Sharing the Recipe: Reproducibility and Replicability in Research Across Disciplines

Rima-Maria Rahal[‡], Hanjo Hamann[§,‡,|], Hilmar Brohmer[¶], Florian Pethig[#]

‡ Max Planck Institute for Research on Collective Goods, Bonn, Germany
§ EBS Law School, Wiesbaden, Germany
| Institute for Globally Distributed Open Research and Education, IGDORE, Germany
¶ Institute of Psychology, University of Graz, Graz, Austria
# University of Mannheim, Mannheim, Germany

Corresponding author: Hanjo Hamann (hanjo.hamann@ebs.edu)

Academic editor: Tamara Heck

Abstract

The open and transparent documentation of scientific processes has been established as a core antecedent of free knowledge. This also holds for generating robust insights in the scope of research projects. To convince academic peers and the public, the research process must be understandable and retraceable (*reproducible*), and repeatable (*replicable*) by others, precluding the inclusion of fluke findings into the canon of insights. In this contribution, we outline what reproducibility and replicability (R&R) could mean in the scope of different disciplines and traditions of research and which significance R&R has for generating insights in these fields. We draw on projects conducted in the scope of the Wikimedia "Open Science Fellows Program" (*Fellowship Freies Wissen*), an interdisciplinary, long-running funding scheme for projects contributing to open research practices. We identify twelve implemented projects from different disciplines which primarily focused on R&R, and multiple additional projects also touching on R&R. From these projects, we identify patterns and synthesize them into a roadmap of how research projects can achieve R&R across different disciplines. We further outline the ground covered by these projects and propose ways forward.

Keywords

reproducibility, replicability, interdisciplinary, open science practices

Introduction

In the quest to gain knowledge and advance scientific discovery, the roles of openness, transparency and free knowledge are increasingly being recognized (United Nations Educational, Scientific and Cultural Organization 2021; Arqus Alliance 2022). In part, the

value of openness in research lies in making accessible to others how this research was conducted, and how data and method insights were derived. Almost like sharing a recipe, this principle of communality (Merton 1942, Anderson et al. 2016) makes it possible for others – be it the public or one's academic peers – to retrace (reproduce) the steps involved in a particular research project, and even to repeat (replicate) them, thereby generating new data. However, the specifics of these steps vary across academic disciplines due to their epistemic diversity. Here, we aim at carving out and depicting these differences based on several research projects conducted within Wikimedia's "Open Science Fellows Program" (Fellowship Freies Wissen).

In philosophy of science and research, replicability and reproducibility (R&R) are discussed as central criteria for robust knowledge (Chambers 2017), despite some criticism of this principle (Stroebe and Strack 2014). Although the spirit of R&R permeates research throughout different disciplines, how R&R are defined and, consequently, the practices of implementing R&R, differ widely (Goodman et al. 2016; Plesser 2018). In this study, we use the term *replicability* to describe the ability to repeat a piece of scholarship by creating new data (with the same or similar materials) and obtaining the same results as the original piece (Nosek et al. 2022). We use the term *reproducibility* to indicate being able to understand how a piece of scholarship has come to a specific conclusion (Nosek et al. 2022). In this sense, reproducibility describes being able to retrace the process of generating insights. Research aims to formulate theories that allow deriving predictions with lawlike characteristics, holding up to long-running attempts of corroboration and falsification (Hempel 1968; Hempel and Oppenheim 1948): To judge the merit and verisimilitude of a theory, it needs to be put to the test repeatedly. In other words, research is considered a credible basis for robust knowledge (only) to the extent that it can be done again (Nosek et al. 2012; Nosek and Errington 2020a).

How these general principles of R&R are defined, discussed, and implemented, however, varies widely between different research disciplines and schools of thought (for an overview, see Goodman et al. 2016; Plesser 2018). This can be summarized under the term of *epistemic diversity,* which has been broadly discussed in the Philosophy of Science (Devezer et al. 2019, Solomon 2012). Here, we aim to bring together perspectives of the life science, natural sciences, social sciences, and the humanities, by giving an overview of a diverse set of research and scholarly projects designed by early-career researchers to contribute to making research more reproducible and replicable. We describe the projects' aims and scopes, and compare how different projects approached the topics of R&R, how they (explicitly or implicitly) defined R&R, and we synthesize these approaches. These findings are discussed in the light of future projects that could make use of and advance R&R.

Challenges for Reproducibility and Replicability across Disciplines

Although R&R are often argued to be important features of research processes to generate *robust* knowledge, it is not always clear whether and how R&R can be achieved. Often, practical obstacles - such as psychological insecurity or licensing uncertainties (Truan and Dressel 2021a) - can make R&R efforts cumbersome. Of central relevance for our discussion of the cross-disciplinary research we reviewed, however, was dealing with two broader, conceptual issues. In the following, we will focus on these two challenges to R&R during data collection and interpretation, as they arise in different research disciplines and traditions.*[4]

Challenges for R&R during Data Collection

Creating new data can sometimes be impossible. For instance, when a lawyer interprets a piece of legislation, reproducibility would – optimally – mean that novel pieces of legislation on the same topic are interpreted while using the same interpretation method or model by different investigators. However, the investigators cannot create a new piece of legislation (the legislator would have to do that), but are limited to reinterpreting the same piece as in the original investigation. However, in this setup replicability could mean that other investigators interpret the same piece and test if they arrive at the same conclusion as the original investigator, which, according to our definition above, would rather count as a reproduction attempt. The original investigator could also attempt to reinterpret the same piece and test if they arrive at the same conclusion as in their first attempt. Again, according to our definition above, we would rather construe this procedure as a reproduction attempt.

When a sociologist interviews people on a certain topic to generate new data for a study, replicability would – optimally – mean that additional interviewees are questioned on the same topic using the same questions and in the same setup, by different investigators. However, a number of possible challenges for replicability could arise: The interviewees may differ in certain characteristics from those originally questioned, because of differences in the sampling strategy (Gilbert et al. 2016). The interviewers themselves may have an effect on the types of answer the interviewees give, even if the questions are standardized. More broadly, it may be impossible to recreate the original interview situation (for instance, because time has elapsed and interviewees see the topic in a different light now than they did at the time of the original investigation).

