DE GRUYTER

Linguistics: An
Interdisciplinary Journal
of the Language Sciences

Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research

Journal:	*Linguistics: An Interdisciplinary Journal of the Language Sciences*
Manuscript ID	LING.2023.0249.R1
Manuscript Type:	Miscellaneous

Abstract

In recent years, numerous fields of research have seen a push for increased reproducibility and transparency. As a result, specific transparency practices have emerged, such as open access publishing, preregistration, sharing data, analyses, and code, performing study replications, and declaring positionality and conflicts of interest. While many agree that open science practices represent a positive step forward in improving scientific rigor, these practices, by and large, have not been adopted in the field of linguistics (Bochynska et al., 2023). Few, if any, researchers have had explicit instruction on the practices of open science as part of their professional training. Nonetheless, today's speech researcher is expected to be up to date on the current protocols of open science in order incorporate the methodological practices aimed at improving reproducibility/replicability. The present work intends to help make open science practices understandable and accessible to researchers in linguistics from all backgrounds and at every stage, from students/early career researchers to senior researchers and advisors. We outline eight specific open science practices that linguists can adopt to make their research more open, transparent, inclusive, and accessible to a wider audience.

Keywords: Open science, Reproducibility, Replicability, Transparency, Positionality, Linguistics

Word count: 7,455

Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research

Introduction - What is open science?

In recent years, numerous fields of research have seen a push for increased reproducibility and transparency practices. These practices, collectively, have been referred to as open science. Parsons et al. (2022) refer to open science as an umbrella term "[…] reflecting the idea that scientific knowledge of all kinds, where appropriate, should be openly accessible, transparent, rigorous, reproducible, replicable, accumulative, and inclusive, all which are considered fundamental features of the scientific endeavor" (p. 11). As a result, specific transparency practices have emerged, such as open access publishing, preregistration, sharing data, analyses, and code, performing study replications, and declaring positionality and conflicts of interest. Though it may come as a surprise to some, these open, transparent research practices have not been the norm in empirical and quantitative sciences, despite painstaking efforts being made in recent years (e.g., Berez-Kroeker et al., 2018; Berez-Kroeker, McDonnell, Koller, & Collister, 2022, among others).

To properly contextualize the need for open science, one must first consider the so-called reproducibility (replication) crisis. In the early 2010's, a team of researchers in Psychology embarked on a large-scale replication project to scrutinize what many considered to be the fields' major findings. Specifically, they attempted to replicate 100 influential studies (Open Science Collaboration, 2015). The endeavor produced astounding results—of note, that approximately 53% of the major findings did not replicate—and inspired similar large-scale replication projects in other fields, yielding similar results.[1] This series of events represents what is now referred to as the replication (or reproducibility) crisis (See also FORRT, 2021). Unsurprisingly, the results

[1] For Economics, see Camerer et al. (2016); for the Social Sciences, see Camerer et al. (2018); and, for cancer research, see Errington et al. (2021).

generated an uproar in the psychological sciences. The alarming findings garnered media attention (e.g., Oliver, 2016) and have led to periods of introspection and self-reflection in many adjacent fields, among them, linguistics (e.g., Berez-Kroeker et al., 2018; Bochynska et al., 2023).

Researchers have pointed to questionable research practices (QRPs), such as p-hacking–knowingly manipulating an analysis until a significant p-value is obtained (See Head, Holman, Lanfear, Kahn, & Jennions, 2015)–and HARKing–hypothesizing after the results are known (See Murphy & Aguinis, 2019)–, along with small sample sizes, poor theory, lack of transparency, misguided incentive structure in academia, etc., as factors that ultimately led to the replication crisis, though it is likely that many factors are/were simultaneously at play. For instance, the aforementioned QRPs may be an unfortunate consequence of misaligned incentive structures in academia, where publication is the universal currency. The pervasive pressure to publish likely leads many researchers to focus on quantity over quality. Couple this with the difficulty of publishing negative or null results, and the result is a research landscape in which many fields suffer from publication bias with little or no incentive to prioritize time consuming open science practices. Taking this into account, it is not hard to understand why some researchers may turn to QRPs. While it is difficult to quantify how prevalent QRPs are in a given field, in a survey of applied linguists, Isbell et al. (2022) found that 94% reported having engaged in one more, and 17% admitted to having committed some form of fraud.

In the aftermath of the aforementioned crisis, there has been a push for increased transparency and reproducible methodology to help mitigate the effects of QRPs. The clearest example of this is the Transparency and Openness Promotion Guidelines (TOP), author guidelines for journals that aim to help evaluate adherence to open science principles (See Nosek et al., 2015, as well as https://www.cos.io/initiatives/top-guidelines). The resulting methodological framework and associated techniques have reshaped research methods in Psychology, and, slowly but surely, are making their way into related fields. While many agree that open science practices represent a positive step forward in improving scientific rigor, these practices, by and large, have not been adopted in the field of linguistics (Bochynska et al., 2023). One reason for the slow adoption in linguistics may be related to the fact that engaging in open science is no trivial feat. On the contrary, it often requires learning new skills, thoughtful planning, as well as an openness and willingness to share materials, code, and data. Many researchers need to implement new techniques with limited pedagogical resources and embrace alternative methods of disseminating their research, all of which can constitute a steep learning curve. That being said, what engaging in open science ultimately entails is sure to be field-specific and vary accordingly. In some disciplines, for instance, it may only involve a few of the practices we outline in the present work without the need for innovative methodologies. Nonetheless, given how new open science practices are, it is reasonable to assume that current senior researchers were not trained in these innovative methodologies. As a consequence, many early career researchers (ECR) find themselves at a crossroads in which they are forced to learn open science on their own, often without institutional support. Ironically, there is also a growing expectation that ECRs implement these novel tools in order to be successful in their programs, on the job market, or to advance in their careers.

The present work intends to both highlight and contribute to a line of research focused on making open science practices understandable and accessible to researchers in linguistics from all backgrounds and at every stage, from students/ECRs to senior researchers and advisors. We identify the following three areas of stance, workflow and dissemination, in which linguists can engage in open science (see Figure 1). The first area, *stance*, refers to practices that focus on the researchers position or attitude towards openness and transparency. The second area, *workflow*, deals with methods and techniques researchers can implement to make their research projects more open and transparent. Finally, *dissemination* refers to novel ways in which researchers can help ensure that their research products are accessible and free from QRPs. While our coverage of these areas cannot be exhaustive, we highlight eight open science practices within these areas: positionality statements and declarations of conflict of interest, open data and materials[2], literate programming, reproducible code/projects, shareable computational environments, preregistration, registered reports, and pre-prints. We provide practical examples and detailed descriptions of the aforementioned practices with the goal of helping the interested linguist commence their journey of engaging in open science practices in their own research. Importantly, the present work should be considered a complement to the extant work fomenting open science practices in the speech sciences.

[2] The idea of open data and materials can be viewed as both a stance (i.e., one's willingness to make materials available), and part of the workflow/dissemination process (i.e., how one goes about making materials open and accessible). We touch on this idea in each of the following sections.

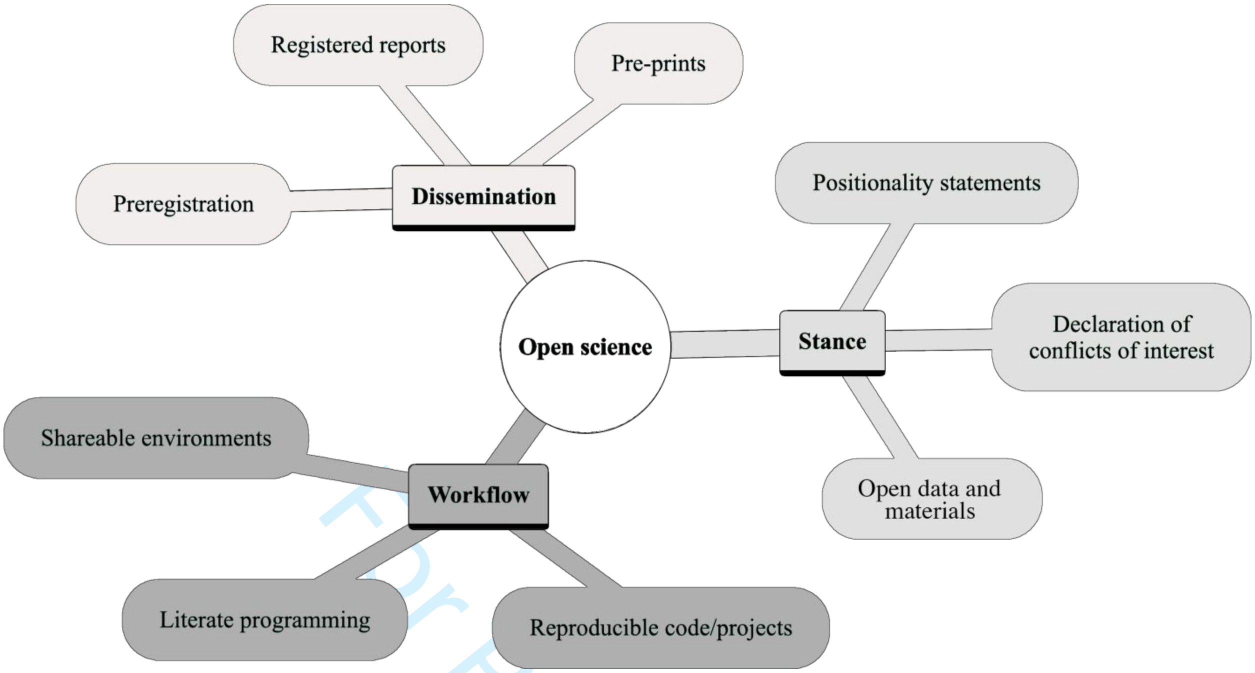

Figure 1: Some open science practices amenable to research in linguistics as they pertain to one's stance, workflow, and the dissemination of research products.

Stance

Positionality statements

A positionality statement is a reflective piece of writing that acknowledges a researcher's stance/position, toward a research topic, framework, and even participants. Similar to a statement of conflict of interest, a positionality statement can influence how results are interpreted (Rowe, 2014). One's positionality differs from a statement of conflict of interest in that it can also influence how research is undertaken and can encompass the researcher's social, cultural, and personal identity, as well as their biases and assumptions (Holmes, 2020). Among others, relevant personal characteristics that may be included in a positionality statement are gender and racial identity, age, sexual orientation, immigration status, and ideological stances (Berger, 2015). These traits may indirectly impact research endeavors, since participants may be more willing to engage in a study if they perceive the researcher as sympathetic (De Tona et al., 2006), or may even offer different responses based on the researcher's perceived identity (Berger, 2015). While positionality statements, due to their reflexive nature, may encompass larger pieces of writing, they can also take the form of short paragraphs that illustrate a few personal characteristics deemed relevant for the particular research endeavor. For instance, "Gabriela is a white immigrant cis-gender woman from Romania whose research focuses on how non-native speakers are ideologically framed as linguistically deficient in comparison to native speakers who are characterized by their linguistic authority and expertise". When submitting a study for publication, the positionality statement can be included in additional materials if the word limit is a concern.

Though positionality statements have been adopted in some disciplines of the humanities and social sciences as a means to recognizing the various ways in which researchers' backgrounds and identities may intersect with their research endeavors, they are a relatively new incorporation in the field of linguistics, appearing primarily in subfields, such as applied linguistics, linguistic anthropology, and linguistic ethnography (Bucholtz et al., 2023). Savolainen, Casey, McBrayer, & Schwerdtle (2023) draw connections between positionality statements and relatively more common statements of conflict of interest, arguing that, while researchers are required to disclose any and all financial gains associated with a research project, "positionality statements grant authors the freedom to decide which parts of their biography they choose to share and how they choose to frame it" (p. 1334). While statements of conflict of interest are notably underused in linguistic research (See Bochynska et al., 2023)[3], positionality statements are likely even less common. Nonetheless, they are considered by some to be

[3] Bochynska et al. (2023) surveyed open and transparent practices in linguistics and found that only 10% of the articles sampled included statements of conflict of interest, and, among those 10%, none declared any conflicts. For a clear example of what a declaration of conflicts of interest can entail in linguistics, the interested reader is directed to Bochynska et al. (2023). Of particular value are the *Competing interests* section and the coding form available at https://escholarship.org/uc/item/6m62j7p6#main and https://osf.io/ehyx3, respectively. Additionally, Cristea & Ioannidis (2018), Hardwicke et al. (2022), and Hardwicke et al. (2020) represent illustrative examples in psychiatry, psychology, and the social sciences more broadly.

increasingly crucial components of the research process, as they increase transparency into research practices (Steltenpohl, Hudson, & Klement, 2022) and contextualize the environment in which studies take place, or, in other words, they "[define] the boundaries within which research was produced" (Jafar, 2018, p. 1). Traditionally, positionality statements have been more prevalent in qualitative research. Our stance is that, when appropriate, they should be considered equally important in quantitative research as well. Aside from contributing to ongoing efforts to promote transparency and openness in research practices, recognizing and addressing one's positionality can, in some instances, support a study's quantitative validity by helping to reduce notions of bias (See Jafar, 2018, for a discussion in the field of medicine).

The support and advocacy for the inclusion of positionality statements in research publications is increasing (Bucholtz et al., 2023; Jafar, 2018; Steltenpohl et al., 2022). Bucholtz et al. (2023) note that considering a researcher's positionality may be especially important in linguistic research on certain language communities, such as indigenous communities, "[…] which relies on racially minoritized communities as sources of data yet lack adequate (if any) representation of those communities among faculty researchers" (p. 2). Nonetheless, others contest this practice. For example, some investigators point to the universalism of research, that is, the belief that scholarly endeavors should be assessed on their inherent merits, regardless of the status or personal identity of the person making the contribution (Savolainen et al., 2023). In addition and in opposition to Bucholtz et al. (2023), the self-identification associated with a positionality statement may also place some individuals, particularly women and individuals from marginalized groups, in a vulnerable position (Massoud, 2022). Specifically, in the field of law and society, Massoud (2022) posits that the pressure to state one's positionality can lead to increased anxiety, as well as cause readers to question the researcher's neutrality, and, ultimately, shift the focus away from the contributions of the research.

How can one marry the aforementioned benefits of including one's positionality with the legitimate counterpoints related to marginalized individuals? We believe researchers should only consider the option of including their positionality if they feel comfortable doing so. Roberts, Bareket-Shavit, Dollins, Goldie, & Mortenson (2020), for instance, argue against mandating one's positionality. Some journals have started to encourage authors to include positionality statements with their submissions (e.g., the Journal of Social and Personal Relationships) as a means to show their commitment to Diversity, Equity, Inclusivity, and Belonging (DEIB) initiatives. No journals, to the best of our knowledge, require positionality statements.

In sum, we believe positionality statements can be productive in linguistic research, as they promote critical self-reflection, increase transparency, can potentially help address diversity and inclusion concerns, and may increase the validity of findings in quantitative research. By reflecting on who it is that does the research, linguistics can become a more diverse, inclusive, and transparent field. That being said, it is important to consider the impact and potential burden of disclosing positionality on marginalized researchers, particularly in collaborative research settings. In the end, regarding one's positionality, *what* and *how* to share are fundamental considerations that cannot be overlooked by investigators, journals, publishing houses, and consumers of academic research. It is our stance that researchers should reflect on their positionality before starting a project, and, if and when it makes sense, consider including a positionality statement. For additional information and examples of positionality statements in linguistic research, the interested reader is directed to Bochynska et al. (2023), Weissler et al. (2023), and https://fosil-project.github.io/posts/positionality-statements/.

Open research data and materials

Recent efforts have pushed for researchers to make their materials (data, code, instruments, etc.) open to the public. Open data, specifically, refers to data collected for research that is freely and easily available to anybody interested in accessing it for any purpose (Open Knowledge, 2023). In academic research, statements such as "data available upon request" are commonplace (See Hardwicke & Ioannidis, 2018). In spite of such assurances, we now know they do not typically result in adequate sharing of research materials (Hardwicke & Ioannidis, 2018; Spellman, Gilbert, & Corker, 2017; Wicherts, Borsboom, Kats, & Molenaar, 2006). Researchers are increasingly encouraged to make linguistic data open and accessible via servers. An illustrative example is the IRIS database (https://www.iris-database.org), a language sciences digital repository that is freely accessible and permits the up- and downloading of research instruments and materials. Additional efforts include open science badges–visual symbols offered by some journals (e.g., *Language Learning*, *Language and Speech*) on published articles. These badges are awarded to researchers for adhering to certain open science principles, such as sharing code, data, or preregistering a study. In arguably more extreme cases, other journals have made data sharing a requirement for publication (e.g., *Applied Psycholinguistics*). Nonetheless, open sharing of research materials is still the exception rather than the norm in linguistics (Bochynska et al., 2023). In this section, we provide more detail regarding the benefits of 'openness' and consider the specific challenges researchers face in the field of linguistics. Our primary focus is on data, but we also underscore the importance of making all research materials open.

The underlying motivation for open data is relatively straightforward, particularly in the wake of the reproducibility crisis. Though researchers may understandably hesitate to share their data, we believe understanding the benefits of open data can help alleviate many concerns. Among researchers, there can exist anxieties unrelated to technical considerations about sharing data (Stieglitz et al., 2020). Stieglitz et al. (2020), in a study investigating 995 researchers from 13 universities in Germany across various fields, found that there were anxieties about competitive pressures, such as losing the opportunity to publish again from the same data set before another researcher does. In this case, anxieties can be quelled with the knowledge that the data can be made available after all research inquiries by the original researchers have been completed. Making linguistic data freely available improves credibility in our findings, to other researchers, and the general public, and may help develop more accurate generalizations and theories (See Berez-Kroeker et al., 2022). Prohibiting or impeding access to data collected for publicly funded research is, in many cases, unethical and can be a detriment to inclusivity. Open data is fundamental for cumulative science in numerous ways. It affords third parties the opportunity to scrutinize original findings, which promotes reproducibility and reduces errors, such as those related to statistical analyses and reporting of outcomes (e.g., Roettger, 2021b). Furthermore, it allows for published data to be reanalyzed in novel ways and utilized in meta-analyses. Revisiting old data sets using innovative techniques can support or contradict past narrative conclusions. For instance, using meta-analytic techniques, Casillas (2021) reexamined extant research regarding ‘compromise categories’ in early bilinguals. This line of research posits that bilingual individuals produce speech sounds intermediate to those produced by monolingual speakers of either language. By systematically reevaluating prior data and incorporating new acoustic analyses of coronal stops from early Spanish-English bilinguals, Casillas (2021) suggested that the cumulative evidence for ‘compromise’ stop categories was negligible. In lieu of intermediate phonetic categories, the study proposed early bilinguals can exhibit performance mismatches resulting from dynamic interlingual interactions. This reanalysis contradicted earlier assumptions about bilingual phonology and provided in-depth scrutiny of statistical power and evidence accumulation in bilingualism research. In short, open data is a cornerstone of scientific research in the 21st century that enables wider access to research information, which, in turn, facilitates validation, motivates replication, promotes reproducibility, and makes possible future scientific progress.

