Opening open science to all:

Demystifying reproducibility and transparency practices in linguistic research

Joseph V. Casillas[1], Gabriela Constantin-Dureci[1], Iván Andreu Rascón[1], Jiawei Shao[1], Stephanie A. Rodríguez[1], Adrija Gadamsetty[1], Alexandria Minetti[1], Krishita Laungani[1], John Thatcher[1], Rhode-Taina Gardere[1], Katherine Taveras[1], Isabelle Chang[1], Nicole Rodríguez[1], & Kyle Parrish[2]

Rutgers University

Goethe University Frankfurt

Author note

Correspondence concerning this article should be addressed to Joseph V. Casillas, Rutgers University - Department of Spanish and Portuguese, 15 Seminary Place, New Brunswick, NJ 08904, USA . E-mail: joseph.casillas@rutgers.edu.

Abstract

In recent years, numerous fields of research have seen a push for increased reproducibility and transparency practices. 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/ERCs 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: 7,441

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.

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 generated an uproar in the psychological sciences. The alarming findings garnered media

[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).

attention (e.g., Oliver, 2016) and have led to periods of introspection and self-reflection in many adjacent fields, among them, linguistics.

Researchers have pointed to questionable research practices (QRPs), such as p-hacking and HARKing, along with small sample sizes, poor theory, lack of transparency, etc. as factors that ultimately led to the replication crisis, though it is likely that many factors are/were simultaneously at play. 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 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. To wit, it requires learning new skills, thoughtful planning, as well as an openness and willingness to share materials, code, and data. 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. 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 (ERC) 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/careers.

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/ERCs to

senior researchers and advisors. To this end, we identify three areas, stance, workflow, and dissemination, in which linguists can engage in open science. 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. In total, we describe eight open science practices, illustrated in Figure 1, within these areas: positionality statements and declarations of conflict of interest, open data, reproducible code/projects and literate programming, 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.

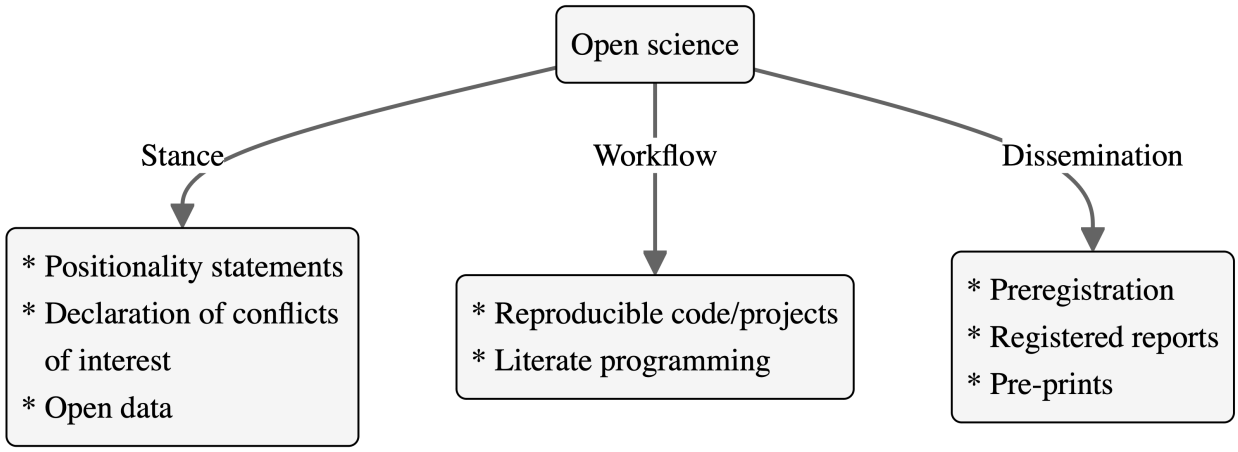

Figure 1: Eight open science practices amenable to research in lingusitics.