Some argue that no such thing as a exact replication, i.e., precisely repeating a previous piece of scholarship, exists “because there are always differences between the original study and the replication [...] (like small differences in reagents or the execution of experimental protocols). As a consequence, repeating the methodology does not mean

an exact replication, but rather the repetition of what is presumed to matter for obtaining the original result." (Nosek and Errington 2017, p. 1).

However, even if exact replication is not possible, *close or direct replication* (Brandt et al. 2014) attempts may prove sufficient, when there is no foreseeable reason why the results of the replication attempt should deviate from the original study. For example, evidence from the social sciences suggests that deviating from the mode of data collection in the original study (lab vs. online) had only little influence on replication success (Klein et al. 2014). While some claim that the contextual dependence of the studied effect determines its replication likelihood (Van Bavel et al. 2016; but see Inbar (2016) for a critical discussion), others argue that, if contextual dependence is so dominant that an effect cannot be replicated, its merit for theoretical advancement in science generally is questionable (also see Zwaan et al. (2017)).

Challenges for R&R during data interpretation

Further, even if new data was generated in a manner that resembled exactly that of the original investigation, this data must be interpreted. This process of interpreting the data can be more or less objective (Gunton et al. 2021; Reiss and Sprenger 2020), but different people may interpret the same data differently, and because even the same person may interpret the same data differently at different points in time, restrictions for replicability follow. Therefore, arguably, the views of the person interpreting the data, or the zeitgeist and methods available when the interpretation takes place, can play an enormous role in how the data is handled, potentially rendering attempts to reproduce or replicate futile (also see Feest (2019)). In the social sciences, however, there is some evidence that specifics of the research teams who attempted the replication study (Open Science Collaboration 2015) mattered little for whether a finding could be replicated successfully.

In a similar vein, comparing the results of a reproduction or replication attempt to the findings of the original investigation, and defining whether the attempt was successful in showing the same result, is subject to interpretation. Reproducibility means being able to retrace how the original finding was achieved. But what if reproduction attempts are only partly successful, or if reproduction teams disagree with the methodological decisions of the original author? Such ambiguities may pose a challenge to cumulative scientific efforts, as it is unclear if the original and the replication finding should be treated separately or combined into meta-analytic evidence (Armbruster 2021; Mathur and VanderWeele 2019; Muradchanian et al. 2021).

Interpreting what the absence of R&R success means is not straightforward either. Failing to reproduce or replicate could mean that the original finding under scrutiny was not representative for a real effect. – but not necessarily so, because of the multitude of things that could stand in the way between R&R success (see discussion above). For empirical research, for instance, because of regressive shrinkage in larger samples and measurement uncertainties, the effect sizes (e.g., the difference between an intervention group and a control group) are often expected to be much smaller in replication attempts

than they had been in the original studies (Fiedler and Prager 2018; Maxwell et al. 2015; Patil et al. 2016). And precisely because many effects are potentially small in reality, there is a high chance that they are missed in replication attempts, when they fail to reach a large enough sample. In conclusion, it may take more than one attempt to test R&R of a finding, and even then it remains difficult to dismiss the merit and verisimilitude of the original finding. Only if accumulating evidence lets the original finding seem unlikely or even as an exaggerated claim should one reconsider its contribution.

Method

Recognizing the discipline-specific difference in how R&R are defined and addressed, we sought to obtain an overview of how recent scholarly work treated R&R. To so do, we drew on projects conducted within Wikimedia's "Open Science Fellows Program" (*Fellowship Freies Wissen, which we will refer to as "Wikimedia Fellowship"*), an interdisciplinary, long-running funding scheme for projects contributing to open research practices in Germany, Switzerland and Austria (Behrens et al. 2022). A total of 90 projects were funded through this scheme following a review procedure, comprising a substantive body of work on Open Science proposed by early-career researchers. Although each project addressed a different topic of Open Science and Scholarship, covering a broad range of methods and specific goals, the projects shared the general aim to advance Open Science.

Twelve of these projects mentioned R&R explicitly in their project titles or descriptions, which implied that they either conducted a replication or tried to enhance R&R in their specific domain by, for instance, improving infrastructure (for a complete list, see Table 1). We used these projects, which stemmed from a range of academic disciplines, as the basis for a qualitative review on how researchers think about R&R, and how their work advances the principles of R&R .

For each of of these 12 projects, we contacted the person who received funding through the fellowship, asking for project-specific publications or deliverables. Where available, these publications were analyzed here, and supplemented using each project's documentation on dedicated Wikiversity pages required by the funder.

For each of the projects, one of the present authors read the existing documentation, created a short summary, and coded basic project characteristics. The dataset with our codings has been provided as a supplement to this publication. Specifically, we coded the primary research area (humanities, engineering, life science, natural sciences, or social sciences), whether the projects produced infrastructure to advance R&R, and whether they tested R&R empirically. Following this initial assessment, we defined three broad categories of projects focusing on similar aspects of R&R:

1. opening research processes by providing infrastructure,
2. improving methods and data, standardizing knowledge, and
3. making knowledge accessible through education and science communication.

Finally, we synthesized from the project materials how R&R were defined or conceptualized therein; and which challenges for R&R were mentioned.

Categorizing the Projects

As described previously, we derived three broad categories to organize the projects along their contributions to R&R. In the following, we will briefly introduce the three categories along with the projects. An overview of the projects is presented in Table 1 .

Opening Research Processes by Providing Infrastructure

The four projects in this category aimed at providing guidance and practical solutions to the challenge of reproducing data and analyses of research projects within fields where large amounts of data and sophisticated (pre-)processing are common. These projects were motivated by the fact that, for example, code may not be easily linked to the output reported in the paper (especially after some time has passed), or computational environments lacked certain dependencies (i.e., software libraries) that were used in the original analyses. In the following, we will briefly summarize each project.

Felix Hoffmann's project "Code, Data and Reproducibility - Open Computational Research" was designed to facilitate research publications in accordance with three important criteria:

1. documentation of data and code,
2. documentation of software libraries used, and
3. provisioning of a computational environment.

