:151–179. https://doi.org/10.1146/annurev-psych-070920-023638
Winter B (2014) Spoken language achieves robustness and evolvability by exploiting degeneracy and neutrality. Bioessays 36(10):960–967. https://doi.org/10.1002/bies.201400028

Acknowledgements

This work was supported by multiple grants from the German Research Foundation (DFG) to XS (464350745), JB (464291459), MH (441890184), and AH (464552782), as part of the DFG priority programme, "META-REP: A Metascientific Programme to Analyse and Optimise Replicability in the Behavioural, Social, and Cognitive Sciences" (SPP 2317, project number 441890184). X.S. was further supported by a grant from the German Research Foundation (456356582). We have no other funding or competing interests to declare.

Author contributions

All authors contributed equally to this work. XS initiated the collaboration and was listed as the first author. All other authors are listed in alphabetical order. Correspondence to XS.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

This work has been sponsored by the German Research Foundation (DFG). We have no other funding or competing interests to declare.

Ethical approval

Ethical approval was not required as this article does not contain any studies with human participants performed by any of the authors.

Informed consent

Informed consent was not required as this article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Xenia Schmalz.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.