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 have been adopted in some disciplines of the humanities and social sciences as a way of 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). Positionality statements are increasingly considered 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. We believe they should be considered equally important in quantitative research, since, aside from contributing to ongoing efforts to promote transparency and openness in research practices, recognizing and addressing a researcher's positionality can increase the validity of their findings (Jafar, 2018). 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).

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). Nonetheless, others contest this practice, pointing 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, Casey, McBrayer, & Schwerdtle, 2023). Savolainen et al. (2023) draw connections between positionality statements and conflict of interest statements, 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). Importantly, statements of conflict of interest are underused in linguistic research. 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). Positionality statements are likely even less common.

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." Lastly, in showing their commitment to Diversity, Equity, Inclusivity, and Belonging (DEIB) initiatives, journals have

started to encourage authors to include positionality statements with their submissions (e.g., the Journal of Social and Personal Relationships). It is our stance that researchers should reflect on their positionality before starting a project and consider including a positionality statement. When submitting a study for publication, the positionality statement can be included in additional materials if the word limit is a concern. For examples of positionality statements in linguistic research, the interested reader is directed to Bochynska et al. (2023) and Weissler et al. (2023).

In sum, we believe positionality statements are essential in linguistic research as they promote critical self-reflection, increase transparency, and help address diversity and inclusion concerns. Furthermore, including positionality statements in quantitative research can increase the validity of findings. By reflecting on who it is that does the research, linguistics can become a more diverse, inclusive, and transparent field.

Open data

The term *open data* 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. In spite of such assurances, we now know they do not typically result in adequate sharing of research materials (Spellman, Gilbert, & Corker, 2017; Wicherts, Borsboom, Kats, & Molenaar, 2006). Recent efforts have attempted to encourage researchers to make linguistic data open and accessible via servers (e.g., the IRIS database). Some journals offer open science badges (e.g., *Language Learning*, *Language and Speech*), and, in some cases, have made data sharing a requirement for publication (e.g., *Applied Psycholinguistics*). Nonetheless, open data 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 open data and consider the specific challenges researchers face in the field of linguistics.

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 any concerns. Making linguistic data freely available improves credibility in our findings, to other researchers and the general public. Prohibiting or impeding access to data collected for publicly funded research is unethical and 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 [TIMO]. 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 (e.g., Casillas, 2021). 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 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). That is to say, the majority of linguistic research is concentrated on specific language, mainly Indo-Germanic, in overrepresented communities, by privileged scholars. 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.

Having stated all the above, it is necessary to recognize that the field of linguistics faces unique challenges with regard to open data. At the heart of these challenges are ethical concerns that must be considered with care. The privacy and consent of participants must be safeguarded. Linguistic data often include personal information, which can be especially difficult to anonymize. Though written and behavioral data do not typically pose many issues, audio and video recordings constitute a large portion of linguistic research materials. In such cases, anonymizing participant information can be troublesome, particularly when working with minority languages or smaller communities. Sociolinguistic interviews, for instance, represent a substantial and valuable contribution to linguistics and often contain sensitive information. 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. 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, 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).

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. Researchers should include a codebook or README file explaining the metadata of the file(s), as well as an explanation of the variables it contains and the context in which it was obtained. Once the data has been prepared for sharing, the researcher must decide where to share it. 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 data sets (and other 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.

To summarize, open data is important because it facilitates transparency, rigor, reproducibility, replication, accumulation of knowledge, and, importantly, it makes participating in the scientific endeavor more inclusive. Linguistics, in general, does not engage in open science practices (Bochynska et al., 2023), including sharing data. 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.

Reproducible code/projects and literate programming

As we have seen, reproducibility is now 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. At worst, a lack of reproducibility can lead to irreproducible results and wasted resources. This can have serious implications for public health and policy decisions based on research findings. In order to ensure reproducibility, it is necessary to be transparent about the methods used in research. This 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 many researchers are beginning to take steps to improve the reproducibility of their 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 R markdown files (Rmd) and Quarto markdown files (qmd). These formats use the R package `knitr` (Xie, 2014, 2015, 2023) to tangle (also "knit" or "render") the output file. In fact, the present manuscript was generated using literate programming and is available for download here: https://osf.io/bsu2q/.