The outcome of this project is a hands-on guide to combining Docker and Sumatra to achieve the aforementioned goals. Future work points to the application of this practical approach to Hoffman's own research on computational neuroscience.

Ludmilla Figueiredo's project "Computational Notebooks as a Tool for Productivity, Transparency, and Reproducibility" provides a starter-kit for computational notebooks so that calculations performed as part of a paper can be traced and understood by others (Figueiredo et al. 2022). Similar to the previous project, the author of this project focuses on providing a workflow that will easily allow researchers to structure their work to improve its reproducibility. In contrast to the previous project, this project employs the tool Jupyter Notebook and cites the advantage of combining "descriptive text, as well as code and its outputs, in a single, dynamic and visually appealing file." Future work aims at implementing the workflow in the author's work on biodiversity.

Jana Lasser's project "Executable papers: Werkzeug für mehr Reproduzierbarkeit und Transparenz in den Naturwissenschaften" (Executable papers: Tools for more reproducibility and transparency in the natural sciences) taps into a similar problem and develops an executable paper, i.e., "dynamic pieces of software that combine text, raw data, and the code for work" (Lasser 2020 p. 1), on pattern formation in salt deserts. The

author finds that there “is currently not much to build on” in terms of how executable papers should be set up. In turn, the author develops their own approach to an executable paper, also using Jupyter Notebook. One major outcome of this project is a journal article that documents the process and challenges of developing executable papers.

Hans Henning Stutz’ project “The Glass Tool - from its development to its usage and to research data” tackles the issue of widespread “dark data” in the field of geotechnical engineering stemming from unique tools that produce intransparent data. He develops an experimental device to determine soil structure and resistance, making his construction drawings and monitoring software openly available. This open-method approach may enable other researchers to build on his solution and improve it in the future.

Improving Methods and Data and Standardizing Knowledge

The four projects in this category have in common that previous work in their respective domains might have lacked scrutiny and best practices to draw general conclusions. Hence, these projects aim at improvements in terms of methods or data quality by ensuring that knowledge generated from new data will be standardized and more reliable.

Charlotte Oertel’s project “Acceleration of quality in the humanities” developed a case study of how flawed art-historical analysis may propagate and get reinforced through subsequent citations. As an attempt to bar such forward-propagation, the author developed and tested an approach she called “citation genealogy analysis” (Thiery and Oertel 2021): By reconstructing a “complete bibliography of an exemplary argument” and “presenting all bibliographical data online”, she sought to enable researchers to trace “citation lines from modern publications back through referenced sources”, thereby ensuring that misinformation ultimately unsupported by evidence would get weeded out.

In his project on the “Effects of Generic Masculine and Its Gender-fair Alternatives”, *Hilmar Brohmer**[3] is trying to replicate a classic social-psychological experiment, where it was shown that gender-fair language prompts people to think more about women compared to when the generic masculine form is used. This project is conducted as a multi-lab study, which has the advantage that the same experiment will be conducted several times, enhancing explanatory and statistical power. Theory and methods were preregistered and peer-reviewed before data collection started.

Richard Höchenberger’s project “Quick estimation of taste sensitivity” collects data meant “to be of larger practical use for clinical diagnostics”. To this end, they aimed at data standardization in order to build a “norm database” based on “measurements of healthy participants, i.e. a large set of reference data.” In order to construct such a reference database, the project implemented a method that will enable researchers to collaboratively collect and share data, namely deploying software tools that allow

“researchers from different institutions to work collaboratively very easily”. A publication is forthcoming, but was not available at the time of writing (D'Alessandro et al. in press).

Florian Pethig’s*[3] project “Data Version Control: Best Practice for Reproducible, Shareable Science?” explores the issue of version control of intermediate datasets that precede the dataset for the final analysis (e.g., as is common for the pre-processing of natural language). He argues that these pre-processing steps are often not properly documented in research papers and analyzes the status quo by conducting a non-representative survey to understand data versioning practices of other researchers. Finally, he discusses the tool DVC as one such way to track changes even for larger datasets.

Making Knowledge Accessible Through Education and Science Communication

Beside their main goal of making research more transparent for peers within the field (see Robson et al. (2021)), open science and R&R may also aim at making knowledge accessible to people in other fields and even to people outside of academia. To this end, the following projects either communicated understandable research output in innovative ways (i.e., they went beyond a purely scientific publication) or promoted scientific education for both people in and outside the field.

Ruben Arslan’s project “Reproducible websites for everyone” was special, as he faced the issue of making open-science practices compatible with ethical and legal standards: his project’s data contained sensitive information on Swedish men’s reproductive behavior and offspring over time. As data sharing was not an option in this context, he worked on a solution to make the results available on a reproducible website, which he and his collaborators launched in 2017 (see Arslan (2017) and Arslan et al. (2017)). Moreover, in interaction with different software (statistical computer language R) and online repositories (Git, GitHub, and Zenodo), Arslan provided a tutorial for other researchers who face similar issues.

Nate Breznau’*s* project “Giving the Results of Crowdsourced Research Back to the Crowd” also made use of a reproducible website. Together with a team of many independent researchers, he analyzed the same data with the same underlying hypothesis: Does immigration undermine citizens’ support for social policies? The results differed a lot throughout the labs, depending on their analysis strategies. To make the results of this multi-lab project accessible to the public, Breznau created a reproducible website, which contained dynamic figures and graphs of the key findings.

Rima-Maria Rahal’s*[3] project “Reproduzierbare Forschung durch offene und transparente Wissenschaft” (Reproducible practices make open and transparent research) developed an online course on methodological foundations of scientific experimentation, integrating Open Science practices (Rahal 2020) to “enable not just students to experiment independently and openly, but also to convey to the general public an elementary understanding of these methods”. Specifically, the course

advocated that students do not rely on just a single experimental finding, because there's a chance it was just a fluke determined entirely by chance. Rahal would ask her students whether they thought a repetition of this experiment would yield the same finding. This is what she called "replicability", referring to "law-like characteristics derived from their long-run frequency of corroboration", but noting that "one-time failure to replicate does not mean that we can be sure our initial finding was a fluke".