Open materials are particularly important for the field of linguistics, for all of the aforementioned reasons, and also because some linguists have described the state of the field, as far as English-language publications are concerned, as being Western, Educated, Industrialized, Rich, and Democratic (WEIRD, see Bochynska et al., 2023; Faytak et al., 2024; Nagle, Baese-Berk, Amengual, & Casillas, 2024). That is to say, the majority of linguistic research appears to be concentrated on specific languages, mainly Indo-Germanic, in overrepresented communities, by privileged scholars. Making materials in linguistic research accessible to all researchers can promote participation *in* and *with* underrepresented communities. Furthermore, it can increase the study of diverse and underreported languages by affording more researchers the opportunity to interact and learn from data that would otherwise not be available to them, which, in turn, can foster a more inclusive and comprehensive understanding of the global linguistic landscape.

Having stated all the above, it is necessary to recognize that linguistics faces a unique set of challenges with regard to data, as there are a multitude of subfields, each of which potentially works with a variety of data formats. Due to such diversity, one must determine which aspects of open science are relevant to their data. For example, a neurolinguistic study investigating event related potentials (ERPs) could share raw data for transparency, as well as preprocessed data with the code used to transform the raw data and a corresponding description for facilitation of reanalysis. In another field, the creation of a corpus will benefit from open access and the use of standardized file formats; the analysis of a corpus will benefit from sharing the search queries, the analysis code, and a description of the analysis code. At the heart of these challenges are ethical concerns that must be considered with care. First and foremost, the privacy and consent of participants must be safeguarded. Linguistic data often include personal information, which can be especially difficult to anonymize. While on the surface written and behavioral data may not appear to pose as many issues as audio and video recordings, which constitute a large portion of linguistic research materials, it is imperative that one consider the sources from which all types of data are derived. As expressed by Holton, Leonard, & Pulisifer (2022), if we haphazardly take language to represent trivial data points and lose focus on the individual embedded within a community, as well as the values of said community, we are doomed to "dehumanize and decontextualize" it (p. 50). This is particularly true when working with minority languages and/or marginalized communities. In cases such as these, the researcher must be held accountable, not only for the anonymization of participant information, but also for respecting and upholding the specific goals and restrictions put forth by the community. This includes, but is not limited to, the use, access, and storage of all collected data. In sum, careful consideration of the priorities of the researcher and the researched, which often do not align, is paramount (for more detailed views, see Adetula, Forscher, Basnight-Brown, Azouaghe, & IJzerman, 2022; Holton et al., 2022; Hudley, Mallinson, & Bucholtz, 2020; Leonard, 2021; Mufwene, 2020; Singh, Killen, & Smetana, 2023; Tsikewa, 2021, among others). In addition, generative artificial intelligence technologies, such as Large Language Models, are burgeoning. These technologies will certainly pose currently unknown challenges in the near future and may necessitate additional steps to secure the protection of sensitive data against misuse, particularly regarding adherence to the original agreement of informed consent, and, importantly, in upholding the conditions of use put forth by the stakeholders in marignalized communities.

While these challenges are substantial, we believe acceptable solutions exist in many, if not all, cases. When primary data, such as audio or video files, cannot be shared, derived data in the form of tabular files can take its place. For instance, if institutional policies prohibit the sharing of audio files, a comma-separated or tab-separated file (csv, tsv) containing the variables of interest (e.g., formant values, response times, etc.) can be made public instead. Tabular data files can be anonymized easily using arbitrary identification codes. Online data collection platforms, such as Prolific (https://www.prolific.com), typically remove identifying information by default and provide participant-specific identification numbers. In more uncommon cases in which institutional policies do not permit the sharing of derived data sets, synthetic data containing the same statistical properties can be generated and shared freely (See Quintana, 2020).[4] To quote the Directorate-General for Research & Innovation of the European Commission (2016), we believe the field can follow the principle that data should be "as open as possible, as closed as necessary" (p. 4).

Another substantial hurdle that cannot be overlooked revolves around the fact that researchers must learn to use new technologies to participate in open, transparent research. Making materials open *and* accessible is not as simple as merely uploading a data file. Ideally, researchers should include relevant information to contextualize the data set at the project-level (i.e., a project-summary document), the data-level (i.e., a README file explaining the data set),

[4] In short, the method consists of capturing the statistical properties of the original data set and using them to simulate new data that preserve the relationships between the variables of interest. A tutorial on the method described in Quintana (2020) is freely available on github (https://github.com/elifesciences-publications/synthpop-primer), and an online RStudio instance can be accessed at https://mybinder.org/v2/gh/dsquintana/synthpop-primer/master?urlpath=rstudio. All relevant materials are available on the OSF: https://osf.io/z524n/.

and the variable-level (i.e., a data dictionary) (C. Lewis, 2024). The inclusion of resources at these three levels is the optimal way for authors to provide the necessary context for an independent researcher to access and utilize their materials Unfortunately, most publicly available materials do not adhere to this standard. For this reason, we direct the interested reader to templates provided in C. Lewis (2024) for documentation at the project-level (https://osf.io/q6g8d, https://osf.io/d3pum), data-level (https://osf.io/tk4cb), and variable-level (https://osf.io/ynqcu). In addition, the reader is referred to the project, data, and variable level documentation of the present project, all of which are freely available on the OSF: https://osf.io/bsu2q/?view_only=68d1e41b327f4a28a9fcd0fc6537ecaf.

Once the research materials have been prepared for sharing, the researcher must decide where to share them. Platforms such as google drive, dropbox, etc. are not recommended because they are linked to personal accounts that may change or become unavailable over time. Free repositories designed for the purpose of sharing research materials, such as the Open Science Framework (Foster & Deardorff, 2017), GitHub, etc., are preferable and can be accessed simply by sharing a link. These repositories represent stable, long-term solutions with ample storage capacity. The materials can be downloaded directly, free of any kind of payment or exchange of personal information (such as an email address) by the user. For relevant examples, we direct the interested reader to https://osf.io/zx9ky/, https://osf.io/3bmcp/, or https://github.com/RAP-group/empathy_intonation_perc. Table 1 summarizes some of the options used by researchers and describes which features are available on each platform.

Table 1: Common data-sharing platforms and their respective features.

Platform	Long-term Support	Version Control	DOI Assignment	Anonymous Sharing	Key Features
Open Science Framework (OSF)	+	+	+	+	Project management and collaboration
GitHub	+	+	Integrates with Zenodo	+ (public repositories)	Project management and collaboration, ideal for coding
GitLab	+	+	Integrates with Zenodo	Limited	Project management and collaboration, ideal for coding
Bitbucket	+	+	Integrates with Zenodo	Limited	Project management and collaboration, ideal for coding
Zenodo	+	+ (via GitHub integration)	+	+	Supports range of file types
Figshare	+	Limited	+	+	Sharing datasets and figures
Box	−	−	−	Limited	Basic file storage and sharing
Google Drive	−	−	−	Limited	Basic file storage and sharing

To summarize, open materials are important because they facilitate transparency, rigor, reproducibility, replication, accumulation of knowledge, and, importantly, they make participating in the scientific endeavor more inclusive. According to some accounts, linguistics, in general, does not engage in open science practices, including sharing research materials (See Bochynska et al., 2023), though others characterize its participation in different terms. For instance, as stated in Berez-Kroeker et al. (2018) “Practitioners in different subfields ‘do transparency’ differently, and these practices could serve as models for an eventual amalgamated standard” (p. 9). While linguistics does face legitimate, field-specific challenges related to non-WEIRD communities, ultimately, the benefits of open materials outnumber many of these challenges. Researchers should take the stance to share what is reasonable and ethically responsible all the while holding at the forefront the priorities of the individuals from which the materials are derived, especially regarding data from marginalized communities.

Workflow

Having seen the consequences from the reproducibility crisis in other fields, reproducibility must be a crucial aspect of any scientific study. Researchers must be able to provide a clear and transparent account of their findings, including the methods used to obtain them. Reproducibility can help to ensure that research results are valid, reliable, and can be used by others to build on existing knowledge. In this section, we explore the importance of reproducibility, what we know about it in the field of linguistics, and how researchers can make their code and projects more reproducible.

In general, reproducibility helps to increase the credibility of research findings and allows other researchers to verify and build on existing work. A lack of reproducibility can lead to findings that cannot be replicated, resulting in wasted resources, and, conceivably, downstream impacts on public health and policy decisions that are often grounded in funded research. For these reasons, among others, transparency in research methods are essential to ensure reproducibility, which includes not only the data collection and analysis methods, but also the code used to conduct the analysis. In linguistics there is increasing awareness of the importance of reproducibility and how a lack thereof could potentially impede advancements in linguistic theory and theories of language acquisition, in addition to having implications for education and language policy decisions based on research findings. As a consequence, many investigators are showing heightened interest in safeguarding the reproducibility of their research.

Literate programming

For quantitative research, there are several steps that researchers can take to make their code and projects more reproducible. One approach is to create reports that document the research process by including descriptions of the data, the methods used to analyze the data, and the results. This documentation can then be made publicly available and used by third parties to retrace the steps to reproduce the research findings. While better than nothing at all, a more complete approach includes the analysis code in the same document in which the very manuscript is written. This integration of analysis code and prose into a single, dynamic document is known as literate programming (Knuth, 1984, 1992). Under the hood, a series of macros and functions are used to tangle the code and prose of the document into a separate file, usually a word document or a pdf, which can then be submitted for publication. Literate programming reduces the likelihood of copy and paste errors that often occur when passing the results of a statistical analysis from the analysis software to the word processing program. If the analysis changes in any way, e.g., more data is included, a different analytic strategy is applied, etc., the document is retangled to update the output file. Currently there are several implementations of literate programming for research purposes, the most common of which are RMarkdown files (.Rmd) Quarto markdown files (.qmd), and Jupyter notebooks (.ipynb). In the case of the former two, RMarkdown and Quarto, the R package `knitr` (Xie, 2015, 2023) tangles (also "knit" or "render") the output file. Jupyter notebooks require a front-end web page and a back-end kernel. The present manuscript was generated using literate programming via Quarto and is available for download on the Open Science Framework: https://osf.io/bsu2q/?view_only=68d1e41b327f4a28a9fcd0fc6537ecaf. Additionally, a brief tutorial in R is available at https://fosil-project.github.io/posts/literate-programming/.

Reproducible code/projects

While the implementation of literate programming into a research workflow is ideal, the gold standard is to use literate, dynamic documents in conjunction with reproducible projects. These projects include all of the data, code, and documentation necessary to reproduce the research findings, not only in a single report, but rather in many reports and/or presentations, simultaneously. This approach makes it easier for others to reproduce research findings and build on previous work because it obviates the complications involved with user-specific file paths and differing operating systems. Ideally, if the project works on one user's computer, it should work on any computer running the same software. In this sense, any researcher could theoretically download an entire project and reproduce the analyses and reports at the click of a button. A popular choice for reproducible projects is the open source software Posit (formerly RStudio), which utilizes `.Rproj` files called RStudio projects. Posit has recently released a new integrated development environment (IDE) called Positron that also works at the project-level, but has the added benefit of being relatively language-neutral. That is to say, one can use R, Python, Julia, Stan, and a number of other programming languages within a single IDE. More information and examples of completed reproducible projects are available to the interested reader here: https://osf.io/un45x/, https://osf.io/cp9bs/, and https://fosil-project.github.io/posts/reproducible-code-projects/. Additionally, the project files of the present work, including data, code, and markdown files, are publicly available on the OSF: https://osf.io/bsu2q/?view_only=68d1e41b327f4a28a9fcd0fc6537ecaf.

Shareable computational environments

Exciting, new technology that facilitates open science is coming out at a rapid pace. This is excellent news for anybody interested in learning the new tools, but also creates other issues, particularly with regard to outdated software. There is no way to completely future-proof code or projects. Researchers must continually strive to maintain the reproducibility of their work. This may imply updating code and documentation as needed, and, where feasible, testing projects on different operating systems to ensure that they can be run in different environments. Dependency management tools like renv (Ushey & Wickham, 2023) and targets (Landau, 2021) can be helpful in future-proofing projects and ensuring reproducibility. These tools help to manage the dependencies that are necessary to run code by providing specific versions of the software used originally by the researchers. Computational reproducibility platforms like Binder, Code Ocean, and Nix can also be used to create instances of virtual environments in which projects can be reproduced online. Thus, these platforms allow researchers to share their code and data in ways that can be easily reproduced by anybody with an internet connection. As an example, the present project is also available online in a stable Code Ocean container that captures the original computational environment: LINK TO BE PROVIDED AFTER PEER REVIEW. The interested reader is encouraged to re-run our code and re-render our files to further their understanding of how computational reproducibility platforms work in conjunction with literate programming.

Summarizing, reproducibility is a crucial aspect of scientific research. It helps to ensure that research findings are valid, reliable, and can be used by others to build on existing knowledge. In linguistics there is increasing awareness of the importance of reproducibility, and many researchers are taking steps to improve the transparency of their research. Instances of shareable computational environments can make research projects available to anybody with an internet connection, independent of operating systems and software preferences.

By creating dynamic reports using literate programming and integrating them into reproducible projects in conjunction with dependency management tools, linguists can make their projects more reproducible and accessible.

Dissemination

In this section, we will consider three open science innovations that are making a profound impact on how academic research is conducted, evaluated, and, ultimately, disseminated to the public. These innovations, preregistrations, registered reports, and pre-prints, were designed with the goal of reducing QRPs and publication bias.

Preregistration

A preregistration is a time-stamped document that provides comprehensive detail about a study, including, but not limited to, research questions, hypotheses, methodologies, and analytic strategies (Mellor & Nosek, 2018). Preregistrations are written prior to data collection and do not undergo peer review. The depth of content detail within a preregistration spans a spectrum: in the simplest case, a preregistration can comprise merely a hypothesis or perhaps a brief description of the methods; on the other extreme, a detailed preregistration can include code, power analyses, participant exclusion criteria and beyond. In this section, we provide information regarding the various components of a preregistration, centering on their advantageous impact on linguistic research. Specifically, we focus on *who* might want to consider preregistrations, *why* they might want to do so, *what* content they can include, and *how* they can complete a preregistration for a linguistics research project.

Linguistic research is multifaceted and spans diverse areas such as phonetics, phonology, syntax, morphology, sociolinguistics, natural language processing, and conversation/discourse analysis, to name just a few. These areas range from purely theoretical to quantitative and experimental, with many falling somewhere in between. Importantly, as highlighted by Roettger (2021a), researchers are human and humans have evolved to filter the world in irrational ways, which can lead to QRPs and other problems that may affect the replicability of published research. Preregistration emerged as a powerful instrument empowering linguists to bolster the trustworthiness and credibility of their inquiries by establishing a systematic and predefined methodology. We believe the practice of preregistration extends its benefits to researchers at all levels, including students and ECRs, senior academics, and professionals alike.

Researchers face vital decisions while engaging in research, with inherent flexibility involved in the process of designing and carrying out projects, as well as in the analysis of the data and interpretation of the results (Simmons, Nelson, & Simonsohn, 2011). This type of flexibility, termed "researcher degrees of freedom", can have serious down-stream consequences in quantitative research, particularly in linguistics. For instance, Coretta et al. (2023) provided the same speech-production data set to different research teams and asked them to answer the same research question. They found substantial variability in both the acoustic analyses and the analytic strategies, neither of which could be explained by analysts' prior beliefs, expertise, or the perceived quality of their analyses. Crucially, these decisions, both acoustic and analytic, impacted the teams' answers to the research question. To provide a simple example, a researcher studying lexical stress could concentrate on distinct acoustic cues typically associated with stress, i.e., pitch, duration, and intensity. Beyond selecting acoustic cues to measure, she must also select a domain for these measurements, such as the mid-point of stressed/unstressed syllables or an average value over the entirety of the syllable. Choices such as these, i.e., the researcher degrees of freedom, can wield significant influence on subsequent outcomes. Preregistration serves the purpose of meticulously documenting these choices *a priori*, thus acting as a deterrent against QRPs, like HARKing or p-hacking (Wicherts et al., 2016). This is because the researcher establishes what decisions will be made, such as measurement choices and analytic strategies, before data collection commences. A benefit of including a high level of specificity in the preregistration is that is forces researchers to consider facets of their study that might usually be deferred to a later stage, e.g., specific statistical tests. This proactive approach demands more initial time investment from the researcher, but also increases the likelihood of uncovering crucial flaws in the study design.

The scope of preregistration extends to any facet of research deemed worthy of temporal documentation preceding the initiation of the study. The essential components often include research questions/hypotheses, methodological framework, and analytic approaches. The specific elements that will comprise a preregistration can be considerably diverse, as they will depend on the specific domain within linguistics and the nuanced nature of the study in question. Consider, for instance, a psycholinguist conducting a self-paced reading study. In this context, the focus of the preregistration might include the formulation of hypotheses, as well as a complete description of the experimental paradigm. Additionally, the researcher may include a characterization of participant demographics, recruitment strategies, sample size considerations, independent variable manipulations, data transformations, and analytic strategies to test hypotheses. Importantly, not all of the aforementioned components are equally prioritized in all preregistrations. In sum, one can preregister any aspect of their research that they deem worthy of documenting *a priori*.

It is important to acknowledge that in many cases incorporating the entirety of these components into a preregistration represents a formidable challenge, as it front loads large portions of work that often take place after a study has begun, e.g., determining sample size, statistical models, etc. In such instances, researchers are encouraged to commence with elements they perceive as most valuable to their study. Many concerns about preregistration revolve around the potential burden of ‘extra work’. Conversely, preregistration is intended to streamline the workflow, fostering efficiency both in the short term and the long term, as it provides the researcher with complete control over the level of detail she chooses to include. The depth of preregistration directly correlates with the effort invested; the more comprehensive the preregistration, the greater the initial workload, leading to reduced effort in subsequent stages.

