The Social Dynamics of Reform

The Citation Advantage of Replications and the Explosion of Prosocial Language in the Psychological Science Reform Movement

Victor Møller Poulsen (201707639)

Department of Linguistics, Cognitive Science and Semiotics
Aarhus University (AU)
Supervisor: Riccardo Fusaroli
June 1st, 2022

Abstract

The replication crisis has left psychological science in a crisis of faith. Both conceptual (Ioannidis, 2005) and empirical (Open Science Collaboration, 2015) results have suggested that reproducibility issues are wide-spread. This challenges the core notion of science as a cumulative and productive behavior and has damaged the credibility and public image of psychological science. A heterogeneous movement of reformers has emerged to address the apparent issues of reproducibility and replicability (Flis, 2022). In this thesis, we address two critical and contested issues. We ask (i) whether reproducible research is disincentivized and (ii) whether the reform movement is promoting a prosocial culture. We address (i) by comparing the rate of citations for replication studies and matched non-replication studies in psychology. We use a robust and multi-level Bayesian modeling approach, which improves on standard practice in the scientometrics domain. We address (ii) with a novel combination of topic modeling, semantic analysis, and network analysis on a large-scale Twitter corpus. In part (i) we show that replication studies are cited at higher rates than non-replication studies, and we suggest that the citation advantage is increasing over time. In part (ii) we show that the discourse in the reform community is increasingly prosocial, and we document a shift in the reform agenda. The results from (i) suggest that psychologists care about reproducible research and that incentives might increasingly favor reproducible research. We suggest that better-aligned incentives could accelerate the adoption of reform practice. The results from (ii) show that the reform movement is addressing the historical lack of diversity and inclusion in psychological science. This is important because diversity and inclusion are desirable properties, both politically (Gruber et al., 2021) and epistemically (Devezer et al., 2019). The two analyses jointly suggest that the reform movement in psychological science is accelerating and that it is increasingly focusing on promoting a prosocial culture.

Keywords— Reform Psychology, Replication Crisis, Open Science, Bayesian Modeling, Social Networks, Topic Modeling

Contents

1 Introduction

Science is an amazing enterprise that continually extends our knowledge of the physical world through a cycle of theory development and empirical tests. A common conception is that science is cumulative, and that we make progress by standing on the shoulders of giants (Scotchmer, 1991; Zeigler, 2012). The replication crisis in psychology has challenged this notion in psychological science since it is unclear whether a field of inquiry can be thought of as cumulative if key results do not replicate (Nosek et al., 2022). Issues of reproducibility and replicability in psychological science are well-documented and wide-spread (Open Science Collaboration, 2015). Although the replication crisis has highlighted deep challenges, the replication crisis can also be viewed as an opportunity to introspect and reform practices (Munafò et al., 2022; Murphy et al., 2020). Depending on perspective, some have called the 2010s a decade of "crisis", "revolution" or even "renaissance" for psychological science (Nosek et al., 2022).

This thesis aims to extend our understanding of what I will call the "Reform Psychology" movement, which I take to be the "progressive" (Flis, 2022) response to the replication crisis in psychology. While methodological concerns are not new, it is a combination of recent conceptual and empirical contributions which has put issues of replicability and reproducibility to the front of psychological science (Nosek et al., 2022). Some evidence suggests that more than half of the findings in the published psychological science literature could be null findings (Ioannidis, 2005; Open Science Collaboration, 2015), and collectively ($n = 77$) only 56% of multi-site replications in psychology have found a significant effect in the same direction as original studies (Nosek et al., 2022). This is problematic because science advances by building on top of previous work, and this cumulative notion suffers if the effects in the literature cannot generally be trusted (Nosek et al., 2022). Given the overwhelming evidence that psychological science suffers from issues of reproducibility, I believe that it is unsustainable to perpetuate current practices. A lack of scientific credibility could result in less funding for psychological science (Rodgers & Shrout, 2018), and I believe that the field has an obligation to improve practices in order to credibly advance our understanding of the mind and human behavior. In this thesis, we ask two broad questions, which we will operationalize and investigate in depth.