Naomi Truan's project "Digitale Daten — meine, deine, unsere?" (Digital data – mine, yours, ours?) was designed as a didactic intervention within a research-based linguistics seminar on "Grammar in the Digital Age". The author and a colleague assigned creative tasks in the course of two iterations of this class and surveyed their students, *inter alia*, on their willingness to publish academic posters in Open Access and getting taught through Open Educational Resources (OER). Their data show that 12 out of 15 student groups were willing to share their posters and were motivated by feeling included within a "community of practice" even outside the course (Truan and Dressel 2021b: 389). The authors specifically determined "that key motivators are a sense of belonging, personal reward, and an active contribution to a culture of collaboration, whereas apprehensions are grounded in concerns about the quality of their work, uncertainties about licensing, and fear of vulnerability through visibility." (Truan and Dressel 2021a).

Reproducibility and Replicability in the Selected Projects

Reproducibility

Especially in the quantitative scientific projects, reproducibility often means reusing the materials and data of a study and being able to recreate the results of the original study. In psychology and the social sciences, quantitative studies are made reproducible by a transparent source code that is compatible with the data at hand and produces the same statistical results as those reported in the original manuscript. Researchers in these disciplines frequently used programming languages, such as R (R Core Team 2020) or Python (Van Rossum and Drake Jr. 1995) to preprocess their raw data. However, the understanding of such a code requires a lot of expertise. Thus, several projects, such as Figueiredo's, aimed at making papers reproducible even for non-experienced outsiders by combining code and explanations in one document. In a similar vein, the data preprocessing steps require transparency, as interested researchers may want to learn from these methods. The use of version control - as investigated in Pethig's project - is a helpful tool, as interim pre-processing steps will not get lost. If a proper documentation of these interim steps is achieved, reproducibility of complex methods and data analyses is possible.

By contrast, Arslan and Breznau aimed to address this problem in a somewhat different fashion: Arslan recognized that sharing data of his project brought about ethical issues (see section on challenges). As sharing methods and data was not possible, his reproducible website draws anonymized data and code from different repositories and

presents the results in figures accompanied by textboxes, making them easily understandable. Moreover, Arslan provided a transparent workflow, which partly constitutes an infrastructure for other researchers to achieve such undertakings. Breznau and his collaborators have taken this idea one step further: First, they showed how several researchers achieved different results for a study, using the same data (hence, highlighting that results may differ if the original statistical code is not shared). Importantly, this multiverse of results was then also presented on a reproducible website, making it accessible for everyone. On this website, they not only presented the aggregated results, but also demonstrated how results change, based on decisions of individual labs. Taken together, both research projects included a dimension of accessibility for making a study reproducible for everyone, thereby communicating results effectively to outsiders.

In the context of qualitative scholarship and the humanities, reproducibility followed a similar reasoning, in that reproduction attempts retrace the path by which the original insight was achieved. This can be achieved by attempting to follow the logic described by the original investigators, for instance with regard to their arguments or interpretation of the data. As one example from our corpus, Oertel's project presents a case study of an instance where misleading to erroneous interpretations became accepted wisdom because the discipline (in this case, art history) proceeds citation-by-citation, continually building on earlier work. With Oertel's online tool, researchers should be able to trace citation lines from modern publications back through referenced sources. This approach, which the author describes as "citation genealogy analysis", bears a ready resemblance to reproducibility as applied to humanities research.*[1]

Replicability

In the quantitative sciences, replicability often means running a new experiment, which generates new data either using the same materials (e.g., instructions, hardware, software) as the original study or novel materials.*[2] However, only if most of these parameters are similar to an original study could this new study qualify as a close replication (see Challenges for R&R above), which can directly be compared to similar findings in meta-analyses. In this sense, a main part of Brohmer's project can be seen as a close replication of an older study by Stahlberg et al. (2001), despite notable differences: for instance, the original study was conducted in the lab via paper and pencil utilizing a sample of students, whereas the replication is done online with different convenience samples as part of multi-lab setting. However, it still qualifies as a close replication, as the materials were closely modeled after the original study and approved by the original authors (see also Nosek and Errington (2020b)).

In the realm of engineering, replicating previous study results might not be as crucial as in the more basic social-scientific research. Instead, the standardization of data output and results is important to achieve comparability. As a lot of engineering tools and devices produce "dark data", which is data stemming from intransparent internal processes, Stutz wants to avoid dark data for future geotechnical engineering projects by

providing construction drawings, code, and a comprehensive documentation for his soil structure tools. Hence, he provides an infrastructure for generating data transparently, which other researchers in the field can profit from in the future.

In the scope of qualitative scholarship and the humanities, replicability follows a similar reasoning, in that new data is generated to assess if, based on this new data, the original insights can again be obtained (Peels 2019). However, the data elicitation process (e.g., sourcing data from interviews, qualitative text analyses, and interpretation or situational observations) is often more situated in the context of the original investigation. An example of replicability in qualitative scholarship was not present among the selected projects, which may highlight an important avenue for future research.

Education as a Prerequisite for R&R

Raising awareness for the issue of R&R is crucial - especially during undergraduate education - because this is when potential future researchers are exposed to scientific practices for the first time. As, for instance, in Rahal's online courses, understanding the importance of replicability of a finding in new studies can enhance students' critical reflection about individual studies. This critical reflection may be accompanied by emphasizing the importance of open-science practices in comparison to questionable research practices (QRPs), which many older publications may suffer from.

Likewise, the reproducibility of previous findings is equally important. For instance, taking openly available results from Breznau's or Arslan's project can be a valuable starting point for students in methods classes trying to reproduce other findings, where data and code are available. Thanks to available infrastructure and software, teaching can also involve handling online repositories (e.g., Zenodo, GitHub), version control (e.g., Git), and even reproducible scripts for semester papers or bachelor theses (e.g., via Jupyter Notebook).