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. This allows a researcher to 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. Examples of completed reproducible projects are available to the interested reader here: https://osf.io/un45x/ and https://osf.io/cp9bs/.

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. 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 and Code Ocean can also be used to create 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 replicated by anybody with an internet connection.

At this juncture it is important to note that there is no way to completely future-proof code or projects. Researchers must continually strive to maintain the reproducibility of their work. 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.

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. 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.

Preregistration and registered reports

In this section, we will briefly consider two 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 and registered reports, 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. 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 corpus analysis, conversation/discourse analysis, experimental research, and more. However, as highlighted by Roettger (2021), 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 conducting experiments, as well as analyzing 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 (e.g., Coretta et al., 2023). For instance, 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. These choices 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. 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.

It is important to acknowledge that 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 run, 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. We provide examples of preregistrations at the following links: https://osf.io/nprgz and https://osf.io/qvjzy.

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 data, analyze data, interpret results, and submit their findings for publication. 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 statistical technique that combines the results of multiple studies to increase the power of analysis. 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. 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 2 provides a side-by-side comparison of the standard publishing model and RRs.

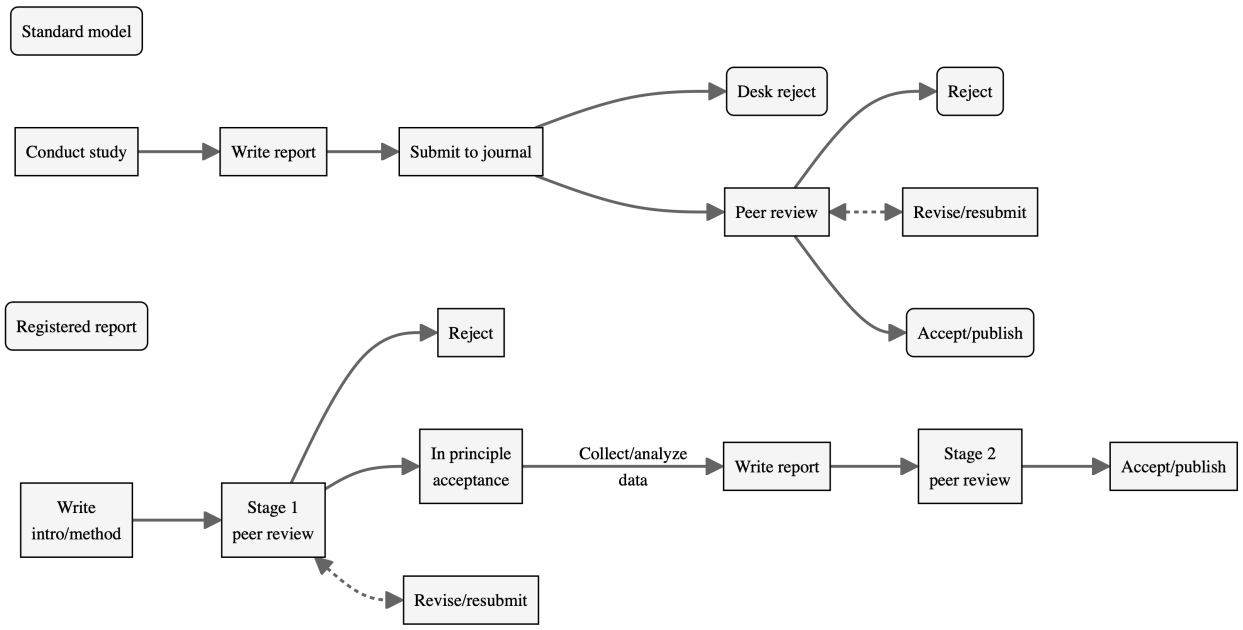

Figure 2: Flow chart of the standard publication model and registered reports.