The Open Science Framework allows researchers to preregister a study.[5] Since its inception, the amount of preregistrations has grown each year (See Figure 2, left panel) and the cumulative number of registrations totals over 163,253 at the time of writing this text (See Figure 2, right panel). We provide useful guides and examples of preregistrations at the following links: https://osf.io/nprgz, https://osf.io/qvjzy, and https://fosil-project.github.io/posts/preregistration/.

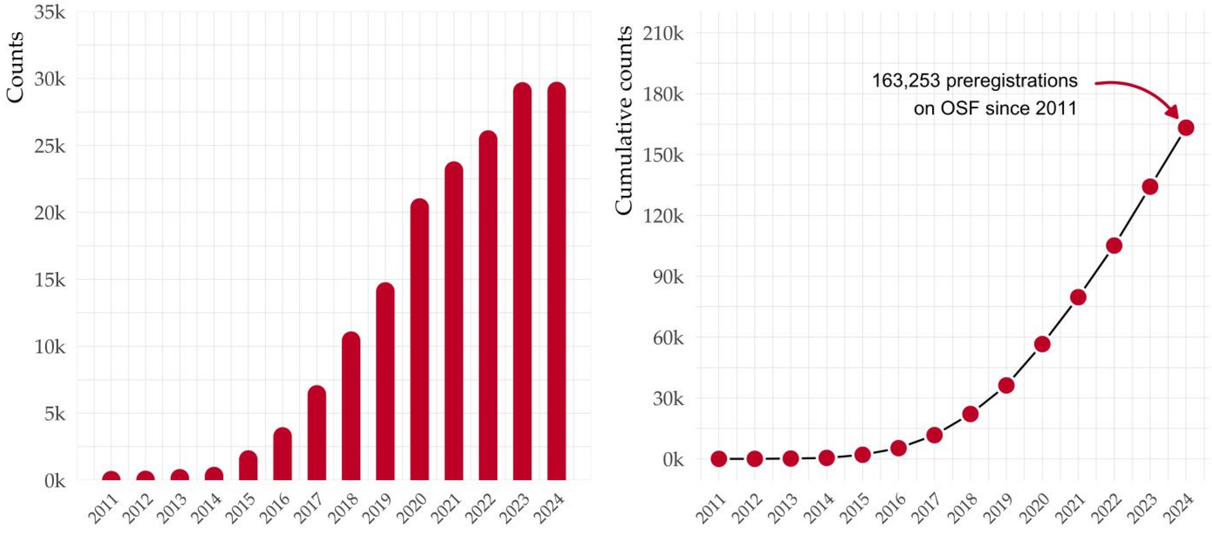

Figure 2: Preregistrations on the Open Science Framework. The left panel plots pregistrations as a function of year. The right panel plots cumulative preregistrations since 2011. Data scraped from https://osf.io/search on 12-12-2024.

Registered reports

The reproducibility crisis has drawn attention to the shortcomings of the traditional model of publishing scientific research. In the current model, researchers generate hypotheses, design studies, collect and analyze data, interpret results, and submit their findings for publication.

[5] See also https://aspredicted.org.

However, this model has been criticized for lending itself to QRPs, such as p-hacking and HARKing, which can result in publication bias.

To address these issues, researchers have attempted various reforms, such as meta-analysis and preregistration. Meta-analysis is a analytic technique that combines the results of multiple studies to increase the statistical power. Preregistration, as we have seen, involves publicly registering a study's design and methods before collecting data, to mitigate QRPs. Registered reports (RRs) represent a new publication model that conceptually combines preregistration with peer review (Nosek & Lakens, 2014). Preregistration is often confused with RRs, but they differ in that preregistration is a separate step that occurs before the traditional publishing pipeline, whereas RR is integrated into the publishing process. In this model, researchers submit a detailed proposal of their study, including their hypotheses, methods, and analyses, for review before data collection. If the proposal is accepted, the study is guaranteed publication, regardless of the results. This incentivizes rigorous methodology and reduces QRPs, as researchers cannot manipulate their analyses to obtain significant results. Figure 3 provides a side-by-side comparison of the standard publishing model and RRs.

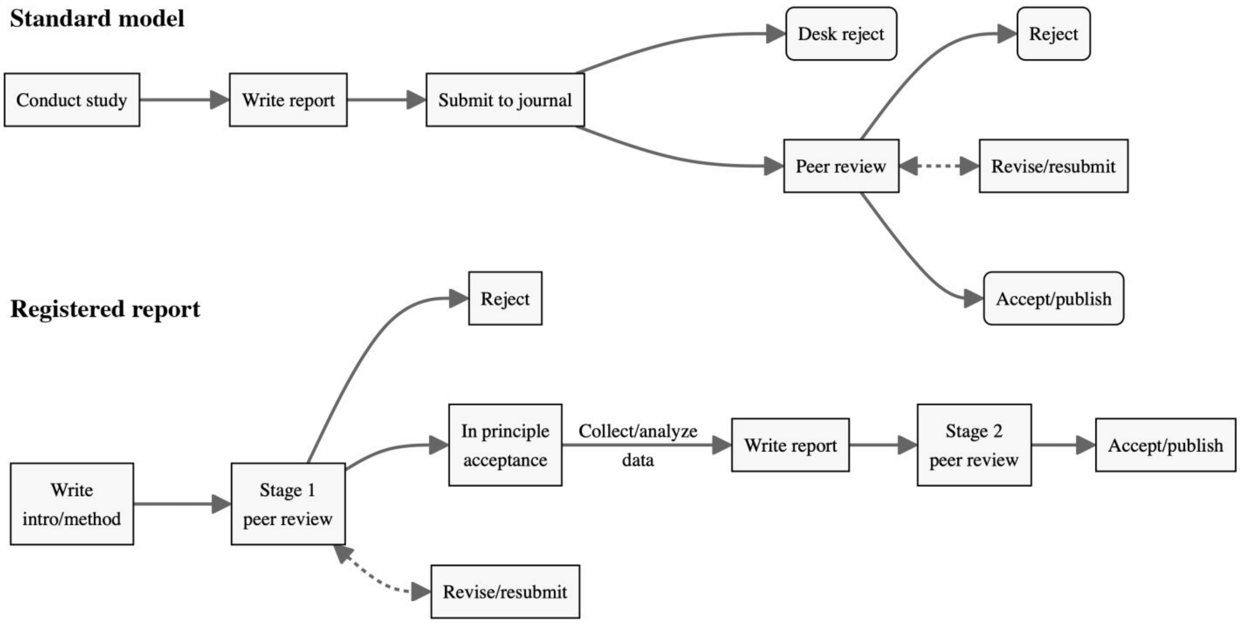

Figure 3: A comparison flow chart of the standard publication model and registered reports.

RRs were first introduced in 2013 by the Center for Open Science (COS), and have since been adopted by many journals across various fields, including psychology, neuroscience, and medicine. In 2019, there were approximately 156 journals offering RRs. This number has jumped to 318 at the time of writing this manuscript, an increase of 104%. Of those 318, only 14 are journals related to language or linguistics. Table 2 lists the language/linguistics journals along with information regarding relevant restrictions.

Table 2: Journals related to 'language' or 'linguistics' that include registered reports as a possible article submission type. The data were retrieved from the Open Science Framework on 12-14-2024. The complete list of journals is freely available at https://www.cos.io/initiatives/registered-reports. Empty cells indicate missing/unavailable data and TBA implies that a pending decision is 'to be announced'.

Journal	Permanence	Permits replication studies	Permits meta-analytic studies	Permits use of existing data	Requires data deposition
Bilingualism: Language and Cognition	Indefinite	✓			✓
Biolinguistics	Indefinite	✓			✓
Cognitive Linguistics	Indefinite	✓			✓
Glossa Psycholinguistics	Indefinite	TBA	TBA	TBA	TBA
Journal of Child Language	Indefinite	✓			
Journal of Memory and Language	Special issue	✓			
Journal of Speech, Language, and Hearing Research					
Language and Cognition					
Language and Speech	Indefinite	✓		✓	✓
Language Learning	Indefinite	✓	✓	✓	✓
Linguistics	Indefinite	✓			✓
Neurobiology of Language	Indefinite	TBA	TBA	TBA	TBA
Second Language Research					
Journal of Memory and Language	Special issue	✓			

The majority of the listed journals plan to offer RRs as a possible submission option indefinitely (n = 9). Likewise, nine of these 14 journals permit RRs as an option for replication studies. Only one journal (Language Learning) specifically states that it will consider RRs that plan to conduct meta-analyses, and two of the journals consider RRs as an option for studies that propose analyzing data sets that already exist. Finally, for six of the 14 journals, a public data deposition is a requirement for RRs.[6]

[6] A public data deposition implies that the authors must publish 'all relevant data collected as part of the research within a freely accessible repository' (See Center for Open Science, 2024).

RRs cannot solve all the problems with the current model, but they can help reduce QRPs and increase transparency in scientific research. RRs are gaining popularity, but some fields, such as linguistics, have been slow to adopt them. RRs may particularly benefit ECRs, who can use them to increase their chances of publication and build a reputation for rigor. However, more senior researchers may be resistant to change and may need to be convinced of the benefits of RRs for the field as a whole. In sum, registered reports represent a promising new model for publishing scientific research that can help reduce QRPs and increase transparency. As more journals adopt RRs, the scientific community can move towards a more rigorous and trustworthy publishing model.

Pre-prints

A pre-print is a version of a research article, open and accessible, that has not yet undergone peer review but is publicly available online, through a pre-print server. The general process consists of an initial screening process, followed by a posting of the manuscript on the preprint server within a few days of submission, bypassing peer review, and making the research findings freely accessible online (Puebla, Polka, & Rieger, 2021). Pre-prints allow researchers to share their findings with the scientific community and get feedback before their work is published in a traditional academic journal. This process can speed up the dissemination of knowledge and facilitate collaboration between researchers.

One of the primary benefits of pre-prints is that they allow researchers to share their findings quickly and easily. This can be especially important in fields where research moves quickly, such as biology or computer science. Pre-prints also allow researchers to receive feedback on their work from their peers, which can help to improve the quality of their research. The provision of commentary and reviews of pre-prints yields benefits, not only to the authors, but also to reviewers, journals, publishers, and the readership. This inclusive process allows more researchers and reviewers to participate in discussing the research findings and can reduce the need for repeated rounds of re-review or extensive revisions. Recognizing these benefits, more major publishers have either launched pre-print platforms or entered partnerships over the past 5-7 years, allowing pre-prints to be incorporated into the workflow (Puebla et al., 2021). By making research findings available to the public before peer-review, pre-prints not only improve the accuracy and reliability of research findings, but also encourage collaborative efforts to identify potential errors, refine methodologies, and accelerate knowledge dissemination.

Another benefit of pre-prints is that they can help to reduce publication bias, a widespread challenge in traditional publishing. Publication bias occurs when positive results are more likely to be published than negative results (Matosin, Frank, Engel, Lum, & Newell, 2014). This can skew the scientific literature and lead to a misunderstanding of the state of the research. Pre-prints address this obstacle by openly sharing all research findings, regardless of outcome, creating a fairer and more accurate representation of the current scientific landscape of that field.

Pre-prints have become increasingly popular in recent years, particularly in fields such as biology, physics, and computer science. The adoption of pre-prints has been slower in some fields, such as the social sciences and humanities, but this is changing as more researchers become aware of the benefits of open science, and new national and regional platforms by open science advocates continue to emerge (Gawne & Styles, 2022). Figure 4 illustrates the growth of pre-prints on the Open Science Framework since 2016.

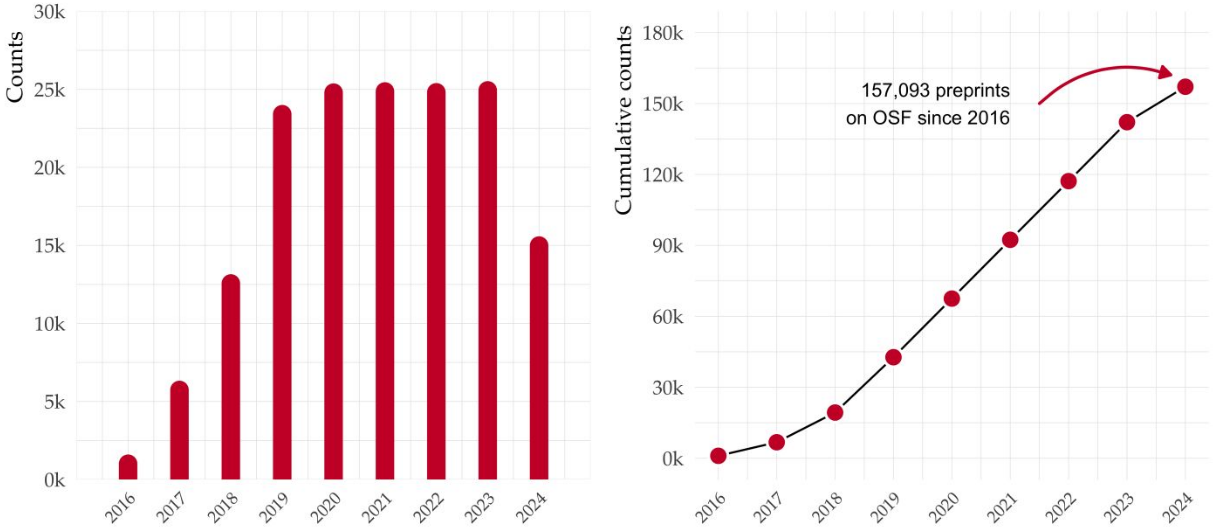

Figure 4: Preprints on the Open Science Framework. The left panel plots preprints as a function of year. The right panel plots cumulative preprints since 2016. Data scraped from https://osf.io/search on 12-15-2024.

Despite the clear benefits, some researchers remain hesitant to use pre-prints. One concern is that publishing a pre-print may harm their chances of being published in a traditional academic journal. However, this concern is becoming less relevant as more journals are accepting pre-prints as a legitimate form of publication. According to Liu & De Cat (2021), who conducted a survey asking as to the barriers in sharing preprints and discovered that the following were raised as additional barriers: peer review, journal policy, lack of knowledge of the process, confidentiality issues, data types, utility of sharing preprints, time constraints, and issues in pre-print management. Fortunately, researchers interested in making a pre-print publicly available will find the process to be quite simple. One must first select a pre-print server that aligns with the course of research (See Table 3). Next, the pre-print is likely to undergo a short screening, confirming author background, basic research content, and compliance with the ethical standards of the pre-print platform. Once the pre-print passes the screening process, the content is made available online in open access format.

Table 3: Available pre-print servers related to language and/or linguistics.

Server	Discipline(s)	Year Created	URL
LingBuzz	General Linguistics	2006	https://ling.auf.net/lingbuzz
Open Science Framework	Multidisciplinary (includes Linguistics)	2011	https://osf.io/preprints
PsyArXiv	Psychology, Cognitive Sciences, Psycholinguistics, Linguistics	2016	https://osf.io/preprints/psyarxiv
Cogprints	Multidisciplinary (includes Cognitive Sciences and Linguistics)	1995	https://web-archive.southampton.ac.uk/cogprints.org/
SocArXiv	Social Sciences (includes Sociolinguistics)	2016	https://osf.io/preprints/socarxiv
EdArXiv	Education Research (includes Applied Linguistics	2018	https://osf.io/preprints/edarxiv
Computational Linguistics Open Archive (CLARIN)	Language-Based Research	2012	https://www.clarin.eu/
ACL Anthology	Computational Linguistics and NLP	2004	https://aclanthology.org/
arXiv	Multidisciplinary (includes Computational Linguistics, NLP)	1991	https://arxiv.org/
SciELO Preprints	Research pertinent to Latin America, Spain, Portugal and South Africa	1998	https://preprints.scielo.org/index.php/scielo/preprints
HAL (Hyper Articles en Ligne)	Multidisciplinary (includes Language-Specific French Linguistics)	2001	https://hal.science/

The growing visibility of pre-prints, and their acceptance as valid research outputs by diverse stakeholders, including researchers, funders, and national institutions, has fueled collaborative research efforts and strengthened support for their presence in a variety of research disciplines. Pre-prints play an important role in advancing the tenets of open science by promoting transparency, reproducibility, and collaboration. While some researchers may still be hesitant to use this dissemination paradigm, the benefits of open science are becoming increasingly clear. By embracing pre-prints, linguists can accelerate the dissemination of knowledge, improve the quality of research, and ensure that their findings are available to the widest possible audience.

Concluding remarks

The early 2010's saw the reproducibility crisis take hold of the psychological sciences. As a consequence, there has been a push for increased transparency and reproducible methodology to help mitigate the effects of questionable research practices. The resulting methodological framework and associated techniques, now referred to as open science, have reshaped research methods in psychology and have slowly but surely made their way into adjacent fields, such as linguistics. While open science provides novel techniques and integrates state-of-the-art innovations, it also comes with challenges, particularly with regard to the steep learning curve researchers face when learning these new methods. We advocate for the "buffet" approach, in which select open science practices are integrated into the researcher's workflow slowly over time (e.g., Bergmann, 2018). We have provided descriptions and relevant examples of these practices to accompany the many guides already available for learning open science (e.g., Crüwell et al., 2018; N. A. Lewis, 2020, https://FOSIL-project.github.io, https://book.fosteropenscience.eu/, among many others).

Crucially, the purpose of this article is to help foster open science in linguistics. Important considerations often overlooked in the wake of the open science movement deal with (1) how linguists actually learn open science practices and (2) how senior researchers can train the next generation of linguists. Few, if any, researchers have had explicit instruction on the practices of open science as part of their professional training. Nonetheless, today's speech researcher is expected to be up to date on the current protocols of open science in order incorporate the methodological practices aimed at improving reproducibility/replicability. What does it mean for the field? We believe that researchers—linguists specifically—have to adapt and learn the new methods of open science. Additionally, we must, as a field, concentrate our efforts to train current students/ECRs in open, transparent research practices, and linguistic journals must to adapt to new models of publishing. In the present work we have outlined eight specific open science practices, classified into three areas–stance, workflow, and dissemination–, that researchers in linguistics can adopt to make their research more open, transparent, inclusive, and accessible to a wider audience.

References

Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from–not just to–Africa. *Nature Reviews Psychology*, *1*(7), 370–371. https://doi.org/10.1038/s44159-022-00070-y

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., … Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, *56*(1), 1–18. https://doi.org/10.1515/ling-2017-0032

Berez-Kroeker, A. L., McDonnell, B., Koller, E., & Collister, L. B. (Eds.). (2022). *The open handbook of linguistic data management*. The MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001

Berger, R. (2015). Now I see it, now I don't: Researcher's position and reflexivity in qualitative research. *Qualitative Research*, *15*(2), 219–234. https://doi.org/10.1177/14687941124684

Bergmann, C. (2018). How to integrate open science into language acquisition research? *The 43rd Annual Boston University Conference on Language Development (BUCLD 43), Boston, USA*.