Truan showed that such a systematic introduction to open science and R&R in student courses can be effective: not only do students learn that these practices are potentially important in their own future work, but they feel *committed* to these practices, as they want to demonstrate them to fellow researchers in their community, planting the seed for a cultural change (Nosek et al. 2015).

Contributions to Reproducibility and Replicability: A Synthesis

As illustrated in Fig. 1, we want to synthesize the Wikimedia projects and describe what they can teach about R&R in the context of quantitative research, qualitative research and the humanities, as well as applied research (research domains in the middle part of the figure). Moreover, we aimed to contrast R&R (light boxes on the right side of the figure) to questionable, yet common research practices that may hinder research progress (dark boxes on the left side of the figure).

Many of the Wikimedia Fellowship projects recognized and served a need for efforts to reproduce and replicate, either due to statistical-methodological problems in their field or due to limited comparability of research output more generally. Such problems include underpowered tests, decreasing the likelihood that statistically significant results show true effects (Button et al. 2013), and questionably flexible rather than rigorous study designs and analysis methods (Ioannidis 2005; John et al. 2012; Pashler and Wagenmakers 2012), increasing the likelihood of finding false-positive results (Simmons et al. 2011), which may inflate the literature (Rosenthal 1979). These problems have been shown to lead to worryingly low replication rates across empirical research fields (Baker 2016; Begley and Ellis 2012; Camerer et al. 2016; Cova et al. 2021; Errington et al. 2014; Open Science Collaboration 2015; Rodgers and Collings 2021). However, when knowledge about open science practices, as well as the handling of repositories, useful software, and the documentation of analysis code are systematically taught and applied, these QRPs may dramatically diminish over time, ensuring that mostly trustworthy findings find their way into the literature.

Applied research and egineering might face issues that are different from the questionable research practices known from basic research. Rather, these problems concern intransparency and the subsequent lack of standardization of methods (see the project by Hans Henning Stutz). This is mainly because the development of tools and devices is done by individual researchers or small groups. They may be reluctant to share details about their materials and devices because they perceive their materials as intellectual property and do not see direct benefits in sharing them. Here, too, awareness has to rise that sharing of methods, codes, and construction plans has beneficial effects for the whole field. In the best case, engineers can exchange their knowledge, as it stems from similar software and tools, which increases comparability, but also the chance for collaborative endeavors across countries.

In the humanities, contributions to R&R focus mainly on increasing the digital availability of well-curated collections of artifacts: since humanities research relies on samples of intellectual production to be interpreted, contextualized, and compared, researchers used to elicit data by visiting archives or going on field trips by themselves. This often produces highly idiosyncratic notes that were never released except through the filtered form of published interpretations. This traditional approach faces increasing competition from digital research tools: As libraries and archives digitize their holdings, field trips are less relevant, while materials are accessible to and shared with greater numbers of researchers for mutual scrutiny, thus presumably increasing the reliability of interpretations derived from them.

Conclusion

Overall, our analysis shows that R&R is relevant across scientific disciplines and cultures, be it in the humanities, engineering, life science, natural sciences, or social sciences. Projects dedicated to advancing R&R took demonstrably different approaches, varying

from enhancing step-by-step reproducibility through code-based transparency to tracing the origin of an argument through publication lines.

A majority of the 12 Wikimedia Fellowship projects assessed in detail here stem from social sciences and psychology, which closely mirrors recent developments in this research area (see Open Science Collaboration (2015)), including a trend towards an increased awareness of R&R challenges, as well as activities designed to overcome them. While relatively fewer contributions emerged from humanities and applied sciences, the above projects may serve as a suitable springboard for future initiatives in these areas.

Notable differences were at which stages of the research process R&R becomes most relevant: whereas in basic research transparency remains relevant from the planning phase of a study to its publication and to its replication, in applied research the main focus may lie on the transparency of methods, as its goal may not be a reproducible and reproducible study, but a comparable methodology. In the humanities, a main focus may be to achieve an objective and reliable interpretation of materials and artifacts and to share how one came to this interpretation.

Despite their marked differences, the cross-disciplinary projects reviewed here shared the goal of improving research practices through R&R. Our findings therefore illustrate that R&R recipes require adjustments to fulfill the needs of the respective fields and research traditions. As chefs adjust recipes to their tastes in the kitchen, researchers may need to adjust how they think of and work with R&R against the background of their disciplines. If these specific needs are addressed appropriately, we are optimistic that an open research culture, which holds R&R at its core, may lie ahead in the not-so-distant future.

Acknowledgements

We thank Lilli Wagner and Berit Heling for their help with formatting this manuscript.

The publication of this article was kindly supported by RIO. We would like to thank RIO and Wikimedia Deutschland for enabling this collection.

Author contributions

During data elicitation, each author reviewed, coded, and described three out of the twelve projects that form the basis of our analysis. The other coauthors subsequently reviewed and revised these parts. The remainder of the text was written collaboratively in iterative revisions.

Conflicts of interest

The authors Brohmer, Pethig, and Rahal were also the principal investigators in projects analyzed in this paper. We are not aware of other potential conflicts.