- Is reproducible research in psychological science disincentivized?
- Is the "Reform Psychology" movement addressing issues of diversity and inclusion?

The first question is important to address because incentives influence the adoption of reform practice. Clearly, scientists are not driven exclusively by incentives, and we should expect scientists to conduct research that they believe to have high scientific value (Smaldino & McElreath, 2016). Still, if psychological science is serious about cultural change, it is sensible to align incentives such that reproducible research is encouraged. Currently, reproducible and robust research is thought to be disincentivized. First, rigorous and reproducible research is demanding and might result in fewer publications (Smaldino & McElreath, 2016; Allen et al., 2017). In this trade-off between quantity and quality, the current system is perceived as favoring quantity (Allen & Mehler, 2019). In addition, journals favor ”positive” findings, and in particular ”novel” results (Smaldino & McElreath, 2016; Allen & Mehler, 2019). The ability to publish reproducible research might be challenged in this system since reproducible and robust research is associated with higher rates of null findings, and will generally tend to produce less ”clean” results (Allen & Mehler, 2019; Hummer et al., 2017; Chambers, 2019). Since the number of publications, especially as first and last author, is associated with career advancement (Allen & Mehler, 2019), this can create a system in which there is a ”natural selection of bad science” (Smaldino & McElreath, 2016). Replication studies in particular have been argued to be disincentivized (Nosek et al., 2022), and this could explain the historically low rate of replications in psychological science (Makel et al., 2012).

The second question is important to address because the replication crisis is an opportunity to reform not just the methodology and research practice in psychological science, but also to reform the scientific culture more broadly (Murphy et al., 2020). Science has traditionally been dominated by the mythology of the lone (male) genius and has emphasized competition rather than a collaborative and prosocial culture (Heering, 2017; Jamieson, 2018). It is well documented that there are gendered issues related to career advancement, promotion, retention, citation rates, and role assignment in psychological science (Gruber et al., 2021). We will not want to empower reformers who perpetuate a broken research culture. Instead, we will want to institutionalize practices that promote a more diverse and inclusive psychological science (Murphy et al., 2020). This is important not just for political reasons, but because diversity is associated with innovation, and generally better scientific outcomes (Devezer et al., 2019; Hofstra et al., 2020; Nielsen et al., 2017). There is conflicting evidence as to whether the ”Reform Psychology” movement is promoting a prosocial culture of diversity and inclusion. The reform movement, and especially the dominant ”Open Science” affiliated group, has been criticized for being exclusive (Whitaker & Guest, 2020) and it continues to be dominated by white, male scientists from North America (Coles et al., 2022). On

the other hand, an empirical investigation of the semantics of the ”Open Science” movement has found that it employs ”prosocial” language and that the representation of females within this literature is growing (Murphy et al., 2020). We aim to address and resolve this apparent contradiction, by constructing a large-scale corpus of tweets from the ”Reform Psychology” movement and by analyzing the evolution of prosocial language within this corpus. In order to understand the social dynamics of reform, the investigation will also include a focus on important topics of discourse and the network centrality of key groups of reformers on Twitter.

Although the two questions are approached with distinct methods applied to different domains, the two analyses are complementary. In order to understand the dynamics of reform, we believe that novel combinations of analytic techniques applied to different domains of data are needed. Twitter in particular has been a central platform for the real-time discussion following the replication crisis (Flis, 2022; Derksen & Field, 2021), but I am not aware of any previous empirical investigations of the ”Reform Psychology” movement on Twitter. Our two domains of study (the published literature and Twitter) are interlinked in the sense that developments on Twitter are likely to have a delayed effect on the published record as ideas disseminate and become adopted. This thesis leverages Bayesian statistics, Natural Language Processing (NLP) methods, and network science approaches to construct a holistic view of the ”Reform Psychology” movement. The analysis provides a template for future cross-domain analysis, highlighting the benefits of Bayesian multi-level modeling applied to citation analysis, and an innovative combination of NLP analyses to evaluate the semantics of an online community. All code and analysis for this thesis are documented and openly available on Github at https://github.com/victor-m-p/reform-psychology.