Bochynska, A., Keeble, L., Halfacre, C., Casillas, J. V., Champagne, I.-A., Chen, K., … Roettger, T. B. (2023). Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics*, *2*(1, 18), 1–36. https://doi.org/10.5070/G6011239

Bucholtz, M., Campbell, E. W., Cevallos, T., Cruz, V., Fawcett, A. Z., Guerrero, B., … Reyes Basurto, G. (2023). Researcher positionality in linguistics: Lessons from undergraduate experiences in community-centered collaborative research. *Language and Linguistics Compass*, *17*(4), 1–15. https://doi.org/10.1111/lnc3.12495

Camerer, C. F., Dreber, A., Forsell, E., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf0918

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., et al.others. (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

Casillas, J. V. (2021). Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages*, *6*(1), 9. https://doi.org/10.3390/languages6010009

Center for Open Science. (2024). *Registered reports*. Accessed from the Center for Open Science on 2024-12-12. Retrieved from https://www.cos.io/initiatives/registered-reports

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., … Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–29. https://doi.org/10.1177/25152459231162567

Cristea, I.-A., & Ioannidis, J. P. A. (2018). Improving disclosure of financial conflicts of interest for research on psychosocial interventions. *JAMA Psychiatry*, *75*(6), 541–542. https://doi.org/10.1001/jamapsychiatry.2018.0382

Crüwell, S., Doorn, J. van, Etz, A., Makel, M. C., Moshontz, H., Niebaum, J., … Schulte-Mecklenbeck, M. (2018). *7 easy steps to open science: An annotated reading list*. https://doi.org/10.1027/2151-2604/a000387

De Tona, C. et al. (2006). But what is interesting is the story of why and how migration happened. *Forum: Qualitative Social Research*, *7*(3), 1–12. https://doi.org/10.17169/fqs-7.3.143

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021). Investigating the replicability of preclinical cancer biology. *Elife*, *10*, e71601. https://doi.org/10.7554/eLife.71601

European Commission, Directorate-General for Research & Innovation. (2016). *H2020 programme: Guidelines on FAIR data management in horizon 2020, version 3.0*. Luxembourg, European Commission, Directorate-General for Research & Innovation. https://doi.org/10.25607/OBP-774

Faytak, M., Kadavá, Š., Xu, C., Özsoy, O., Akumbu, PiusW., Cardoso, A., … Roettger, T. B. (2024). *Big team science for language science: Opportunities and challenges*. Open Science Framework. Retrieved from osf.io/3pkj6

FORRT. (2021). Reproducibility crisis (a.k.a. Replicability or replication crisis). Retrieved from https://forrt.org/glossary/english/reproducibility_crisis/

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). *Journal of the Medical Library Association: JMLA*, *105*(2), 203.

Gawne, L., & Styles, S. (2022). Situating linguistics in the social science data movement. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 9–25). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0006

Hardwicke, T. E., & Ioannidis, J. P. (2018). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PloS One*, *13*(8), 1–12. https://doi.org/10.1371/journal.pone.0201856

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

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), 1–10. https://doi.org/10.1098/rsos.190806

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), 1–15. https://doi.org/10.1371/journal.pbio.1002106

Holmes, A. G. D. (2020). Researcher positionality–a consideration of its influence and place in qualitative research–a new researcher guide. *Shanlax International Journal of Education*, *8*(4), 1–10. https://doi.org/10.34293/education.v8i4.3232

Holton, G., Leonard, W. Y., & Pulisifer, P. L. (2022). Indigenous peoples, ethics, and linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 51–60). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0008

Hudley, A. H. C., Mallinson, C., & Bucholtz, M. (2020). Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, *96*(4), e200–e235. https://doi.org/10.1353/lan.2020.0074

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., … Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*(1), 172–195. https://doi.org/10.1111/modl.12760

Jafar, A. J. N. (2018). What is positionality and should it be expressed in quantitative studies? *Emergency Medicine Journal*, *35*(5), 323–324. https://doi.org/10.1136/emermed-2017-207158

Knuth, D. E. (1984). Literate programming. *The Computer Journal*, *27*(2), 97–111. https://doi.org/10.1093/comjnl/27.2.97

Knuth, D. E. (1992). *Literate programming*. Center for the Study of Language; Information, Stanford University, CA: Distributed by Univiversity of Chicago Press.

Landau, W. M. (2021). The targets R package: A dynamic make-like function-oriented pipeline toolkit for reproducibility and high-performance computing. *Journal of Open Source Software*, *6*(57), 2959. Retrieved from https://doi.org/10.21105/joss.02959

Leonard, W. Y. (2021). Centering indigenous ways of knowing in collaborative language work. In L. Crowshow, I. Genee, M. Peddle, J. Smith, & C. Snoek (Eds.), *Sustaining indigenous languages: Connecting communities, teachers, and scholars* (pp. 21–34). Athabasca University Press.

Lewis, C. (2024). *Data management in large-scale education research*. CRC Press.

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

Liu, M., & De Cat, C. (2021). Open science in applied linguistics: A preliminary survey. In L. Plonsky (Ed.), *Open science in applied linguistics* (pp. 1–28). John Benjamins.

Massoud, M. F. (2022). The price of positionality: Assessing the benefits and burdens of self-identification in research methods. *Journal of Law and Society*, *49*, S64–S86. https://doi.org/10.1111/jols.12372

Matosin, N., Frank, E., Engel, M., Lum, J. S., & Newell, K. A. (2014). Negativity towards negative results: A discussion of the disconnect between scientific worth and scientific culture. *Dis Model Mech*, *7*(2), 171–173. https://doi.org/10.1242/dmm.015123

Mellor, D. T., & Nosek, B. A. (2018). Easy preregistration will benefit any research. *Nature Human Behaviour*, *2*(28), 98.

Mufwene, S. S. (2020). Decolonial linguistics as paradigm shift. In A. Deumert, A. Storch, & N. Shephard (Eds.), *Colonial and decolonial linguistics: Knowledges and epistemes* (pp. 289–300). Oxford University Press.

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, *34*, 1–17. https://doi.org/10.1007/s10869-017-9524-7

Nagle, C., Baese-Berk, M., Amengual, M., & Casillas, J. V. (2024). *Sound communities: A quantitative proposal for studying bilingualism in context*. PsyArXiv. https://doi.org/10.31234/osf.io/m67tx

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al.others. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab237

Nosek, B. A., & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology*, *45*(3), 137–141. https://doi.org/10.1027/1864-9335/a000192

Oliver, J. (2016). Scientific studies: Last Week Tonight with John Oliver. Retrieved from https://youtu.be/0Rnq1NpHdmw?si=6tIMWkEbOY47rhaE

Open Knowledge. (2023). The Open Definition. Retrieved from https://opendefinition.org

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Parsons, S., Azevedo, F., Elsherif, M. M., Guay, S., Shahim, O. N., Govaart, G. H., et al.others. (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

Puebla, I., Polka, J., & Rieger, O. Y. (2021). *Preprints: Their evolving role in science communication*. MetaArXiv. https://doi.org/10.31222/osf.io/ezfsk

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, *9*, 1–12. https://doi.org/10.7554/eLife.53275

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927

Roettger, T. B. (2021a). Preregistration in experimental linguistics: Applications, challenges, and limitations. *Linguistics*, *59*(5), 1227–1249. https://doi.org/10.1515/ling-2019-0048

Roettger, T. B. (2021b). Toward transparent and reproducible speech sciences. *Séminaires de Recherches En Phonétique Et Phonologie, CNRS, Paris*.

Rowe, W. E. (2014). Positionality. In D. Coghlan & M. Brydon-Miller (Eds.), *The SAGE encyclopedia of action research* (pp. 627–628). Sage.

Savolainen, J., Casey, P. J., McBrayer, J. P., & Schwerdtle, P. N. (2023). Positionality and its problems: Questioning the value of reflexivity statements in research. *Perspectives on Psychological Science*, *18*, 1331–1338. https://doi.org/10.1177/17456916221144988

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

Singh, L., Killen, M., & Smetana, J. G. (2023). Global science requires greater equity, diversity, and cultural precision. *APS Observer*, *36*. Retrieved from https://www.psychologicalscience.org/observer/gs-equity-diversity-cultural-precision

Spellman, B., Gilbert, E., & Corker, K. S. (2017). *Open science: What, why, and how*. https://doi.org/10.31234/osf.io/ak6jr

Steltenpohl, C., Hudson, S., & Klement, K. (2022). How to begin writing a positionality statement. Retrieved from https://vimeo.com/675236573/741e24aab7

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger, B., L'opez, A., & Rehwald, S. (2020). When are researchers willing to share their data?–Impacts of values and uncertainty on open data in academia. *PLoS One*, *15*(7), 1–20. https://doi.org/10.1371/journal.pone.0234172

Tsikewa, A. (2021). Reimagining the current praxis of field linguistics training: Decolonial considerations. *Language*, *97*(4), e293–e319. https://doi.org/10.1353/lan.2021.0072

Ushey, K., & Wickham, H. (2023). *Renv: Project Environments*. Retrieved from https://CRAN.R-project.org/package=renv

Weissler, R., Drake, S., Kampf, K., Diantoro, C., Foster, K., Kirkpatrick, A., … Baese-Berk, M. M. (2023). Speech perception and production lab: Positionality statements. Retrieved from https://www.speechperceptionproductionlab.com/positionalitystatments

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *American Psychologist*, *61*(7), 726. https://doi.org/10.1037/0003-066X.61.7.726

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Aaert, R. C. van, & Assen, M. A. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Boca Raton, Florida: Chapman; Hall/CRC. Retrieved from https://yihui.org/knitr/

Xie, Y. (2023). *Knitr: A general-purpose package for dynamic report generation in r*. Retrieved from https://yihui.org/knitr/

Supplementary materials

Author contributions

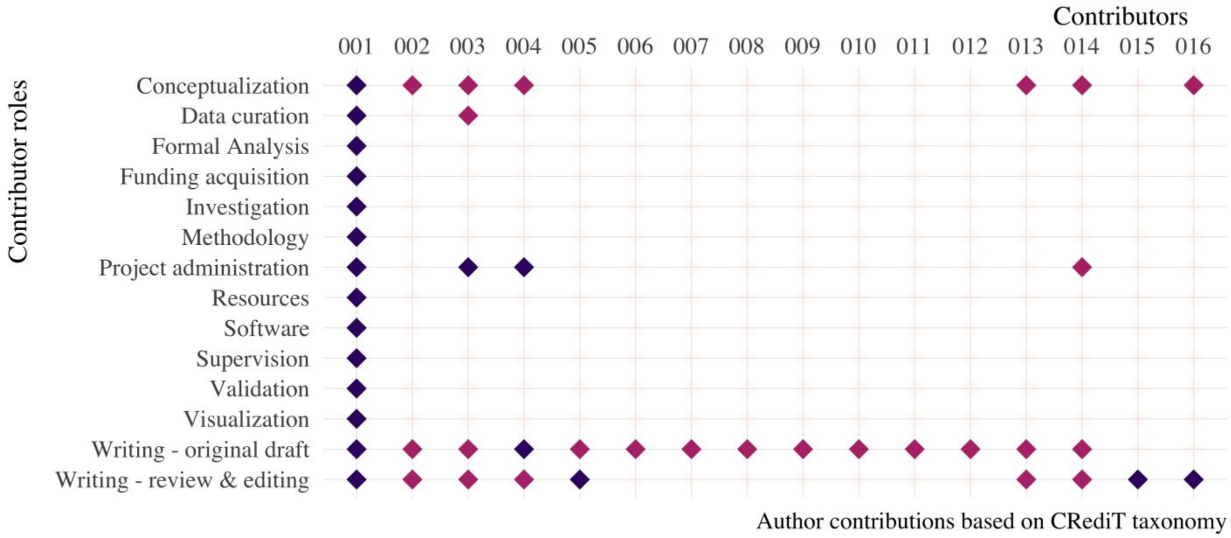

Figure 5: Author contributions according to the CRediT author roles taxonomy. Contributions are indicated as being substantial (dark diamonds) or moderate (light diamonds).

The authors made the following contributions: 001: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing - original draft, Writing - review & editing; 002: Conceptualization, Writing - original draft, Writing - review & editing; 003: Conceptualization, Data curation, Project administration, Writing - original draft, Writing - review & editing; 004: Conceptualization, Project administration, Writing - original draft, Writing - review & editing; 005: Writing - original draft, Writing - review & editing; 006: Writing - original draft; 007: Writing - original draft; 008: Writing - original draft; 009: Writing - original draft; 010: Writing - original draft; 011: Writing - original draft; 012: Writing - original draft; 013: Conceptualization, Writing - original draft, Writing - review & editing; 014: Conceptualization, Project administration, Writing - original draft, Writing - review & editing; 015: Writing - review & editing; 016: Conceptualization, Writing - review & editing.

Reproducibility information

About this document

This document was written in quarto (version 1.6): https://quarto.org.

Session info

```
 setting  value
 version  R version 4.4.1 (2024-06-14)
 os       macOS 15.2
 system   aarch64, darwin20
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/New_York
 date     2024-12-17
 pandoc   3.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/aarch64/ (via
rmarkdown)

              loadedversion       date
archive               1.1.9 2024-09-12
bit                   4.5.0 2024-09-20
bit64                 4.5.2 2024-09-22
cachem                1.1.0 2024-05-16
cellranger            1.1.0 2016-07-27
chromote              0.3.1 2024-12-15
cli                   3.6.3 2024-06-21
colorspace            2.1-1 2024-07-26
contributoR           0.4.0 2024-12-17
crayon                1.5.3 2024-06-20
devtools              2.4.5 2022-10-11
digest               0.6.37 2024-08-19
dplyr                 1.1.4 2023-11-17
ellipsis              0.3.2 2021-04-29
evaluate              1.0.1 2024-10-10
fansi                 1.0.6 2023-12-08
farver                2.1.2 2024-05-13
```

```
fastmap          1.2.0 2024-05-15
forcats          1.0.0 2023-01-29
fs               1.6.4 2024-04-25
gargle           1.5.2 2023-07-20
generics         0.1.3 2022-07-05
ggplot2          3.5.1 2024-04-23
glue             1.8.0 2024-09-30
googledrive      2.1.1 2023-06-11
googlesheets4    1.1.1 2023-06-11
gtable           0.3.5 2024-04-22
here             1.0.1 2020-12-13
hms              1.1.3 2023-03-21
htmltools      0.5.8.1 2024-04-04
htmlwidgets      1.6.4 2023-12-06
httpuv          1.6.15 2024-03-26
httr             1.4.7 2023-08-15
janitor          2.2.0 2023-02-02
jsonlite         1.8.9 2024-09-20
knitr             1.48 2024-07-07
labeling         0.4.3 2023-08-29
later            1.3.2 2023-12-06
lifecycle        1.0.4 2023-11-07
lubridate        1.9.3 2023-09-27
magrittr         2.0.3 2022-03-30
memoise          2.0.1 2021-11-26
mime              0.12 2021-09-28
miniUI         0.1.1.1 2018-05-18
munsell          0.5.1 2024-04-01
patchwork        1.3.0 2024-09-16
pillar           1.9.0 2023-03-22
pkgbuild         1.4.4 2024-03-17
pkgconfig        2.0.3 2019-09-22
pkgload          1.4.0 2024-06-28
processx         3.8.4 2024-03-16
profvis          0.4.0 2024-09-20
promises         1.3.0 2024-04-05
ps               1.8.0 2024-09-12
```

```
purrr           1.0.2 2023-08-10
R6              2.5.1 2021-08-19
Rcpp            1.0.13 2024-07-17
readr           2.1.5 2024-01-10
remotes         2.5.0 2024-03-17
rlang           1.1.4 2024-06-04
rmarkdown       2.28 2024-08-17
rprojroot       2.0.4 2023-11-05
rstudioapi      0.16.0 2024-03-24
rvest           1.0.4 2024-02-12
scales          1.3.0 2023-11-28
sessioninfo     1.2.2 2021-12-06
shiny           1.9.1 2024-08-01
snakecase       0.11.1 2023-08-27
stringi         1.8.4 2024-05-06
stringr         1.5.1 2023-11-14
tibble          3.2.1 2023-03-20
tidyr           1.3.1 2024-01-24
tidyselect      1.2.1 2024-03-11
timechange      0.3.0 2024-01-18
tzdb            0.4.0 2023-05-12
urlchecker      1.0.1 2021-11-30
usethis         3.0.0 2024-07-29
utf8            1.2.4 2023-10-22
vctrs           0.6.5 2023-12-01
viridisLite     0.4.2 2023-05-02
vroom           1.6.5 2023-12-05
websocket       1.4.2 2024-07-22
withr           3.0.1 2024-07-31
xfun            0.48 2024-10-03
xml2            1.3.6 2023-12-04
xtable          1.8-4 2019-04-21
yaml            2.3.10 2024-07-26
```

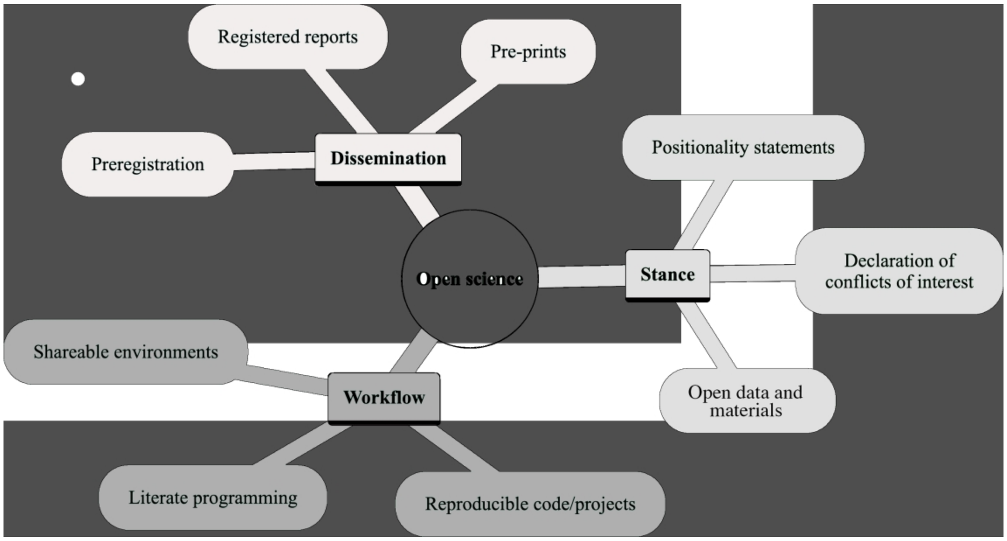

Figure 1: Some open science practices amenable to research in linguistics as they pertain to one's stance, workflow, and the dissemination of research products.