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. Preregistration is often confused with RR, 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.

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. While pre-prints accelerate the dissemination of research, it is vital to remember that this process does not always lead to journal publication. This can occur for several reasons, such as authors may choose not to pursue this route, or the research may be intended for different dissemination avenues (Ettinger et al., 2022). 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.

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 extends support to the authors, but this process also supports reviewers, journals and publishers, and the reader audience. This inclusive process allows more researchers and reviewers to participate in discussing the research findings and reduces the need for repeated rounds of re-review or extensive revisions.

One of the most significant benefits of pre-prints is their “early view” and “open access” effect, which leads to more attention from readers, increasing visibility and development of the overall research (Das Biswas & Biswas, 2023). Pre-prints are not just quick dissemination; they also promote transparency and reproducibility in scientific research. 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. 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.

Despite these 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.

Additional concerns about pre-prints include their ability to secure the steady resources (technologies, expertise, policies, visions, standards, and so on) required to maintain and enhance the value of a service based on a user community's needs (Rieger, 2012). Preprints emerged as a 'public good' and pre-print platforms provide a free service to both authors and readers; at the same time, many of the existing pre-print services lack a scalable and transparent business model. Moreover, data show that more senior researchers had more experience sharing through the format of pre-print than PhD students and ECRs in the 0-4 years group (Liu & De Cat, 2021). Nonetheless, the data collected from these surveys showed positive attitudes and willingness to contribute to open science through the submission of pre-prints. These findings encourage ECRs to submit their research for pre-prints to receive valuable feedback from established scholars in their field of study and to increase visibility of their research, particularly in rapidly evolving fields.

Researchers interested in making a pre-print publicly available can follow these simple steps. First, select a pre-print server that aligns with the course of research. Next, all pre-prints 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, encouraging others to comment and share.

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. In the present work we have outlined eight specific open science practices, classified into three areas, that researchers in linguistics can adopt to make their research more open, transparent, inclusive, and accessible to a wider audience.

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. This necessarily implies providing a framework for senior researchers to learn open science, if we intend for them to train the next generation of linguists. Furthermore, linguistic journals must to adapt to new models of publishing.

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; Lewis, 2020). Crucially, the purpose of this article is to help foster open science in linguistics (FOSIL). We leave the interested reader with a series of tutorials designed for linguists: https://FOSIL-project.github.io. The FOSIL project details many of the open, transparent practices described here with crowd sourced translations into various languages other than English with the goal of

making open science practices clear and accessible to all individuals conducting research in the field of linguistics.

References

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/10.1177/1468794112468

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., et al.others. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, *351*(6280), 1433–1436. https://doi.org/10.1126/science.aaf091

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

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

Das Biswas, M., & Biswas, A. (2023). Open access to scholarly communication through preprints: Accelerating sustainable development in education. In D. Coghlan & M. Brydon-Miller (Eds.), *Digital libraries: Sustainable development in education* (pp. 525–545). National Digital Library of India, IIT Kharagpur.

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

Ettinger, C. L., Sadanandappa, M. K., Görgülü, K., Coghlan, K. L., Hallenbeck, K. K., & Puebla, I. (2022). A guide to preprinting for early-career researchers. *Biology Open*, *11*(7), 1–8. https://doi.org/10.1242/bio.059310

FORRT. (2021). Reproducibility crisis (a.k.a. Replicability or replication crisis). Retrieved from https://forrt.org/glossary/reproducibility-crisis-aka-replicab/

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

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

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

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

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.

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

Rieger, O. Y. (2012). Sustainability: Scholarly repository as an enterprise. *Bulletin of the American Society for Information Science and Technology*, *39*(1), 27–31. https://doi.org/10.1002/bult.2012.1720390110

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

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

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

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

Xie, Y. (2014). Knitr: A comprehensive tool for reproducible research in R. In V. Stodden, F. Leisch, & R. D. Peng (Eds.), *Implementing reproducible computational research*. Chapman; Hall/CRC.

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/