1.1 Replication Crisis

It is tricky to place the replication crisis in a historical context, and impossible to put a definite time or event as the starting point. Criticism of the standard paradigm of psychology goes back to at least the 1960s-1970s, and Null Hypothesis Significance Testing (NHST) became increasingly controversial in psychology in the 1980s-1990s (Gelman, 2016). However, it is not before the 2000s or the early 2010s that the replication crisis can be said to have pervaded psychology broadly (Nosek et al., 2022). There are several important articles in this period, including ”Why Most Published Research Findings Are False” (Ioannidis, 2005), ”Voodoo Correlations in Social Neuroscience” (Vul et al., 2009) and ”False-positive Psychology” (Simmons et al., 2011). Typical of this period, the

papers mainly focus on poor understanding of statistics, questionable methodology, and problematic research practices in psychology and related fields. While these papers signal a growing suspicion in parts of the field, the paper "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect" by Bem (2011) is often referenced as a turning point (Gelman, 2016; Engber, 2017). This article presented evidence for extrasensory perception (ESP) and was published in a top psychology journal. While few believed in the result, the methodology was generally perceived as solid, and it followed a standard NHST paradigm ($p < 0.05$). The fact that the methodological paradigm of psychological science could produce "solid" evidence for such an outrageous effect was broadly perceived as an issue by psychologists (Gelman, 2016; Wagenmakers et al., 2011) and the case also received (negative) publicity outside of academia (Engber, 2017).

If some psychologists had not fully accepted that the field was facing a replication crisis following Bem (2011), many recognized that psychological science had a wide-spread replicability issue with the Open Science Collaboration (2015) publication. This large-scale replication effort sought to replicate 100 findings from three social psychology, developmental psychology, and cognitive psychology flagship journals. The results were discouraging. 97% of original studies reported significant results at the standard threshold of $p < 0.05$ while only 36% of replications found significant results in the same direction. In addition, the mean effect size of the replication studies was only half the magnitude of the mean effect size of the original studies (Open Science Collaboration, 2015). The study continues to be extremely influential and has accrued more than 6000 citations to date. Crucially, the study suggested that issues were not confined to parts of psychology, but that the field had a wide-ranging problem. Social psychology continues to be perceived as the area with the most acute replicability issues in psychological science (Szollosi & Donkin, 2021; Stevens, 2017) but this study showed that issues were not confined to social psychology.

1.2 Reform Psychology

I refer to the progressive response to the replication crisis in psychology as "Reform Psychology". This is not a homogeneous group or movement, but rather a constellation of various groups of voices with different views about how to most effectively improve and reform psychological science (Flis, 2022; Malich & Munafò, 2022). Understanding the social dynamics of the reform movement is a worthwhile effort in its own right, and it is helpful to focus on at least three groups of actors in order to answer our questions. Following the review by Flis (2022) I define (i) "Mainstream Reformers"

(ii) ”Theory Reformers” and (iii) ”Old Guard”. We will pay particular attention to the ”Mainstream Reformers” and the ”Theory Reformers” since understanding the dynamics between these groups will help us answer our questions.

The ”Mainstream Reformers” have focused mainly on issues with methodology and questionable research practices (QRPs). This focus dominated especially leading up to, and in the immediate aftermath of the replication crisis (Flis, 2022; Gelman, 2016). The mainstream reform group has largely been subsumed within the broader Open Science movement (Hesse, 2018) where the Center for Open Science (COS) has been particularly central for psychology (McKiernan et al., 2016). I will argue that this coincides with a shift of focus from diagnosing issues with methodology and research practice to a focus on institutional and behavioral change, critically focusing on aligning publication incentives such that they promote reproducible research. The ”Mainstream Reformers” have in particular pushed for more high-powered replication studies (Open Science Collaboration, 2015; Williamson, 2022) and for the adoption of the Registered Report (RR) format (Chambers, 2019; “OSF Registered Reports”, 2022). As the director and co-founder of the COS, the psychologist Brian Nosek has been central in associating the ”Mainstream Reformers” with the Open Science movement, and pushing for institutional change. Simine Vazire, another reformer, recently emphasized that ”It’s hard to overstate how central Brian Nosek’s role in the reform movement [...] has been. [...] Brian has been a leader in the movement” (Williamson, 2022). The COS has been involved with the ”Many Labs” replication series (Williamson, 2022), the Open Science Collaboration (2015) study, the Psychological Science Accelerator (PSA) (Moshontz et al., 2018), the Open Science Framework (OSF) preregistration platform (“OSF”, 2022) and the push for the adoption of the RR format (“OSF Registered Reports”, 2022).