1346x716mm (57 x 57 DPI)

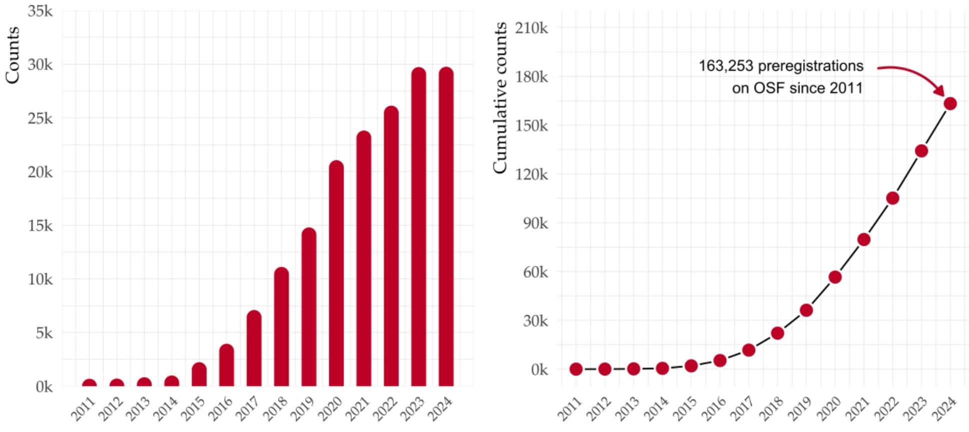

Figure 2: Preregistrations on the Open Science Framework. The left panel plots pregistrations as a function of year. The right panel plots cumulative preregistrations since 2011. Data scraped from https://osf.io/search on 12-12-2024.

1905x857mm (72 x 72 DPI)

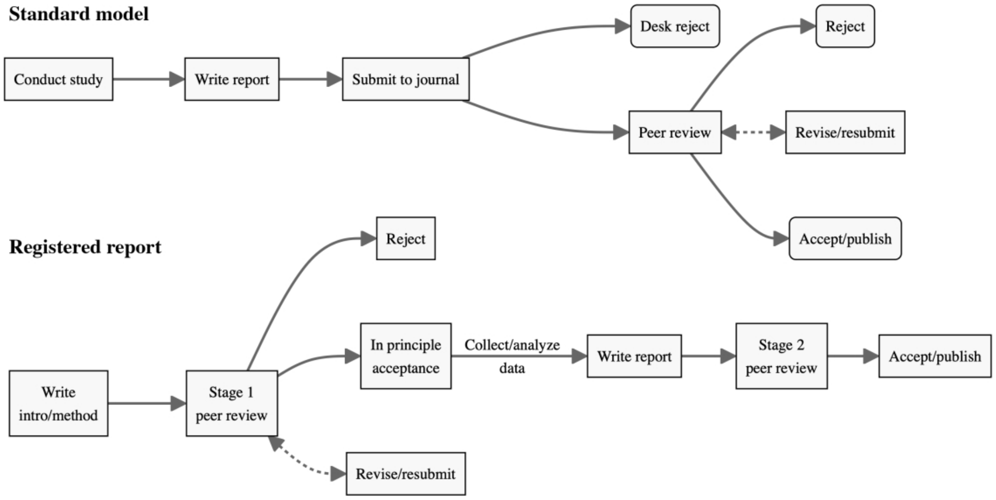

Figure 3: A comparison flow chart of the standard publication model and registered reports.

1310x663mm (57 x 57 DPI)

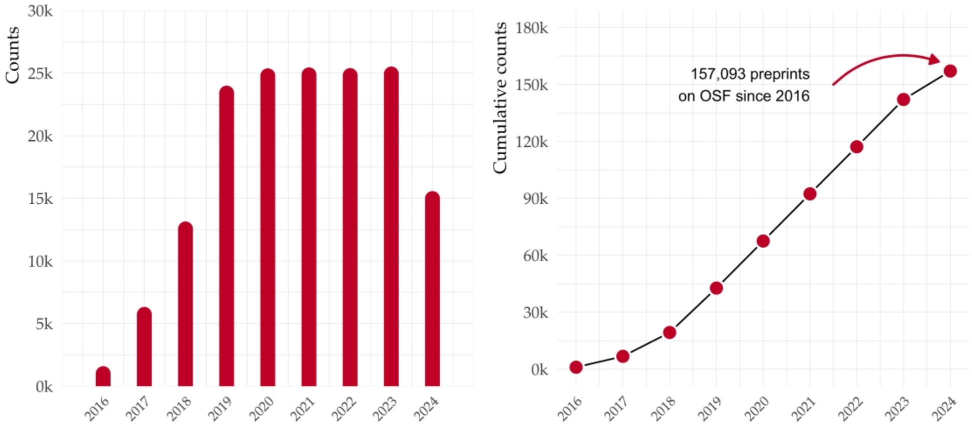

Figure 4: Preprints on the Open Science Framework. The left panel plots preprints as a function of year. The right panel plots cumulative preprints since 2016. Data scraped from https://osf.io/search on 12-15-2024

1905x857mm (72 x 72 DPI)

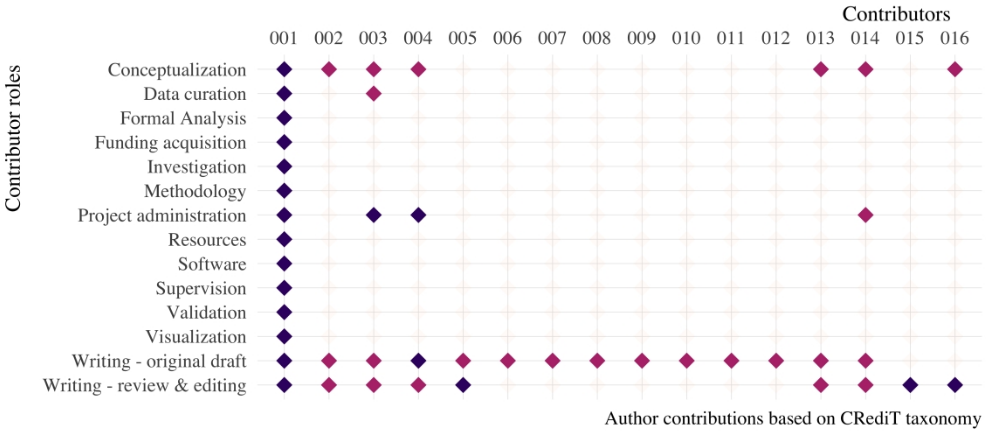

Figure 5: Author contributions according to the CRediT author roles taxonomy. Contributions are indicated as being substantial (dark diamonds) or moderate (light diamonds).

1481x666mm (72 x 72 DPI)

Author response to reviews of

Manuscript LING.2023.0249

Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research

submitted to *Linguistics: An Interdisciplinary Journal of the Language Sciences*

[RC] Reviewer comment Manuscript text

Dear Dr. Volker Gast,

Thank you for taking the time to consider our manuscript *Opening open science to all: Demystifying reproducibility and transparency practices in linguistic research* (LING.2023.0249) for publication in *Linguistics: An Interdisciplinary Journal of the Language Sciences*. Our understanding is that we needed to consider the reviewers' comments and revise accordingly before the manuscript could be reevaluated for publication. We noted three specific areas that were consistently highlighted by the reviewers as needing improvement: a 'superficial' treatment of the core practices, missing key references, and concrete examples related to the field of linguistics. We have considered thoroughly the detailed feedback provided by all three reviewers and resubmit what we believe to be a much improved version of the manuscript. In this letter we address the reviewers' concerns point-by-point. Where feasible, we quote all revised text in this document, otherwise, we refer the reader to the relevant sections of the revised manuscript. We thank you and the three anonymous reviewers for all comments and suggestions and again enthusiastically submit the revised manuscript for consideration in *Linguistics: An Interdisciplinary Journal of the Language Sciences*.

Sincerely,

(corresponding author)

1. Reviewer #1

[RC 1.1.] Let me begin by first clarifying my own positionality with regard to the subject at hand: I am an enthusiastic supporter of open science practices in linguistics, and have been for many years. So in this sense, I am solidly in favor of more exhortations in print to our colleagues to improve their practices, and recognize that many linguists are not trained in these practices and left to more or less fend for themselves. This is a sad state of affairs for our field and others, and needs to be corrected through outreach and education.

This paper, however, seems to me to be quite weak on two fronts. The first is what seems to be a concerning lack of awareness of recent work in linguistics to promote open science, especially with regard to research data. Of particular note is the absence of two fairly recent references in linguistics. The first is a highly-cited, multi-authored position paper calling for reproducible linguistics research that was published in this very journal in 2018:

Berez-Kroeker, Andrea L., Lauren Gawne, Susan Kung, Barbara F. Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard P. Meier, Nicholas Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible research in linguistics: A position statement on data citation and attribution in our field. Linguistics 56(1): 1–18. `https://doi.org/10.1515/ling-2017-0032`.

The second is an open access handbook from the very prominent MIT Press Open, with 56 chapters on open data management from more than 100 international authors across many subfields (which I refer to later in this review as OHLDM):

Berez-Kroeker, Andrea L., Bradley McDonnell, Eve Koller & Lauren B. Collister (eds.). 2022. The open handbook of linguistic data management. Cambridge, MA: MIT Press Open. `https://doi.org/10.7551/mitpress/12200.001.0001`.

And then of course there is the 2021 statement from the Linguistic Society of America on the Scholarly Merit of Open Scholarship in Linguistics: `https://www.linguisticsociety.org/content/statement-scholarly-merit-and-evaluation-open-` scholarship-linguistics.

It seems to me that the authors have not done their due diligence in terms of understanding the state of literature in the field, especially in light of their claims that there is no such literature.

That is the first major weakness of this paper. Misrepresenting the current state of the field makes it very easy for readers to dismiss the topic altogether. Papers urging the research culture of the entire field forward need to carefully and accurately describe the problem at hand, so that readers can recognize the problem for what it is. This is turn makes it harder for readers to ignore the problem, and encourages them to think carefully about how they can help solve it in their own work. My own review is a case in point: I don't recognize the world of linguistics as described in this paper – that is, one in which no linguists have really thought seriously about this issue before today – and I am writing a negative review despite the fact that I hold this topic dear.

We would like to begin by thanking the reviewer for their blunt honesty and the time they have taken to provide references and insight regarding both the topic and our manuscript. Without a doubt, their comments have helped us to reflect on the paper's shortcomings and we believe we have revised accordingly in order to address the two main weaknesses highlighted here. First, we have incorporated all of the suggested references – Berez-Kroeker et al. (2018) and Berez-Kroeker, McDonnell, Koller, and Collister (2022), particularly, but also Zwaan, Etz, Lucas, and Donnellan (2018) and Gawne and Styles (2022), among others – throughout the revised manuscript. Importantly, we wish to express that we have given careful thought to the assertion that we have misrepresented the current state of the field, which, without a doubt, was not our intention, and have kept this concern present as we have added these key references. In an effort to save space and time, we will not copy/paste all of the changed text in the revised manuscript here, but note the inclusion of the aforementioned references (and others) in paragraphs 1, 2, and 4 of the introduction and paragraphs 2, 4, and 6 of the open data section where we have found them to be particularly relevant.

Additionally, we would like to use a bit of space to address the following comment made by the reviewer: *It seems to me that the authors have not done their due diligence in terms of understanding the state of literature in the field, especially in light of their claims that there is no such literature*. While we disagree with the assertion that we have not done our due diligence, we humbly note that we never make the claim that there is no literature on open science in the field of linguistics. We do, in fact, cite a great deal of it. In the revised manuscript we have made it more clear that our work does not intend to present itself as being singular in this area, but rather as a complement to current efforts. There is no doubt that we did not cite *all* the relevant literature in the original version of the manuscript–certainly this is still the case in the revised version–but we maintain, then and now, that “few, if any, researchers have had explicit instruction on the practices of open science as part of their professional training” and that “current senior researchers were not trained in these innovative methodologies”, which has led to a situation in which “[. . .] all parties are forced to learn open science on their own, often without institutional support”. This is the reality of our research team and that of most of the researchers with whom we are acquainted. We know that this is certainly not the experience of everybody. Our intent is merely to address the issue as we understand it head on and to provide a helpful resource that can complement the literature that is currently available. In our humble view, the more books, articles, websites, blogs, voices, etc., pushing for open science practices in the field, the better. We do not claim to be pioneers in any way, but we are certain that our contribution has the potential to reach some individuals that are not familiar with any of the extant literature, and that others can benefit from revisiting familiar issues from our perspective. We view our contribution as additive in nature.

[RC 1.2.] The second weakness is that the eight steps outlined here are handled only superficially, and in some cases, in a way that can be potentially harmful. Each section could be a paper unto itself (and interesting special issue idea, perhaps?). Below I highlight a few of the major oversights in the sections.

Positionality statements. The section on positionality statements does not address the counterpoints that have been made regarding the extra burden that required statements put onto BIPOC authors. See, e.g.:

Massoud, Mark Fathi. 2022. The price of positionality: Assessing the benefits and burdens of self-identification in research methods. Journal of Law & Society 1-23.

Roberts, Steven O., et al. 2020. Racial inequality in psychological research: Trends of the past and recommendations for the future. Perspectives on Psychological Science 15: `https://doi.org/10.1177/1745691620927709`.

Open data. The section on open data, while acknowledging that privacy concerns should be respected, does not mention any discussion of the concerns of Indigenous peoples regarding colonial practices in linguistics and especially around who may own, and therefore share, language data (see eg Holton, Leonard & Pulsifer, chapter 4 of OHLDM and that chapter’s very helpful list of references).

Preregistration and registered reports. The section on registered reports does not take into consideration

very real pressures of the systems of academic reward, which bias university researchers against any additional steps in the research process that increase the time to publication (that is, the authors overlook the need for systemic change in academia to encourage open science – see Alperin et al, chapter 13 of OHLDM).

To address the second weakness pointed out by the reviewer, particularly with regard to our treatment of positionality statements, open data, preregistration and registered reports, we have considered the specific issues pointed out above, and, where relevant, expanded our discussion, reorganized, and included critical, missing references (e.g., Holton, Leonard, & Pulisifer, 2022; Leonard, 2021; Massoud, 2022; Mufwene, 2020; Roberts, Bareket-Shavit, Dollins, Goldie, & Mortenson, 2020; Singh, Killen, & Smetana, 2023; Tsikewa, 2021, among others).

Regarding positionality statements, we have heavily revised and reorganized this section. We have focused on including a discussion of the counterpoints of the inclusion of positionality statements particularly regarding marginalized individuals. As this entire section has been reorganized, we direct the reviewer to pages 7-10 of the revised manuscript.

Regarding open data, we have expanded our discussion regarding marginalized communities. We highlight here some of the changes, but also encourage the reviewer to consider the entirety of the revised section in context (see pages 10-14).

> Having stated all the above, it is necessary to recognize that linguistics faces a unique set of challenges with regard to data, as there are a multitude of subfields, each of which potentially works with a variety of data formats. Due to such diversity, one must determine which aspects of open science are relevant to their data. For example, a neurolinguistic study investigating event related potentials (ERPs) could share raw data for transparency, as well as preprocessed data with the code used to transform the raw data and a corresponding description for facilitation of reanalysis. In another field, the creation of a corpus will benefit from open access and the use of standardized file formats; the analysis of a corpus will benefit from sharing the search queries, the analysis code, and a description of the analysis code. At the heart of these challenges are ethical concerns that must be considered with care. First and foremost, the privacy and consent of participants must be safeguarded. Linguistic data often include personal information, which can be especially difficult to anonymize. While on the surface written and behavioral data may not appear to pose as many issues as audio and video recordings, which constitute a large portion of linguistic research materials, it is imperative that one consider the sources from which all types of data are derived. As expressed by Holton et al. (2022), if we haphazardly take language to represent trivial data points and lose focus on the individual embedded within a community, as well as the values of said community, we are doomed to "dehumanize and decontextualize" it (p. 50). This is particularly true when working with minority languages and/or marginalized communities. In cases such as these, the researcher must be held accountable, not only for the anonymization of participant information, but also for respecting and upholding the specific goals and restrictions put forth by the community. This includes, but is not limited to, the use, access, and storage of all collected data. In sum, careful consideration of the priorities of the researcher and the researched, which often do not align, is paramount (for more detailed views, see Adetula, Forscher, Basnight-Brown, Azouaghe, & IJzerman, 2022; Holton et al., 2022; Hudley, Mallinson, & Bucholtz, 2020; Leonard, 2021; Mufwene, 2020; Singh et al., 2023; Tsikewa, 2021, among others). In addition, generative artificial intelligence technologies, such as Large Language Models, are burgeoning. These technologies will certainly pose currently unknown challenges in the near future and may necessitate additional steps to secure the protection of sensitive data against misuse, particularly regarding adherence to the original agreement of informed consent, and, importantly, in upholding the conditions of use put forth by the stakeholders in marignalized communities.

As well as…

> To summarize, open materials are important because they facilitate transparency, rigor, reproducibility, replication, accumulation of knowledge, and, importantly, they make participating in the scientific endeavor more inclusive. According to some accounts, linguistics, in general, does not engage in open science practices, including sharing research materials (See Bochynska et al., 2023), though others characterize its participation in different terms. For instance, as stated in Berez-Kroeker et al. (2018) "Practitioners in different subfields 'do transparency' differently, and these practices could serve as models for an eventual amalgamated standard" (p. 9). While linguistics does face legitimate, field-specific challenges related to non-WEIRD communities, ultimately, the benefits of open materials outnumber many of these challenges. Researchers should take the stance to share what is reasonable and ethically responsible all the while holding at the forefront the priorities of the individuals from which the materials are derived, especially regarding data from marginalized communities.

Finally, we have also included discussion of incentive structure in academia and its effects on HARKing, p-hacking and early career researchers.