References

- Anderson M, Martinson B, De Vries R (2016) Normative dissonance in science: Results from a national survey of U.S. scientists. Journal of Empirical Research on Human Research Ethics 2 (4): 3-14. https://doi.org/10.1525/jer.2007.2.4.3
- Armbruster S (2021) What makes a replication successful? An investigation of frequentist and Bayesian criteria to assess replication success. Ludwig Maximilian University of Munich https://doi.org/10.5282/ubm/epub.77434
- Arqus Alliance (2022) Arqus openness position paper. https://doi.org/10.5281/zenodo.588190
- Arslan RC (2017) Paternal age effects on offspring fitness in four populations. https://rubenarslan.github.io/paternal_age_fitness/0_krmh_massage.html#authors-acknowledgements. Accessed on: 2022-5-24.
- Arslan RC, Willführ KP, Frans E, Verweij KJH, Bürkner PC, Myrskylä M, Voland E, Almqvist C, Zietsch B, Penke (2017) Paternal age and offspring fitness: Online supplementary website (v2.0.1) . https://doi.org/10.5281/zenodo.838961
- Baker M (2016) 1,500 scientists lift the lid on reproducibility. Nature News 533: 452-454. https://doi.org/10.1038/533452a
- Begley CG, Ellis LM (2012) Raise standards for preclinical cancer research. Nature 483: 531-533. https://doi.org/10.1038/483531a
- Behrens S, Schwarzkopf C, Gödeke A-, Scholl D, Schneider N (2022) Wikimedia Fellow-Programm Freies Wissen 2016 - 2021. Zenodo URL: https://zenodo.org/record/5788379
- Brandt MJ, Ijzerman H, Dijksterhuis A, Farach FJ, Geller J, Giner-Sorolla R, Grange JA, Perugini M, Spies JR, van 't Veer A (2014) The replication recipe: What makes for a convincing replication? Journal of Experimental Social Psychology 50: 217-224. https://doi.org/10.1016/j.jesp.2013.10.005
- Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: Why small sample size undermines the reliability of neuroscience. Nature Reviews Neuroscience 14 (5): 365-376. https://doi.org/10.1038/nrn3475
- Camerer CF, Dreber A, Forsell E, Ho T-, Huber J, Johannesson M, Kirchler M, Almenberg J, Altmejd A, Chan T, Heikensten E, Holzmeister F, Imai T, Isaksson S, Nave G, Pfeiffer T, Razen M, Wu H (2016) Evaluating replicability of laboratory experiments in economics. Science 351 (6280): 1433-1436. https://doi.org/10.1126/science.aaf0918
- Chambers C (2017) The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice. Princeton University Press https://doi.org/10.2307/j.ctvc779w5
- Cova F, Strickland B, Abatista A, Allard A, Andow J, Attie M, Beebe J, Berniūnas R, Boudesseul J, Colombo M, Cushman F, Diaz R, N'Djaye N, van Dongen N, Dranseika V, Earp BD, Torres AG, Hannikainen I, Hernández-Conde JV, Hu W, Jaquet F, Khalifa K,

Kim H, Kneer M, Knobe J, Kurthy M, Lantian A, Liao S, Machery E, Moerenhout T, Mott C, Phelan M, Phillips J, Rambharose N, Reuter K, Romero F, Sousa P, Sprenger J, Thalabard E, Tobia K, Viciana H, Wilkenfeld D, Zhou X (2021) Estimating the reproducibility of experimental philosophy. Review of Philosophy and Psychology 12: 9-44. https://doi.org/10.1007/s13164-018-0400-9

- D'Alessandro A, Grigorescu O, Höchenberger R, Ohla K, Hummel T (in press) A Bayesian adaptive algorithm (QUEST) to estimate olfactory threshold in hyposmic patients. Journal of Sensory Studies.
- Devezer B, Nardin L, Baumgaertner B, Buzbas EO (2019) Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. PLOS ONE 14 (5). https://doi.org/10.1371/journal.pone.0216125
- Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA (2014) An open investigation of the reproducibility of cancer biology research. eLife 3: e04333. https://doi.org/10.7554/eLife.04333
- Feest U (2019) Why replication is overrated. Philosophy of science 86 (5): 895-905. https://doi.org/10.1086/705451
- Fiedler K, Prager J (2018) The regression trap and other pitfalls of replication science-Illustrated by the report of the Open Science Collaboration. Basic and Applied Social Psychology 40 (3): 115-124. https://doi.org/10.1080/01973533.2017.1421953
- Figueiredo L, Scherer C, Cabral JS (2022) A simple kit to use computational notebooks for more openness, reproducibility, and productivity in research. PLOS Computational Biology 18 (9): 1010356. https://doi.org/10.1371/journal.pcbi.1010356
- Gilbert DT, King G, Pettigrew S, Wilson TD (2016) Comment on “Estimating the reproducibility of psychological science.". Science 351 (6277): 1037. https://doi.org/10.1126/science.aad7243
- Goodman SN, Fanelli D, Ioannidis JP (2016) What does research reproducibility mean? Science Translational Medicine 8 (341): 341ps12-341ps12. https://doi.org/10.1126/scitranslmed.aaf5027
- Gunton RM, Stafleu MD, Reiss MJ (2021) A general theory of objectivity: Contributions from the reformational philosophy tradition. Foundations of Science1-15. https://doi.org/10.1007/s10699-021-09809-x
- Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. Philosophy of Science 15 (2): 135-175. https://doi.org/10.1086/286983
- Hempel CG (1968) Maximal specificity and lawlikeness in probabilistic explanation. Philosophy of Science 35 (2): 116-133. https://doi.org/10.1086/288197
- Inbar Y (2016) Association between contextual dependence and replicability in psychology may be spurious. Proceedings of the National Academy of Sciences of the United States of America 113 (34): E4933-E4934. https://doi.org/10.1073/pnas.1608676113
- Ioannidis JP (2005) Why most published research findings are false. PLoS medicine 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124
- John LK, Loewenstein G, Prelec D (2012) Measuring the prevalence of questionable research practices with incentives for truth telling. Psychological Science 23 (5): 524-532. https://doi.org/10.1177/0956797611430953
- Klein RA, Ratliff KA, Vianello M, Adams Jr. RB, Bahník Š, Bernstein MJ, Bocian K, Brandt MJ, Brooks B, Brumbaugh CC, Cemalcilar Z, Chandler J, W. C, Davis WE, Devos T, Eisner M, Frankowska N, Furrow D, Galliani EM, Hasselman F, Hicks JA, Hovermale

JF, Hunt SJ, Huntsinger JR, Ijzerman H, John M-, Joy-Gaba JA, Kappes HB, Krueger LE, Kurtz J, Levitan CA, Mallett RK, Morris WL, Nelson AJ, Jier JA, Packard G, Pilati R, Rutchick AM, Schmidt K, Skorinko JL, Smith R, Steiner TG, Stoybeck J, Van Swol LM, Thompson D, van 't Veer AE, Vaughn LA, Vranka M, Wichman AL, Woodzicka JA, Nosek BA (2014) Investigating variation in replicability. Social psychology 45 (3): 142-152. https://doi.org/10.1027/1864-9335/a000178