The first wave of criticism of the ”Mainstream Reform” group came from what I call the ”Old Guard”, which corresponds to what Flis (2022) calls the ”status-quoers”. The criticism from the ”Old Guard” has mainly focused on two fronts. One line of criticism has questioned the need for reform and denied that psychological science is experiencing a replication crisis (S. T. Fiske et al., 2016; Barrett, 2015). The other line of criticism has focused on the ”tone” of the debate. Reformers have been accused of ”trash talking” and a lack of civility. Perceived attacks by reformers on uncurated platforms (notably Twitter) have received particular criticism (Derksen & Morawski, 2022). A notable example is the article by Susan Fiske (S. Fiske, 2016), which was scheduled for publication in the APS Observer, and provoked an ”online firestorm” (S. T. Fiske, 2016). In the piece, Susan Fiske coined the legendary phrase ”methodological terrorism”, and described reformers as

”data police”. This line of criticism highlights that the initial battle over reform was mainly fought over questions of methodology and statistical practice. Although there continues to be push-back from the ”Old Guard”, we do not observe a strong presence in the Twitter corpus that we gather, and this group has become less vocal as ”the winds have changed” (Gelman, 2016).

A more interesting line of criticism for our purposes is the more recent ”Theory Reform” criticism that has come from within the reform community. There has been a call for diversity at multiple levels from this group of critical reformers. On the one hand, the ”Theory Reformers” have criticized the ”preregistration” agenda which is central to the ”Mainstream Reform” community (Szollosi et al., 2020). The ”Theory Reformers” instead emphasize the importance of better theorizing and computational modeling to improve psychological science (Guest & Martin, 2021; Fried, 2020; van Rooij & Baggio, 2020; van Rooij, 2022; D. J. Navarro, 2021; Smaldino, 2020). This difference in perspective can be attributed, at least partly, to the fact that the ”Mainstream Reform” community is dominated by social, clinical, and experimental psychologists, while ”Theory Reformers” generally come from different backgrounds, such as cognitive science and mathematical modeling (Flis, 2022). There are differences in how theory-heavy particular sub-fields of psychological science are, and a clear confirmatory-exploratory distinction might be more natural for scientists coming from particular research traditions (Szollosi & Donkin, 2021). As such, part of the criticism from the ”Theory Reform” group is that the ”Mainstream Reformers” have conceptualized the challenges of psychological science narrowly, and such that the proposed solutions only address issues that are recognized by parts of the field. In addition to this line of criticism, The ”Theory Reformers” have also criticized the culture of the ”Mainstream Reform” movement more broadly (Flis, 2022). Olivia Guest coined the term #bropenscience in 2017 which started a discussion about cultural problems within the reform movement. In particular, the criticism has focused on the narrow demographics of the reform movement, and on the fear that the reform movement will perpetuate a system of exclusion (Whitaker & Guest, 2020). Importantly, the criticism by this group of female ”Theory Reformers” have emphasized the importance of diversity, not only as a political but as an epistemic goal (Flis, 2022; Devezer et al., 2019). It has been suggested that both the professional background of critics and possibly the gender of critics are related to the marginalization of viewpoints within the broader reform movement (Flis, 2022). In particular, Iris Van Rooij, Berna Devezer, Danielle Navarro and Olivia Guest have been described as outsiders whose contributions are ignored, while Eiko Fried and Paul Smaldino have been recognized for their work on reformist theory (Flis, 2022; Scheel et al., 2021). Taken together, the work by Flis (2022) suggests a reform movement with a

core of ”Mainstream Reformers” and a periphery of marginalized (female) ”Theory Reformers”, who are critical of the proposals and the culture within the mainstream of the movement. We might say that while the ”Mainstream Reformers” have mainly focused on questionable research practice (QRP), this recent line of criticism has highlighted the need to address what we might call a questionable scientific culture (QSC).