> Researchers have pointed to questionable research practices (QRPs), such as p-hacking–knowingly manipulating an analysis until a significant p-value is obtained (See Head, Holman, Lanfear, Kahn, & Jennions, 2015)–and HARKing–hypothesizing after the results are known (See Murphy & Aguinis, 2019)–, along with small sample sizes, poor theory, lack of transparency, misguided incentive structure in academia, etc., as factors that ultimately led to the replication crisis, though it is likely that many factors are/were simultaneously at play. For instance, the aforementioned QRPs may be an unfortunate consequence of misaligned incentive structures in academia, where publication is the universal currency. The pervasive pressure to publish likely leads many researchers to focus on quantity over quality. Couple this with the difficulty of publishing negative or null results, and the result is a research landscape in which many fields suffer from publication bias with little or no incentive to prioritize time consuming open science practices. Taking this into account, it is not hard to understand why some researchers may turn to QRPs. While it is difficult to quantify how prevalent QRPs are in a given field, in a survey of applied linguists, Isbell et al. (2022) found that 94% reported having engaged in one more, and 17% admitted to having committed some form of fraud.

Again, we only highlight some of the changes here, but suggest a reread of the relevant sections in the revised manuscript.

2. Reviewer #2

[RC 2.1.] The argumentation and description of methods is generally very abstract. If I understand the title "opening open science to all" correctly, the authors should include more concrete examples and instructions so that readers who are (still) unfamiliar with the practices described can implement new methods. I added concrete suggestions to do so below; this is necessary for the manuscript to fulfill its promise of opening open science and to become an important resource in research and teaching.

We thank the reviewer for their thoughtful insight. We took a close look at all of the issues raised and have made substantial revisions to the manuscript. Concretely, we have included many more examples and tried to be more inclusive by including in our discussions more of the distinct subfields of linguistics. Furthermore, we have taken the reviewer's advice and included 3 tables related to preprint servers, registered reports and preregistration, along with 2 new figures. We feel that the manuscript is now greatly improved and we appreciate the time and effort spent giving us such thorough feedback. Below we address all of the concerns raise, and, where relevant, copy/paste prose, figures, and tables.

[RC 2.2.] ERC: should be spelled out here for clarity, as it is only explained later in the introduction.

This acronym has been spelled out in the abstract of the revised manuscript.

[RC 2.3.] "Researchers have pointed to questionable research practices (QRPs), such as p-hacking and HARKing" –> HARKing (maybe also p-hacking) needs to be explained briefly here, especially given the goal of the paper.

In the revised manuscript we have defined both terms, included references, and elaborated a bit more on the relationship between QRPs and incentive structures in academia.

> Researchers have pointed to questionable research practices (QRPs), such as p-hacking–knowingly manipulating an analysis until a significant p-value is obtained (See Head et al., 2015)–and HARKing–hypothesizing after the results are known (See Murphy & Aguinis, 2019)–, along with small sample sizes, poor theory, lack of transparency, misguided incentive structure in academia, etc., as factors that ultimately led to the replication crisis, though it is likely that many factors are/were simultaneously at play. For instance, the aforementioned QRPs may be an unfortunate consequence of misaligned incentive structures in academia, where publication is the universal currency. The pervasive pressure to publish likely leads many researchers to focus on quantity over quality. Couple this with the difficulty of publishing negative or null results, and the result is a research landscape in which many fields suffer from publication bias with little or no incentive to prioritize time consuming open science practices. Taking this into account, it is not hard to understand why some researchers may turn to QRPs. While it is difficult to quantify how prevalent QRPs are in a given field, in a survey of applied linguists, Isbell et al. (2022) found that 94% reported having engaged in one more, and 17% admitted to having committed some form of fraud.

[RC 2.4.] "It necessitates that researchers implement new techniques with limited pedagogical resources and embrace alternative methods of disseminating their research, all of which constitutes a steep learning curve." –> Does it always, though? For instance, in the field of language typology, there are a number of qualitative studies that are based on a sample and annotations using reference grammars. Those

studies do not include any code, making them transparent simply means providing a spreadsheet with the languages and the respective annotations and sources that the authors should have in some form anyway. Can we really speak of "innnovative methodologies"?All this goes to say that, yes, some formats (e.g. using OSF) may involve a learning curve, but this is not necessarily so. It may also be about vulnerability, i.e. making all details of the study and analysis public and thus subject to potential criticism, wheareas before, those could be left partially "hidden" and "proctected". I suggest that the authors sightly rephrase this paragraph.

We appreciate the reviewer's thoughtful feedback. We agree that the challenges associated with increasing transparency vary across different fields and methodologies. In response, we have rephrased the paragraph to acknowledge that while some researchers may face a learning curve or feelings of vulnerability when adopting new transparency practices, this is not universally the case. The revised paragraph now reads as follows:

> One reason for the slow adoption in linguistics may be related to the fact that engaging in open science is no trivial feat. On the contrary, it often requires learning new skills, thoughtful planning, as well as an openness and willingness to share materials, code, and data. Many researchers need to implement new techniques with limited pedagogical resources and embrace alternative methods of disseminating their research, all of which can constitute a steep learning curve. That being said, what engaging in open science ultimately entails is sure to be field-specific and vary accordingly. In some disciplines, for instance, it may only involve a few of the practices we outline in the present work without the need for innovative methodologies. Nonetheless, given how new open science practices are, it is reasonable to assume that current senior researchers were not trained in these innovative methodologies. As a consequence, many early career researchers (ECR) find themselves at a crossroads in which they are forced to learn open science on their own, often without institutional support. Ironically, there is also a growing expectation that ECRs implement these novel tools in order to be successful in their programs, on the job market, or to advance in their careers.

[RC 2.5.] **"To this end, we identify three areas, stance, workflow, and dissemination, in which linguists can engage in open science." –> rephrase: "... we identify the following three areas of stance, workflow and dissemination, ...**

This change has been included in the revised manuscript.

[RC 2.6.] **Figure 1 –> I am not sure open data is clearly only related to stance; I would also view providing the datasets in a useful form as part of the workflow. Could this point be put in between the two areas?**

The reviewer makes a good point. Open data, in theory, could fit in all three areas depending on what exactly one is referring to, i.e., the stance to make data open or not, how making data open is incorporated into one's workflow, and how one goes about actually disseminating the data. We have overhauled the figure and include it here for convenience (see below).

[RC 2.7.] **Could you give a concrete example of a positionality statement early on? A concrete example is valuable to make this more understandable for researchers not familiar with this practice.**

In the revised manuscript this section has been reorganized to reflect changes suggested by all three reviewers. We follow the reviewers suggestion and now provide an example of a positionality statement in prose in the

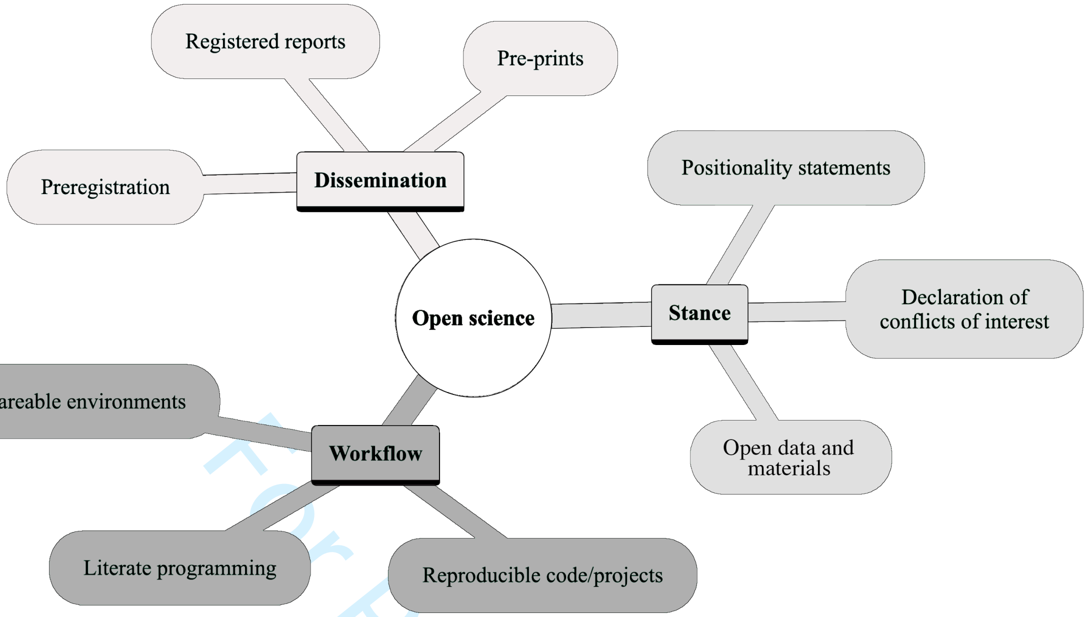

Figure 1: Some open science practices amenable to research in linguistics as they pertain to one's stance, workflow, and the dissemination of research products.

first paragraph, with links to more examples at the end of the section.

[RC 2.8.] "Moreover, Bucholtz et al. (2023) note that considering a researcher's positionality may be especially important in linguistics, "[. . .] which relies on racially minoritized communities as sources of data yet lack adequate (if any) representation of those communities among faculty researchers" (p. 2)." –> Does this quote really apply to linguistics in general? I can see how it is very relevant for certain areas of linguistics, but there is much linguistic research on major languages, where this point does not seem to apply. Please rephrase / explain.

The reviewer's point is duly noted. In the revised manuscript we have rephrased this sentence to make it more clear that we are referring to work on minority languages. The sentence in question now reads as follows:

> The support and advocacy for the inclusion of positionality statements in research publications is increasing (Bucholtz et al., 2023; Jafar, 2018; Steltenpohl, Hudson, & Klement, 2022). Bucholtz et al. (2023) note that considering a researcher's positionality may be especially important in linguistic research on certain language communities, such as indigenous communities, "[. . .] which relies on racially minoritized communities as sources of data yet lack adequate (if any) representation of those communities among faculty researchers" (p. 2). Nonetheless, others contest this practice.

[RC 2.9.] "Many researchers support and advocate for the inclusion of positionality statements in their research publications (e.g., Bucholtz et al., 2023; Jafar, 2018; Steltenpohl et al., 2022)." –> Please specify the research areas here, also those mentioned as arguing against positionality statements.

This section has been heavily edited and reorganized in the revised manuscript. This particular sentence no longer appears as such. That being said, the revised version now expands the discussion around positionality to include more research areas and dedicates more space to discussing the arguements against them.

[RC 2.10.] "Bochynska et al. (2023) surveyed open and transparent practices in linguistics and found that only 10% of the articles sampled included statements of conflict of interest, and, among those 10%, none declared any conflicts (See also Cristea & Ioannidis, 2018; Hardwicke et al., 2022, 2020)." –> Out of curiosity, can you give an example of a statement of conflict of interest from linguistics? Again, concrete examples would help readers less aware of / familiar with these issues understand what those statements can look like.

We thank the reviewer for this welcomed suggestion. We have added references and links to examples in linguistics, psychiatry, psychology, and the social sciences more broadly. We include the revised text for convenience.

> Bochynska et al. (2023) surveyed open and transparent practices in linguistics and found that only 10% of the articles sampled included statements of conflict of interest, and, among those 10%, none declared any conflicts. For a clear example of what a declaration of conflicts of interest can entail in linguistics, the interested reader is directed to Bochynska et al. (2023). Of particular value are the *Competing interests* section and the coding form available at `https://escholarship.org/uc/item/6m62j7p6#main` and `https://osf.io/ehyx3`, respectively. Additionally, Cristea and Ioannidis (2018), Hardwicke et al. (2022), and Hardwicke et al. (2020) represent illustrative examples in psychiatry, psychology, and the social sciences more broadly.

[RC 2.11.] "While positionality statements, due to their reflexive nature, may encompass larger pieces of writing, they can also take the form of short paragraphs that illustrate a few personal characteristics deemed relevant for the particular research endeavor." –> This whole paragraph should be moved to the beginning of the section.

This section was heavily edited and this particular part, in its revised form, is not at the beginning of the section.

[RC 2.12.] "Gabriela is a white immigrant cis-gender woman from Romania whose research focuses on how non-native speakers are ideologically framed as linguistically deficient in comparison to native speakers who are characterized by their linguistic authority and expertise." –> OK, this is what I wanted to see early on in this section. Please add some context: What was the title / topic of the publication that this statement belongs to? Can you give a concrete reference? Or is this a made-up example? Please clarify.

This particular example, which now appears at the beginning of the section, is for one of the authors of the present work. For the sake of anonymity during the peer review process, more details will be included at a later stage.

[RC 2.13.] "For examples of positionality statements in linguistic research, the interested reader is directed to Bochynska et al. (2023) and Weissler et al. (2023)." –> Great pointer to ressources, but if the purpose of this paper really is to make open science practices more accessible to linguists who are not (yet) familiar with them, a concrete example in the beginning of this section is important. It would also help to make

this proposal of including a positionality statement more concrete: In which areas of linguistics is it helpful? In which areas is it necessary? In which areas may it be less relevant? For which methods may this be particularly helpful / important? While I sympathize with the arguments made in this section in general, I think they are too abstract in order to be immediately helpful for a concrete implementation. Please make the argumentation a bit more concrete and geared towards certain / different areas of linguistic research so that the reader has more practical advice on how to include such statements in their future research.

We thank the reviewer for the aforementioned suggestions. They have guided our rewriting/reorganizing of this section, which we believe is now much more informative and useful. As opposed to copy/pasting the entire section here, we encourage the reviewer to consider the revised section in its entirety (approx. pages 7-10).

[RC 2.14.] **“In academic research, statements such as “data available upon request” are commonplace.” –> This may be so, but the statement would clearly look better with references or some data to back it up. I do not mean that the authors should single out studies as negative examples, but maybe it would be possible to mention journals and years of publication where this happens. This makes it somewhat more tangible and related to linguistic research. Please rephrase.**

We have added a reference to the aforementioned sentence (See Hardwicke & Ioannidis, 2018), which contains enlightening statistics from the field of psychology.

[RC 2.15.] **“Recent efforts have attempted to encourage researchers to make linguistic data open and accessible via servers (e.g., the IRIS database).” –> Given the goal of this paper, please add more explanation about the IRIS database here.**

“open science bagdes” –> Can you explain briefly what this is?

In the revised manuscript we have expanded upon our discussion of the IRIS database and open science badges. The relevant text is included below.

> Researchers are increasingly encouraged to make linguistic data open and accessible via servers. An illustrative example is the IRIS database (`https://www.iris-database.org`), a language sciences digital repository that is freely accessible and permits the up- and downloading of research instruments and materials. Additional efforts include open science badges–visual symbols offered by some journals (e.g., *Language Learning*, *Language and Speech*) on published articles. These badges are awarded to researchers for adhering to certain open science principles, such as sharing code, data, or preregistering a study. In arguably more extreme cases, other journals have made data sharing a requirement for publication (e.g., *Applied Psycholinguistics*).

[RC 2.16.] **“Though researchers may understandably hesitate to share their data, we believe understanding the benefits of open data can help alleviate any concerns.” –> This point needs to be elaborated more in that there may be very different motivations for preferring not to share data. Corpus data may contain sensitive personal information (e.g. video, audio) and speakers may not consent to making it public. Authors of the study may not have the rights to publish the data that their study is based. In this case, it may be possible to make processed data publicly available, but not the raw data (i.e. data can mean different things, and should be distinguished in more detail here). In other cases, it may simply be the**

fear of making all data and annotation choices criticiseable by making everything publicly available. Those are very different perspectives that call for very different solutions / approaches. This needs to be discussed and differentiated in more detail.

We thank the reviewer for these useful insights. In the revised manuscript we have reorganized this section and rewritten many paragraphs to better reflect these concerns. In addition, we have addressed how certain choices become criticiseable upon being made public and included a relevant reference (See Stieglitz et al., 2020). Given the large amount of changes, we do not copy/paste any text here.

[RC 2.17.] **"It affords third parties the opportunity to scrutinize original findings, which promotes reproducibility and reduces errors, such as those related to statistical analyses and reporting of outcomes [TIMO]." –> What is [TIMO]?**

We apologize for this oversight. TIMO was a 'note to self' to include a reference that was not deleted before submitting the initial version of the manuscript. This has been corrected in the revised version.

[RC 2.18.] **"Revisiting old data sets using innovative techniques can support or contradict past narrative conclusions (e.g., Casillas, 2021)." –> It would be useful to elaborate this in a few sentences and present the Casillas (2021) study in terms of revisiting old data sets and drawing (new?) conclusions.**

The revised manuscript now includes a brief discussion of this study and its relevance for open data, which is included below.

> Revisiting old data sets using innovative techniques can support or contradict past narrative conclusions. For instance, using meta-analytic techniques, Casillas (2021) reexamined extant research regarding 'compromise categories' in early bilinguals. This line of research posits that bilingual individuals produce speech sounds intermediate to those produced by monolingual speakers of either language. By systematically reevaluating prior data and incorporating new acoustic analyses of coronal stops from early Spanish-English bilinguals, Casillas (2021) suggested that the cumulative evidence for 'compromise' stop categories was negligible. In lieu of intermediate phonetic categories, the study proposed early bilinguals can exhibit performance mismatches resulting from dynamic interlingual interactions. This reanalysis contradicted earlier assumptions about bilingual phonology and provided in-depth scrutiny of statistical power and evidence accumulation in bilingualism research.

[RC 2.19.] **"Open data are particularly important for the field of linguistics, for all of the aforementioned reasons, and also because linguistics is Western, Educated, Industrialized, Rich, and Democratic (WEIRD, Bochynska et al., 2023)." –> I do not disagree, in principle, but there may be one caveat that needs to be mentioned: Bochynska et al. (2023) evaluated publications written in English. This seems slightly self-selecting to me, as it probably excludes relevant linguistic research traditions whose main language of publication is not English (e.g. Chinese linguistics). How sure can we then be that linguistics as a whole is WEIRD? I suggest the authors tone down this statement a bit, as it seems too strong in its current form.**

We agree with the reviewer's assessment and have toned down the message accordingly. In addition, we have added more references that highlight this assertion about the field. The revised text is included below.

> Open materials are particularly important for the field of linguistics, for all of the aforementioned reasons, and also because some linguists have described the state of the field, as far as English-language publications are concerned, as being Western, Educated, Industrialized, Rich, and Democratic (WEIRD, see Bochynska et al., 2023; Faytak et al., 2024; Nagle, Baese-Berk, Amengual, & Casillas, 2024). That is to say, the majority of linguistic research appears to be concentrated on specific languages, mainly Indo-Germanic, in overrepresented communities, by privileged scholars.