- Lasser J (2020) Creating an executable paper is a journey through Open Science. Communication Physics 3 (143): 1-5. https://doi.org/10.1038/s42005-020-00403-4
- Leonelli S (2022) Open Science and Epistemic Diversity: Friends or Foes? Philosophy of Science1-21. https://doi.org/10.1017/psa.2022.45
- Mathur MB, VanderWeele TJ (2019) Challenges and suggestions for defining replication "success" when effects may be heterogeneous: Comment on Hedges & Schauer (2018). Psychological Methods 24 (5): 571-575. https://doi.org/10.1037/met0000223
- Maxwell S, Lau M, Howard G (2015) Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? The American Psychologist 70 (6): 487-498. https://doi.org/10.1037/a0039400
- Merton RK (1942) A note on science and democracy. Journal of Legal and Political Sociology 1: 115-126. URL: https://heinonline.org/HOL/Contents?handle=hein.journals/jolegpo1
- Muradchanian J, Hoekstra R, Kiers H, van Ravenzwaaij D (2021) How best to quantify replication success? A simulation study on the comparison of replication success metrics. Royal Society Open Science 8 (5): 201697. https://doi.org/10.1098/rsos.201697
- Nosek BA, Spies JR, Motyl M (2012) Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspectives on Psychological Science 7 (6): 615-631. https://doi.org/10.1177/1745691612459058
- Nosek BA, Alter G, Bowman SD, Breckler SJ, Buck S, Chambers CD, Chin G, Christensen G, Contestabile M, Dafoe A, Eich E, Freese J, Glennerster R, Goroff D, Green DP, Hesse B, Humphreys M, Ishiyama J, Karlan D, Kraut A, Lupia A, Mabry P, Madon T, Malhotra N, Mayo-Wilson E, McNutt M, Miguel E, Paluck EL, Simonsohn U, Soderberg C, Spellman BA, Turitto J, Vandenbos G, Vazire S, Wagenmakers EJ, Wilson R, Yarkoni T (2015) Promoting an open research culture. Science 348 (6242): 1422-1425. https://doi.org/10.1126/science.aab2374
- Nosek BA, Errington TM (2017) Making sense of replications. ELife 6: e23383. https://doi.org/10.7554/eLife.23383
- Nosek BA, Errington TM (2020a) What is replication? PLOS Biology 18 (3): 3000691. https://doi.org/10.1371/journal.pbio.3000691
- Nosek BA, Errington TM (2020b) The best time to argue about what a replication means? Before you do it. Nature 583 (7817): 518-520. https://doi.org/10.1038/d41586-020-02142-6
- Nosek BA, Hardwicke TE, Moshontz H, Allard A, Corker KS, Dreber A, Fidler F, Hilgard J, Struhl MK, Nuijten MB, Rohrer JM, Romero F, Scheel AM, Scherer LD, Schönbrodt FD, Vazire S (2022) Replicability, robustness, and reproducibility in psychological science. Annual Review of Psychology 73 (1): 719-748. https://doi.org/10.1146/annurev-psych-020821-114157
- Open Science Collaboration (2015) Estimating the reproducibility of psychological science. Science 349 (6251): aac4716. https://doi.org/10.1126/science.aac4716

- Pashler H, Wagenmakers EJ (2012) Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? Perspectives on Psychological Science 7 (6): 528-530. https://doi.org/10.1177/1745691612465253
- Patil P, Peng RD, Leek JT (2016) What should researchers expect when they replicate studies? A statistical view of replicability in psychological science. Perspectives on Psychological Science: A Journal of the Association for Psychological Science 11 (4): 539-544. https://doi.org/10.1177/1745691616646366
- Peels R (2019) Replicability and replication in the humanities. Research Integrity and Peer Review 4 (1): 2. https://doi.org/10.1186/s41073-018-0060-4
- Plesser HE (2018) Reproducibility vs. replicability: A brief history of a confused terminology. Frontiers in Neuroinformatics 11 https://doi.org/10.3389/fninf.2017.00076
- Rahal RM (2020) Open for Insight: An online course in experimentation. PsychArchives. https://doi.org/10.23668/psycharchives.4319
- R Core Team (2020) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.r-project.org/index.html
- Reiss J, Sprenger J (2020) Scientific objectivity. In: Zalta EN (Ed.) The Stanford Encyclopedia of Philosophy (Winter 2020). Metaphysics Research Lab, Stanford University URL: https://plato.stanford.edu/archives/win2020/entries/scientific-objectivity/
- Robson SG, Baum MA, Beaudry JL, Beitner J, Brohmer H, Chin JM, Jasko K, Kouros CD, Laukkonen RE, Moreau D, Searston RA, Slagter HA, Steffens NK, Tangen JM, Thomas A (2021) Promoting open science: A holistic approach to changing behaviour. Collabra: Psychology 7 (1): 30137. https://doi.org/10.1525/collabra.30137
- Rodgers P, Collings A (2021) What have we learned? eLife 10: e75830. https://doi.org/10.7554/eLife.75830
- Rosenthal R (1979) The file drawer problem and tolerance for null results. Psychological Bulletin 86 (3): 638-641. https://doi.org/10.1037/0033-2909.86.3.638
- Simmons JP, Nelson LD, 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
- Solomon M (2012) Norms of epistemic diversity. Episteme 3: 23-36. https://doi.org/10.3366/epi.2006.3.1-2.23
- Stahlberg D, Sczesny S, Braun F (2001) Name your favorite musician: Effects of masculine generics and of their alternatives in German. Journal of Language and Social Psychology 20 (4): 464-469. https://doi.org/10.1177/0261927X01020004004
- Stroebe W, Strack F (2014) The alleged crisis and the illusion of exact replication. Perspectives on Psychological Science 9 (1): 59-71. https://doi.org/10.1177/1745691613514450
- Thiery F, Oertel C (2021) How to increase data quality in the digital humanities? Two linked open data community approaches (preprint on file with authors). Elsevier
- Truan N, Dressel D (2021a) Doing open science in a research-based seminar: Students' positioning towards openness in jigher education. HAL SHS Archives ouvertes URL: https://halshs.archives-ouvertes.fr/halshs-03395171
- Truan N, Dressel D (2021b) Das eigene digitale Schreiben erforschen: Ein sprachwissenschaftliches Seminarkonzept zur Produktion, Analyse und Reflexion eigener digitaler Schreibpraktiken für angehende Deutschlehrkräfte. Herausforderung Lehrer*innenbildung - Zeitschrift zur Konzeption, Gestaltung und Diskussion (HLZ 4 (1): 378-397. https://doi.org/10.11576/hlz-4343