[RC 2.20.] “Making linguistic data accessible to all researchers promotes participation in and with underrepresented communities. Furthermore, it can increase the study of diverse and underreported languages, which fosters a more inclusive and comprehensive understanding of the global linguistic landscape.” –> Please explain why and how accessible linguistic data promotes the researchers’ participation with the (potentially underrepresented) speaker communities? While I am sympathetic to this argument, it needs to be spelled out more and made more concrete in order to be meaningful.

We have tried to make this point more clear in the revised manuscript. We include the relevant paragraph here for convenience.

> Open materials are particularly important for the field of linguistics, for all of the aforementioned reasons, and also because some linguists have described the state of the field, as far as English-language publications are concerned, as being Western, Educated, Industrialized, Rich, and Democratic (WEIRD, see Bochynska et al., 2023; Faytak et al., 2024; Nagle et al., 2024). That is to say, the majority of linguistic research appears to be concentrated on specific languages, mainly Indo-Germanic, in overrepresented communities, by privileged scholars. Making materials in linguistic research accessible to all researchers can promote participation *in* and *with* underrepresented communities. Furthermore, it can increase the study of diverse and underreported languages by affording more researchers the opportunity to interact and learn from data that would otherwise not be available to them, which, in turn, can foster a more inclusive and comprehensive understanding of the global linguistic landscape.

[RC 2.21.] “Having stated all the above, it is necessary to recognize that the field of linguistics faces unique challenges with regard to open data.” –> Are these necessarily unique challenges? I do not think it is important to point this out (and I am not sure if linguistics really faces unique challenges that different from other related disciplines). What would be more important is to go into more detail about which areas of linguistics face which types of challenges, since there are so many different types of data that we as linguists work with (experimental data, corpus data, data from grammars, data from linguistic databases, written, spoken, signed data, historical data, and each oft these types could be distinguished further). This should be discussed in much more detail and made more concrete so that the readers can see their approaches / types of data represented in this discussion.

We again thank the reviewer for this helpful comment. This section has been reworked to address the challenges researchers face in the field of linguistics, and provide a more nuanced view regarding marginalized communities. We refer the reviewer to pages 11-16 (appox.) of the revised manuscript.

[RC 2.22.] “The privacy and consent of participants must be safeguarded. . . ” –> OK, this is the kind of information that I was waiting for earlier; consider moving this paragraph up. Also, not all linguistic studies face ethical issues: What about studies that work with databases / corpora that have been compiled by third

parties, which are publicly accessible and which the authors simply use as is. They are far removed from any speakers in this case and had no influence on any decision of data collection and processing that went into compiling the database. Or what about historical linguistics? The role that participants or informants play in different types of linguistic studies deserves a more fine-grained discussion.

Following the reviewer's suggestion, this section was reorganized to address their concerns. We refer the reviewer to pages 11-16 (appox.) of the revised manuscript.

[RC 2.23.] "Moreover, the advent of generative artificial intelligence technologies, such as Large Language Models, may pose unknown challenges in the near future that necessitate additional steps to secure the protection of sensitive data against misuse." –> Please explain in more detail: Is the idea that LLMs use data compiled by linguists for linguistic research? Or does the potential of misuse lie in the linguistic research with LLMs itself? If the authors wish to mention this, they should discuss in more detail what exactly the risks are. Otherwise, this sentence could be omitted.

We thank the reviewer for recommending clarity regarding this point. We are referring to the fact that LLMs are new and we don't fully understand how they will be used in the future, by both academics and the general public. Given the fact that they operate using/are trained by large amounts of text data, the availability of open data/corpora could lead to potentially sensitive data (e.g., data derived from marginalized communities) being used in ways that do not adhere to the original agreement of informed consent. We have clarified this point in the revised manuscript and include the relevant text below.

> In addition, generative artificial intelligence technologies, such as Large Language Models, are burgeoning. These technologies will certainly pose currently unknown challenges in the near future and may necessitate additional steps to secure the protection of sensitive data against misuse, particularly regarding adherence to the original agreement of informed consent, and, importantly, in upholding the conditions of use put forth by the stakeholders in marignalized communities.

[RC 2.24.] "When primary data, such as audio or video files, cannot be shared, derived data in the form of tabular files can take its place…" –> This paragraph is very important and constructive. Different types of data (primary data vs. processed data and annotations) should be introduced at the beginning of this section, though, and the entire section should be more precise about what types of data particular arguments refer to.

As mentioned above, this section has been rewritten/reorganized to reflect suggestions by all reviewers. We refer the reviewer to pages 11-16 (appox.) of the revised manuscript.

[RC 2.25.] Prolific –> add a link, e.g. in a footnote.

A link to the prolific website has been included in the revised manuscript.

[RC 2.26.] "In more uncommon cases in which institutional policies do not permit the sharing of derived data sets, synthetic data containing the same statistical properties can be generated and shared freely (See Quintana, 2020)." –> Please elaborate how this would work.

In the revised manuscript we have included more information regarding how the process works and provided links to code and an example tutorial.

> In more uncommon cases in which institutional policies do not permit the sharing of derived data sets, synthetic data containing the same statistical properties can be generated and shared freely (See Quintana, 2020). In short, the method consists of capturing the statistical properties of the original data set and using them to simulate new data that preserve the relationships between the variables of interest. A tutorial on the method described in Quintana (2020) is freely available on github (`https://github.com/elifesciences-publications/synthpop-primer`), and an online RStudio instance can be accessed at `https://mybinder.org/v2/gh/dsquintana/synthpop-primer/master?urlpath=rstudio`. All relevant materials are available on the OSF: `https://osf.io/z524n/`.

[RC 2.27.] **“A substantial hurdle that cannot be overlooked revolves around the fact that researchers must learn to use new technologies to participate in open, transparent research. Making data open and accessible is not as simple as merely uploading a data file.” –> I disagree in that it depends on the linguistic subdiscipline. In large-scale typological studies that are qualitative in nature and use no statistics but are based on a language sample, making research transparent simply consists of adding a supplementary spread sheet file with the names of the languages, all relevant annotations / examples and their sources. Impressionistically, even this is still not done in all studies, so improving on transparency in this case only involves adding a supplementary file to the publication. Also, if the researcher does not provide such supplementary files on their own, many journals offer to host supplementary materials on the journal website together with the publication. This may not be the best solution, but it is a solution where the researcher really does not need to know or consider platforms to host data. Please clarify this.**

The reviewer’s disagreement regarding this issue is duly noted. We believe we failed to fully describe and contextualize the difficulties involved with making data open and accessible. In the revised manuscript we have further elaborated, providing examples and templates. For the sake of completeness, we describe the issue here. The logic behind our argument revolves around the fact that simply providing access to data (or a data file) is not sufficient. To elaborate using the reviewer’s example, we will consider the case in which a supplementary spread sheet file with the names of languages represent the data in question. In this case, uploading the file, be it on an open, independent server or on the journal’s website, does not provide the relevant context for a non-expert to use the data. This is the case because, when publicly sharing data, it is ideal that one also provide three distinct levels of documentation so that an independent researcher has the proper context to use data. It is helpful to have in mind a researcher that may not be completely familiar with the norms of a particular subfield. This part is key because the objective is that any individual be able to use the data. The levels of documentation are *project-level* (i.e., a project summary document), *data-level* (i.e., a README file explaining the data set), and variable-level (i.e., a data dictionary). The different levels will be more or less relevant depending on the subfield. In our example, it could theoretically be critical for an independent researcher to understand how and why the data are relevant to the broader goals of the project (project-level), as well as how and why the contents of the data file (in our case the languages) came to be included and any relevant relationships or hierarchies among them (data-level, e.g., language families, varieties of a single language, etc). In this particular case, a data dictionary (variable-level) could also be relevant to understand how the languages are listed, i.e., whether or not abbreviations are used, or if the names are interpretable to speakers of other languages. If it wasn’t obvious, this particular example is not in the area of our (the authors) collective expertise. Nonetheless, we believe this makes the example even more relevant, as the topics we have covered represent aspects of our ignorance in this area that we believe we may need to know in order to make full use of a data set of this nature. Importantly, the type of thoroughness we outline here is virtually non-existent in most data sets publicly available. For this reason we also have included links to templates any researcher could use and adapt for their specific purposes. In the revised manuscript there relevant text reads as follows:

> Another substantial hurdle that cannot be overlooked revolves around the fact that researchers must learn to use new technologies to participate in open, transparent research. Making materials open *and* accessible is not as simple as merely uploading a data file. Ideally, researchers should include relevant information to contextualize the data set at the project-level (i.e., a project-summary document), the data-level (i.e., a README file explaining the data set), and the variable-level (i.e., a data dictionary) (Lewis, 2024). The inclusion of resources at these three levels is the optimal way for authors to provide the necessary context for an independent researcher to access and utilize their materials Unfortunately, most publicly available materials do not adhere to this standard. For this reason, we direct the interested reader to templates provided in Lewis (2024) for documentation at the project-level (`https://osf.io/q6g8d`, `https://osf.io/d3pum`), data-level (`https://osf.io/tk4cb`), and variable-level (`https://osf.io/ynqcu`). In addition, the reader is referred to the project, data, and variable level documentation of the present project, all of which are freely available on the OSF: `https://osf.io/bsu2q/?view_only=68d1e41b327f4a28a9fcd0fc6537ecaf`.

[RC 2.28.] "Free repositories designed for the purpose of sharing research materials, such as the Open Science Framework (Foster & Deardorff, 2017), github, etc., are preferable and can be accessed simply by sharing a link." –> Here, it would actually be helpful to spell this out a bit more and mention e.g. zenodo as another option. For readers less familiar with these options, it would be valuable to add a table with different platforms / systems and provide information about long-term support, version control, possibility of DOI, anonymous links (e.g. what OSF offers and what is very useful for sharing data and code during the revision stage of a study), etc.

We thank the reviewer for this idea. In the revised manuscript we have included a table summarizing some of the available resources. We include the table here for convenience.

Platform	Long-term Support	Version Control	DOI Assignment	Anonymous Sharing	Key Features
Open Science Framework (OSF)	+	+	+	+	Project management and collaboration
GitHub	+	+	Integrates with Zenodo	+ (public repositories)	Project management/collaboration, ideal for coding
GitLab	+	+	Integrates with Zenodo	Limited	Project management/collaboration, ideal for coding
Bitbucket	+	+	Integrates with Zenodo	Limited	Project management/collaboration, ideal for coding
Zenodo	+	+ (via GitHub integration)	+	+	Supports range of file types
Figshare	+	Limited	+	+	Sharing datasets and figures
Box	-	-	-	Limited	Basic file storage and sharing
Google Drive	-	-	-	Limited	Basic file storage and sharing

[RC 2.29.] "While linguistics does face legitimate, field-specific challenges, ultimately the benefits of open data outnumber these challenges, and researchers should take the stance to share what is ethically reasonable." –> My impression when reading this is that the only field-specific challenge mentioned was the point on sensitive speaker data. I generally agree with the statement, but in order to make this statement at the end of the section, the section would need to be much more explicit about the different challenges taking into account the different areas of linguistics.

Another point to consider is the following, although it may not immediately relate to data and more to the inclusion of lesser studied languages and integration of speaker communities in linguistic research: There are linguistic journals that allow / require an additional abstract of the papers in the target language or another local language besides the usual abstract in English. From personal experience, I know that Linguistics Vanguard offers this as an option (although I am not sure if you can find this information other than submitting a paper). Another journal that makes use of this practice is the Journal of African Languages and Linguistics (just to give some examples from the most recent volume: `https://www.degruyter.com/document/doi/10.1515/jall-2023-2008/html`, `https://www.degruyter.com/document/doi/10.1515/jall-2023-2011/html`)

Again, the reviewer's point is duly noted. We have reorganized and rewritten large portions of this section to include more field-specific challenges. For the sake of convenience we highlight but one here and recommend a reread of the section in its entirety.

> Having stated all the above, it is necessary to recognize that linguistics faces a unique set of challenges with regard to data, as there are a multitude of subfields, each of which potentially works with a variety of data formats. Due to such diversity, one must determine which aspects of open science are relevant to their data. For example, a neurolinguistic study investigating event related potentials (ERPs) could share raw data for transparency, as well as preprocessed data with the code used to transform the raw data and a corresponding description for facilitation of reanalysis. In another field, the creation of a corpus will benefit from open access and the use of standardized file formats; the analysis of a corpus will benefit from sharing the search queries, the analysis code, and a description of the analysis code. At the heart of these challenges are ethical concerns that must be considered with care.

[RC 2.30.] "As we have seen, reproducibility is now a crucial aspect of any scientific study." –> Where did we see that? Also, I agree that it should be, but reproducibility is not yet a crucial aspect of any scientific study, unfortunately. Please rephrase.

We have rephrased this sentence in the revised manuscript. It now reads as follows:

> Having seen the consequences from the reproducibility crisis in other fields, reproducibility must be a crucial aspect of any scientific study.

[RC 2.31.] "At worst, a lack of reproducibility can lead to irreproducible results and wasted resources." –> Is this really the worst consequence? Is it not worse (or equally bad) that we may end up with results that are taken as a given for decades in a given linguistic field, because the original study was not reproducible and never replicated but influential for some reason?

We have rewritten this paragraph. In the revised manuscript it now appears as follows:

> In general, reproducibility helps to increase the credibility of research findings and allows other researchers to verify and build on existing work. A lack of reproducibility can lead to findings that cannot be replicated, resulting in wasted resources, and, conceivably, downstream impacts on public health and policy decisions that are often grounded in funded research. For these reasons, among others, transparency in research methods are essential to ensure reproducibility, which includes not only the data collection and analysis methods, but also the code used to conduct the analysis. In linguistics there is increasing awareness of the importance of reproducibility and how a lack thereof could potentially impede advancements in linguistic theory and theories of language acquisition, in addition to having implications for education and language policy decisions based on research findings. As a consequence, many investigators are showing heightened interest in safeguarding the reproducibility of their research.

[RC 2.32.] "This can have serious implications for public health and policy decisions based on research findings." –> Is this statement about linguistics or research in general? Please clarify.

"In linguistics there is increasing awareness of the importance of reproducibility, and many researchers are beginning to take steps to improve the reproducibility of their research." –> too many reproducibil-

ities in one sentence, rephrase.

These concern were addressed in the rewriting of the original paragraph (see response directly above).

[RC 2.33.] **"There are several steps that researchers can take to make their code and projects more reproducible." –> A large portion of linguistic work is qualitative and does not rely on any code. If the paper is meant to be relevant to all linguistic approaches in general, please be more inclusive. Alternatively, the topic of the paper could be specified more and refer to "quantitative linguistics"?**

Again, the reviewer's point is duly noted. There are some areas of open science that are clearly more applicable to certain subfields of linguistics. As the reviewer points out, this particular section is more relevant for quantitative research. We have rewritten the beginning of this paragraph to highlight this fact. The revised text now reads as follows.

> For quantitative research, there are several steps that researchers can take to make their code and projects more reproducible. One approach is to create reports that document the research process by including descriptions of the data, the methods used to analyze the data, and the results.

Notably, other sections of the manuscript are more applicable to qualitative research (e.g., preregistration and registered reports) and highlighted at a later juncture.

[RC 2.34.]

- **regarding the sharing of code, dependencies, versions & reproducibility: It may be helpful to explain in more detail why it is important to not only share the code, i.e. that different operating systems can react to code differently, functions can differ in different versions of packages, etc. The authors may consider referring to Roberts et al. (2015), who show that functions from the lme4 R package produce different results when used on different operating systems (althouh fixed since then). Roberts, Seán, James Winters & Keith Chen. 2015. Future tense and economic decisions: Controlling for cultural evolution. PLOS ONE 10(7). e0132145. `https://doi.org/10.1371/journal.pone.0132145`.**
- **The literate coding section is very much focused on R. This may be warranted given that R seems to be the main programming language used for statistical analyses in many (?) linguistic areas, but python is certainly not uncommon either. If the authors go into detail for literate programing in R (e.g. citing the knitr package), they should also do this for python, and potentially mention Julia, which is also, although less commonly, used in linguistics. Here again, a table with programing languages, approaches / formats / packages as well as platforms with virtual environments would be very helpful.**
- **"This includes updating code and documentation as needed and testing projects on different operating systems to ensure that it can be run in different environments." –> This statement seems to assume that researchers should have access to different OSs, which may not always be the case. Also, it is hardly possible to test for all compatibilities and eventualities on the side of the authors. It is more reasonable to create and publish an, e.g., Docker environment for an R project, so that other researchers can reproduce the original analysis with the exact OS and package versions as in the original study.**

The reviewer brings up excellent points. We address them here together, as they relate to a common theme. In the revised manuscript we have taken care to explain *why* it is important to share more than just code and how reproducibility can be affected by system and platform specific issues. With regard to the literate programming section, we have made sure to discuss more of the available languages and platforms as well. In this particular case, we opted not to include a table. Additionally, we rephrased our treatment of updating and testing code (we do not intend to imply that a researcher has the obligation to test their projects on all conceivable platforms). To underscore the usefulness of creating public instances of projects (e.g., via Docker), we have made our project available there as well. We mention this in this section and will provide the working link upon completing the project. This is also true for the OSF and github repositories. For the time being, the anonymous version of the OSF repo is available here: `https://osf.io/bsu2q/?view_only=68d1e41b327f4a28a9fcd0fc6537ecaf`.

[RC 2.35.] "Linguistic research is multifaceted and spans diverse areas such as corpus analysis, conversation/discourse analysis, experimental research, and more." –> This is a very reduced list of linguistic research areas, please add more areas and discuss in more detail for which areas preregistration makes sense. For instance, can it be used in theoretical linguistics? It is mentioned in the beginning that "we focus on who might want to consider preregistrations", but this does not become very clear. Please add at least a paragraph about what types of linguistic studies can benefit from preregistrations.

We agree with the reviewer's assessment of this sentence. We have rewritten it to better represent some of the subfields of linguistics.

> Linguistic research is multifaceted and spans diverse areas such as phonetics, phonology, syntax, morphology, sociolinguistics, natural language processing, and conversation/discourse analysis, to name just a few. These areas range from purely theoretical to quantitative and experimental, with many falling somewhere in between.

In addition, we have added more detail throughout this section regarding *who* can/should use preprints and for *what* purpose.

[RC 2.36.] "Researchers face vital decisions while engaging in research, with inherent flexibility involved in the process of designing and conducting experiments, as well as analyzing the results (Simmons, Nelson, & Simonsohn, 2011)." –> Does this mean you only consider experimental linguistics research? This needs to be framed differently, or, the scope of the paper should be reduced to experimental linguistics.