- United Nations Educational, Scientific and Cultural Organization (2021) UNESCO recommendation on open science (SC-PCB-SPP/2021/OS/UROS). URL: https://unesdoc.unesco.org/ark:/48223/pf0000379949.locale=en
- Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA (2016) Contextual sensitivity in scientific reproducibility. Proceedings of the National Academy of Sciences of the United States of America 113 (23): 6454-6459. https://doi.org/10.1073/pnas.1521897113
- Van Rossum G, Drake Jr. FL (1995) Python reference manual. Centrum voor Wiskunde en Informatica Amsterdam URL: https://www.python.org/
- Zwaan RA, Etz A, Lucas RE, Donnellan MB (2017) Making replication mainstream. Behavioral and Brain Sciences 41: e120. https://doi.org/10.1017/S0140525X17001972

Endnotes

*1 Note that the project used the terms “reproducing” and “replicating” in exclusively negative connotation: It described art history as a field where flawed historical speculation gets carried forward by “continuous replication”, by being “still reproduced today”, even when newer evidence is available and past falsehoods have been corrected. This language construes reproduction as an ailment, so it may not appear intuitive to seek reproducibility as its cure.

*2 Novel materials may be used either because the original materials are unavailable (e.g., because they have been lost or because they are not shared openly with the replication team), unsuitable (e.g., because they are written in a different language than that used by the replication team), or systematic variation is required (e.g., because boundary conditions should be tested; conceptual replication sensu (Nosek and Errington 2017)).

*3 This is a project by one of the co-authors.

*4 In this context, we acknowledge that even the term "data" could mean very different things in different epistemic cultures and (within and across) academic fields (see Leonelli 2022). For the sake of practicality, we understand data as all kinds of research output that is beneficial for achieving the goal of the research project (i.e., answering a research question in an unbiased fashion or building a device to solve research-related problems).

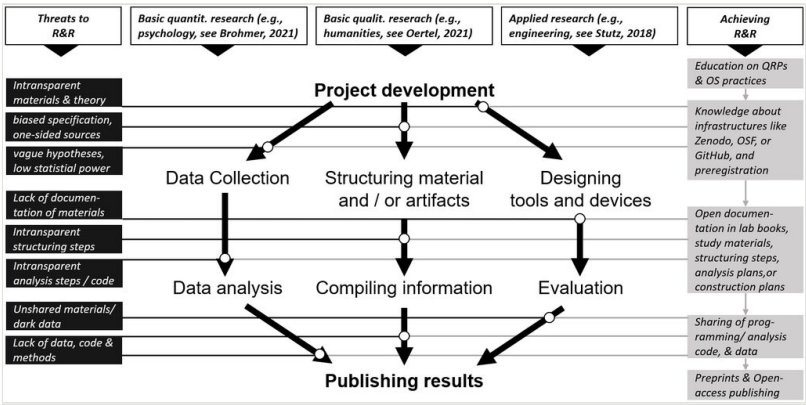

Figure 1.

Threats to R&R and how to achieve it in three research domains. Note that horizontal lines connect different research steps (marked by circles on arrows) with boxes (“threats” and “achieving”) for R&R; light boxes are interconnected because R&R is an incremental process; domains are prototypical only and can overlap with other domains in reality. R&R = reproducibility and replicability; QRPs = questionable research practices; OS = open science.

Table 1.

Overview of all Wikimedia Fellowship projects with focus on R&R. A complete list with annotations and more details on the projects can be retrieved from the OSF: https://osf.io/kqw3h/?view_only=296cae9077d146ee92d2372eed15d6c8.

Project leader	Funding year	Project title	Academic discipline	Link
Arslan, R.C.	2016/17	Reproducible websites for everyone	Social Sciences	https://rb.gy/nkh65t
Breznau, N.	2019/20	Giving the Results of Crowdsourced Research Back to the Crowd. A Proposal to Make Data from 'The Crowdsourced Replication Initiative' Reliable, Transparent and Interactive	Social Sciences	https://rb.gy/ydez7u
Brohmer, H.	2020/21	Effects of Generic Masculine and Its Gender-fair Alternatives. A Multi-lab Study	Social Sciences	https://rb.gy/bwqhdu
Figueiredo, L.	2020/21	Computational notebooks as a tool for producitvty, transparency, and reproducibility	Life Sciences	https://rb.gy/jkv1zp
Hoffman, F.Z.	2017/18	Code, Data and Reproducibility – Open Computational Research	Engineering	https://rb.gy/hjaxg7
Höchenberger, R.	2017/18	Quick estimation of taste sensitivity	Life Sciences	https://rb.gy/ysbjlx
Lasser, J.	2019/20	Executable papers: Tools for more reproducibility and transparency in the natural sciences	Natural Sciences	https://rb.gy/7dgwka
Oertel, C.	2020/21	Acceleration of quality in the humanities – chances of open source implementation in research and training	Humanities	https://rb.gy/arpqf2
Pethig, F.	2020/21	Data Version Control: Best Practice for Reproducible, Shareable Science?	Social Sciences	https://rb.gy/ihzetc
Rahal, R.-M.	2018/19	Reproducible practices make open and transparent research: An online course	Social Sciences	https://rb.gy/jeppln
Stutz, H.	2017/18	The Galss Tool – from development, to use, to the data set	Engineering, Natural Sciences	https://rb.gy/bi7lxf

Truan, N.	2020/21	Digital data – mine, yours, ours? Linguistic resources about digital communication as Open Data and Open Educational Resources for (higher) education	Humanities	https://rb.gy/trir4h