We thank the reviewer for this helpful comment. We have taken care to be more inclusive in our wording, and, in turn, more faithful to the point being made by Simmons, Nelson, and Simonsohn (2011). The revised text now reads as follows:

> Researchers face vital decisions while engaging in research, with inherent flexibility involved in the process of designing and carrying out projects, as well as in the analysis of the data and interpretation of the results (Simmons et al., 2011).

[RC 2.37.] Coretta et al. 2023: Please explain in a few sentences what the study did and what it showed; it is impossible to understand (and appreciate) that without knowing the study.

We have included a description of the Coretta et al. (2023) study in order to facilitate understanding/appreciating "researcher degrees of freedom". The revised text is included below.

> This type of flexibility, termed "researcher degrees of freedom", can have serious down-stream consequences in quantitative research, particularly in linguistics. For instance, Coretta et al. (2023) provided the same speech-production data set to different research teams and asked them to answer the same research question. They found substantial variability in both the acoustic analyses and the analytic strategies, neither of which could be explained by analysts' prior beliefs, expertise, or the perceived quality of their analyses. Crucially, these decisions, both acoustic and analytic, impacted the teams' answers to the research question. To provide a simple example, a researcher studying lexical stress could concentrate on distinct acoustic cues typically associated with stress, i.e., pitch, duration, and intensity. Beyond selecting acoustic cues to measure, she must also select a domain for these measurements, such as the mid-point of stressed/unstressed syllables or an average value over the entirety of the syllable. Choices such as these, i.e., the researcher degrees of freedom, can wield significant influence on subsequent outcomes. Preregistration serves the purpose of meticulously documenting these choices *a priori*, thus acting as a deterrent against QRPs, like HARKing or p-hacking (Wicherts et al., 2016).

[RC 2.38.] This section needs more concrete pointers to specific platforms, journals, spaces for preregistration. Linguists outside of experimental research may want to do this, but are less likely to know how and where exactly they can preregister their study. The abstract motivations are spelled out clearly in this section, but it is not of much concrete help to those linguists who are convinced about the method but have no experience with it. It would be helpful to see an example of how this can be done outside of experimental studies, e.g. a corpus study, a typological study, etc. What about e.g. a theoretical morphological / phonological / syntactic analysis based on information from grammars? Is preregistration even possible or sensible? Please add more concrete details regarding different areas of linguistics so that this section becomes usefull to the readers

We thank the reviewer for this excellent suggestion. In the revised manuscript we have included more resources related to preregistration in distinct subfields of linguistics. Though multiple platforms exist for preregistering studies, we have opted not to include a table summarizing the options. We made this decision based on the fact that there are two primary platforms used generally, OSF and aspredicted.org, the former much more commonly used than the latter. The remaining platforms (e.g., ClinicalTrials.gov, European Union Clinical Trials Register, PROSPERO, SRCD, etc.) are all field-specific, and, therefore, not relevant for linguistics.

[RC 2.39.] Please give examples of journals / platforms that allow for registered reports.

We again thank the reviewer for a excellent comment that resulted in the inclusion of extremely relevant information. The revised manuscript now includes data and a summary table regarding registered reports in linguistics. The reviewer is directed to pages 25-26 (approx.) of the revised manuscript.

[RC 2.40.] "First, select a pre-print server that aligns with the course of research." –> Please give examples of suitable pre-print servers for different linguistic areas. Again, a table would be very helpful.

At the reviewers request, the revised manuscript now includes examples of suitable pre-print servers, as well as a table summarizing some of the interesting details.

Server	Discipline(s)	Database Size	Year Created	URL
LingBuzz	General Linguistics	+8,000	2006	https://ling.auf.net/lingbuzz
Open Science Framework	Multidisciplinary (includes Linguistics)	+3M	2011	https://osf.io/preprints
PsyArXiv	Psychology, Cognitive Sciences, Psycholinguistics, Linguistics	+30,000	2016	https://osf.io/preprints/psyarxiv
Cogprints	Multidisciplinary (includes Cognitive Sciences and Linguistics)	+4,000	1995	https://web-archive.southampton.ac.uk/cogprints.org/
SocArXiv	Social Sciences (includes Sociolinguistics)	+10,000	2016	https://osf.io/preprints/socarxiv
EdArXiv	Education Research (includes Applied Linguistics	+1,000	2018	https://osf.io/preprints/edarxiv
Computational Linguistics Open Archive (CLARIN)	Language-Based Research	N/A	2012	https://www.clarin.eu/
ACL Anthology	Computational Linguistics and NLP	+10,000	2004	https://aclanthology.org/
arXiv	Multidisciplinary (includes Computational Linguistics, NLP)	+2,5M	1991	https://arxiv.org/
SciELO Preprints	Research pertinent to Latin America, Spain, Portugal and South Africa	+140,000	1998	https://preprints.scielo.org/index.php/scielo/preprints
HAL (Hyper Articles en Ligne)	Multidisciplinary (includes Language-Specific French Linguistics)	+1M	2001	https://hal.science/

[RC 2.41.] "We have provided descriptions and relevant examples of these practices to accompany the many guides already available for learning open science (e.g., Crüwell et al., 2018; Lewis, 2020). Crucially, the purpose of this article is to help foster open science in linguistics (FOSIL)." –> The current version of this manuscript provided detailed motivations and descriptions, but not sufficient examples and concrete instructions for linguists who are less familiar with these practices. Please add more concrete details, examples, instructions as indicated above. –> The FOSIL tutorials should be mentioned much earlier whenever relevant, given that they do include more concrete information.

Following the reviewer's suggestions, we have included more concrete examples pertaining to linguistics throughout the revised manuscript, and we have referenced the FOSIL tutorials where relevant. The examples are too many to reproduce again here, but we are confident they are more "visible" in the updated version.

[RC 2.42.] Some references contain incomplete author lists –> all authors should be cited.

We have carefully scrutinized there reference section and made the appropriate changes.

3. Reviewer #3

[RC 3.1.] Since QRPs appear to be a critical component of the commentary here, I would spend at least a paragraph in the introduction explaining them / the original few papers on them and how this interwove with the time period of the 2015 collaboration to lead to TOP guidelines.

We thank the reviewer for this comment, which was common among the other reviewers as well. In the revised manuscript we have dedicated more space to discussing/explaining QRPs and incentive structure in academia. Particularly relevant for the reviewer's comment is the revised text we include below:

> Researchers have pointed to questionable research practices (QRPs), such as p-hacking–knowingly manipulating an analysis until a significant p-value is obtained (See Head et al., 2015)–and HARKing–hypothesizing after the results are known (See Murphy & Aguinis, 2019)–, along with small sample sizes, poor theory, lack of transparency, misguided incentive structure in academia, etc., as factors that ultimately led to the replication crisis, though it is likely that many factors are/were simultaneously at play. For instance, the aforementioned QRPs may be an unfortunate consequence of misaligned incentive structures in academia, where publication is the universal currency. The pervasive pressure to publish likely leads many researchers to focus on quantity over quality. Couple this with the difficulty of publishing negative or null results, and the result is a research landscape in which many fields suffer from publication bias with little or no incentive to prioritize time consuming open science practices. Taking this into account, it is not hard to understand why some researchers may turn to QRPs. While it is difficult to quantify how prevalent QRPs are in a given field, in a survey of applied linguists, Isbell et al. (2022) found that 94% reported having engaged in one more, and 17% admitted to having committed some form of fraud.
>
> In the aftermath of the aforementioned crisis, there has been a push for increased transparency and reproducible methodology to help mitigate the effects of QRPs. The clearest example of this is the Transparency and Openness Promotion Guidelines (TOP), author guidelines for journals that aim to help evaluate adherence to open science principles (See Nosek et al., 2015, as well as `https://www.cos.io/initiatives/top-guidelines`). The resulting methodological framework and associated techniques have reshaped research methods in Psychology, and, slowly but surely, are making their way into related fields. While many agree that open science practices represent a positive step forward in improving scientific rigor, these practices, by and large, have not been adopted in the field of linguistics (Bochynska et al., 2023). One reason for the slow adoption in linguistics may be related to the fact that engaging in open science is no trivial feat. On the contrary, it often requires learning new skills, thoughtful planning, as well as an openness and willingness to share materials, code, and data. Many researchers need to implement new techniques with limited pedagogical resources and embrace alternative methods of disseminating their research, all of which can constitute a steep learning curve. That being said, what engaging in open science ultimately entails is sure to be field-specific and vary accordingly. In some disciplines, for instance, it may only involve a few of the practices we outline in the present work without the need for innovative methodologies. Nonetheless, given how new open science practices are, it is reasonable to assume that current senior researchers were not trained in these innovative methodologies. As a consequence, many early career researchers (ECR) find themselves at a crossroads in which they are forced to learn open science on their own, often without institutional support. Ironically, there is also a growing expectation that ECRs implement these novel tools in order to be successful in their programs, on the job market, or to advance in their careers.

[RC 3.2.] I love this figure. Given the pervasiveness of TOP guidelines, maybe make the connection between them direct. I think you could comment on how those guidelines are heavily quant focused and ignore the intersection of the researcher and the research (and how positionality statements acknowledge these things).

Though this figure has changed significantly, we thank the reviewer for their praise nonetheless. We have also made the connection with the TOP guidelines more clear and heavily revised the positionality statement section.

[RC 3.3.] When you talk about repositories, mention that there are lists of "trusted repositories" that are better than researcher websites, are more accessible, etc.

In the revised manuscript we now include a table that highlights some of the more common data sharing platforms and highlight some of the key features.

[RC 3.4.] Organization: I would make the big headers each of your sections from the figure and then subheader the points in each.

We thank the reviewer for this excellent idea. We have made this organizational change to the revised manuscript.

[RC 3.5.] Maybe provide a link to tutorials that show people how to do literate programming? I think there are a few good markdown / rmarkdown / quarto ones that may be beneficial for further reading. You could also mention code ocean as options for researchers who don't want to recreate entire environments but want to test the code.

Oh here's code ocean ha. Maybe a touch earlier.

We again thank the reviewer for this practical and useful idea. We have provided examples of literate programming, highlighting this very manuscript as an example tutorial. We make the entire project available on the OSF, GitHub, as well as an online server instance via Code Ocean.

[RC 3.6.] I think the section on pre-reg should start with a quite note that they are not only for experimental studies, you can pre-reg ideas, etc. I think there's been a lot of push back against them because it doesn't feel like it fits in qual research or more exploratory studies.

We wholeheartedly agree with the reviewer's observation. In the revised text we have tried to emphasize that pre-registrations are adaptable to many subfields of linguistics by providing more concrete examples throughout this section.

[RC 3.7.] In the pre-print section, this is a good place for TOP as well – many of the guidelines encourage them, so that's a helpful signal.

It is not entirely clear to us how exactly the reviewer envisions promoting TOP in this section. Which of the guidelines encourage pre-prints? We use the OSF (and corresponding publication, Nosek et al., 2015) as our guide, but do not see a clear connection. We are happy to include this in a future revision of the manuscript once we have a better idea of what exactly the reviewer means.

References

Adetula, A., Forscher, P. S., Basnight-Brown, D., Azouaghe, S., & IJzerman, H. (2022). Psychology should generalize from–not just to–Africa. *Nature Reviews Psychology*, *1*(7), 370–371. https://doi.org/10.1038/s44159-022-00070-y

Berez-Kroeker, A. L., Gawne, L., Kung, S. S., Kelly, B. F., Heston, T., Holton, G., ... Woodbury, A. C. (2018). Reproducible research in linguistics: A position statement on data citation and attribution in our field. *Linguistics*, *56*(1), 1–18. https://doi.org/10.1515/ling-2017-0032

Berez-Kroeker, A. L., McDonnell, B., Koller, E., & Collister, L. B. (Eds.). (2022). *The open handbook of linguistic data management*. The MIT Press. https://doi.org/10.7551/mitpress/12200.001.0001

Bochynska, A., Keeble, L., Halfacre, C., Casillas, J. V., Champagne, I.-A., Chen, K., ... Roettger, T. B. (2023). Reproducible research practices and transparency across linguistics. *Glossa Psycholinguistics*, *2*(1, 18), 1–36. https://doi.org/10.5070/G6011239

Bucholtz, M., Campbell, E. W., Cevallos, T., Cruz, V., Fawcett, A. Z., Guerrero, B., ... Reyes Basurto, G. (2023). Researcher positionality in linguistics: Lessons from undergraduate experiences in community-centered collaborative research. *Language and Linguistics Compass*, *17*(4), 1–15. https://doi.org/10.1111/lnc3.12495

Casillas, J. V. (2021). Interlingual interactions elicit performance mismatches not "compromise" categories in early bilinguals: Evidence from meta-analysis and coronal stops. *Languages*, *6*(1), 9. https://doi.org/10.3390/languages6010009

Coretta, S., Casillas, J. V., Roessig, S., Franke, M., Ahn, B., Al-Hoorie, A. H., ... Roettger, T. B. (2023). Multidimensional signals and analytic flexibility: Estimating degrees of freedom in human-speech analyses. *Advances in Methods and Practices in Psychological Science*, *6*(3), 1–29. https://doi.org/10.1177/25152459231162567

Cristea, I.-A., & Ioannidis, J. P. A. (2018). Improving disclosure of financial conflicts of interest for research on psychosocial interventions. *JAMA Psychiatry*, *75*(6), 541–542. https://doi.org/10.1001/jamapsychiatry.2018.0382

Faytak, M., Kadavá, Š., Xu, C., Özsoy, O., Akumbu, PiusW., Cardoso, A., ... Roettger, T. B. (2024). *Big team science for language science: Opportunities and challenges*. Open Science Framework. Retrieved from osf.io/3pkj6

Gawne, L., & Styles, S. (2022). Situating linguistics in the social science data movement. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 9–25). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0006

Hardwicke, T. E., & Ioannidis, J. P. (2018). Populating the data ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PloS One*, *13*(8), 1–12. https://doi.org/10.1371/journal.pone.0201856

Hardwicke, T. E., Thibault, R. T., Kosie, J. E., Wallach, J. D., Kidwell, M. C., & Ioannidis, J. P. A. (2022). Estimating the prevalence of transparency and reproducibility-related research practices in psychology (2014–2017). *Perspectives on Psychological Science*, *17*(1), 239–251. https://doi.org/10.1177/1745691620979806

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), 1–10. https://doi.org/10.1098/rsos.190806

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), 1–15. https://doi.org/10.1371/journal.pbio.1002106

Holton, G., Leonard, W. Y., & Pulisifer, P. L. (2022). Indigenous peoples, ethics, and linguistic data. In A. L. Berez-Kroeker, B. McDonnell, E. Koller, & L. B. Collister (Eds.), *The open handbook of linguistic data management* (pp. 51–60). The MIT Press. https://doi.org/10.7551/mitpress/12200.003.0008

Hudley, A. H. C., Mallinson, C., & Bucholtz, M. (2020). Toward racial justice in linguistics: Interdisciplinary insights into theorizing race in the discipline and diversifying the profession. *Language*, *96*(4), e200–e235. https://doi.org/10.1353/lan.2020.0074

Isbell, D. R., Brown, D., Chen, M., Derrick, D. J., Ghanem, R., Arvizu, M. N. G., . . . Plonsky, L. (2022). Misconduct and questionable research practices: The ethics of quantitative data handling and reporting in applied linguistics. *The Modern Language Journal*, *106*(1), 172–195. https://doi.org/10.1111/modl.12760

Jafar, A. J. N. (2018). What is positionality and should it be expressed in quantitative studies? *Emergency Medicine Journal*, *35*(5), 323–324. https://doi.org/10.1136/emermed-2017-207158

Leonard, W. Y. (2021). Centering indigenous ways of knowing in collaborative language work. In L. Crowshow, I. Genee, M. Peddle, J. Smith, & C. Snoek (Eds.), *Sustaining indigenous languages: Connecting communities, teachers, and scholars* (pp. 21–34). Athabasca University Press.

Lewis, C. (2024). *Data management in large-scale education research*. CRC Press.

Massoud, M. F. (2022). The price of positionality: Assessing the benefits and burdens of self-identification in research methods. *Journal of Law and Society*, *49*, S64–S86. https://doi.org/10.1111/jols.12372

Mufwene, S. S. (2020). Decolonial linguistics as paradigm shift. In A. Deumert, A. Storch, & N. Shephard (Eds.), *Colonial and decolonial linguistics: Knowledges and epistemes* (pp. 289–300). Oxford University Press.

Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, *34*, 1–17. https://doi.org/10.1007/s10869-017-9524-7

Nagle, C., Baese-Berk, M., Amengual, M., & Casillas, J. V. (2024). *Sound communities: A quantitative proposal for studying bilingualism in context*. PsyArXiv. https://doi.org/10.31234/osf.io/m67tx

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., et al.others. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. https://doi.org/10.1126/science.aab237

Quintana, D. S. (2020). A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, *9*, 1–12. https://doi.org/10.7554/eLife.53275

Roberts, S. O., Bareket-Shavit, C., Dollins, F. A., Goldie, P. D., & Mortenson, E. (2020). Racial inequality in psychological research: Trends of the past and recommendations for the future. *Perspectives on Psychological Science*, *15*(6), 1295–1309. https://doi.org/10.1177/1745691620927

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

Singh, L., Killen, M., & Smetana, J. G. (2023). Global science requires greater equity, diversity, and cultural precision. *APS Observer*, *36*. Retrieved from https://www.psychologicalscience.org/observer/gs-equity-diversity-cultural-precision

Steltenpohl, C., Hudson, S., & Klement, K. (2022). How to begin writing a positionality statement. Retrieved from https://vimeo.com/675236573/741e24aab7

Stieglitz, S., Wilms, K., Mirbabaie, M., Hofeditz, L., Brenger, B., L'opez, A., & Rehwald, S. (2020). When are researchers willing to share their data?–Impacts of values and uncertainty on open data in academia. *PLoS One*, *15*(7), 1–20. https://doi.org/10.1371/journal.pone.0234172

Tsikewa, A. (2021). Reimagining the current praxis of field linguistics training: Decolonial considerations. *Language*, *97*(4), e293–e319. https://doi.org/10.1353/lan.2021.0072

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Aaert, R. C. van, & Assen, M. A. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid *p*-hacking. *Frontiers in Psychology*, *7*, 1–12. https://doi.org/10.3389/fpsyg.2016.01832

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, *41*(e120), 1–61. https://doi.org/10.1017/S0140525X17001972