1.3 Incentives of Reform

To understand the incentives of reform, we will focus on two key instruments that are proposed by the ”Mainstream Reformers” to improve the reproducibility and replicability of psychological science. The mainstream response to the replication crisis has featured a strong emphasis on replication studies (Nosek et al., 2022), both as a means to estimate the reproducibility of psychological science (Open Science Collaboration, 2015) and as a means to rid the literature of false findings (Nosek et al., 2022). There has been a particular focus on large-scale and multi-lab replication studies (Nosek et al., 2022; Williamson, 2022) one prominent example being the ”Many Labs” series (Klein et al., 2014; Klein et al., 2018; Klein et al., 2019; Ebersole et al., 2016; Ebersole et al., 2020). Of course, replication studies are not a novel phenomenon, but they are uncommon (Nosek et al., 2022), with estimates at just above 1% of the literature (Makel et al., 2012). Registered Reports (RR) have emerged as another key proposal to improve the reproducibility of psychological science (Flis, 2022; Hummer et al., 2017; Chambers, 2019). RR is a format where a study protocol including hypotheses and methods is peer-reviewed prior to data collection, and where an agreement exists such that results are published conditional on the preregistered study plan being followed. Importantly, whether ”positive” results are then obtained does not influence whether the work is published (Chambers, 2019). As such RRs integrate preregistration and publishing incentives (Hummer et al., 2017) promoting reproducible research rather than positive results. An analysis has shown that RRs of novel studies report null findings in 55% of cases whereas estimates of conventional papers range between $5 - 20\%$ (Allen & Mehler, 2019), suggesting both that issues of publication bias are wide-spread in the traditional literature, but also that the RR format does successfully combat this issue (Chambers, 2019). A survey of 350 reviewers found that RRs were rated higher than conventional articles on a number of desirable properties, including ”rigor of methods” and ”overall quality of paper” (Soderberg et al., 2020). Initial evidence suggests that RRs are cited at comparable, or even higher, rates than conventional journal articles (Hummer et al., 2017). This indicates that scientists care about reproducibility and robustness and that incentives

are not universally aligned against this type of research. The evidence from Hummer et al. (2017) is preliminary, however (e.g. small sample size) and we aim to critically strengthen the findings using a much larger sample of studies, more sophisticated modeling, and by extending the findings to the domain of replication studies.

1.4 Culture of Reform

To understand the culture of the ”Reform Psychology” movement, two lines of evidence are important. Criticism of the culture in the ”Mainstream Reform” group has come from a group of (female) ”Theory Reformers” who have questioned the ”preregistration” agenda (Szollosi et al., 2020), and highlighted issues of diversity and representation (Whitaker & Guest, 2020). A conflicting line of evidence comes from empirical work by Murphy et al. (2020). Focusing on a corpus of abstracts from ”Open Science” publications, they show that the ”Open Science” language is prosocial as measured against abstracts from the ”Reproducibility” literature. They also show that the representation of females is growing in the ”Open Science” literature and that the network of authors who have published in the ”Open Science” literature is dense and well-connected. The results in Murphy et al. (2020) are taken to show that the ”Open Science” community is prosocial and communal and that this fosters diversity and a greater representation of traditionally marginalized groups (e.g. females). However, the findings in Murphy et al. (2020) suffer from both low sample-size, and they survey science broadly, questioning their validity for psychological science specifically (see e.g. A.1.7). We aim to resolve the contrasting evidence from Murphy et al. (2020) and Whitaker and Guest (2020) by gathering and analyzing a large-scale and longitudinal Twitter corpus of ”Reform Psychology” discourse.

1.5 Research Questions

The present work will be structured as two sub-studies that aim at understanding the ”Reform Psychology” movement from two different angles. The first study (S1) will attempt to answer our first question; ”is reproducible research in psychological science disincentivized?”. It will do so by creating two matched sample data sets of replication studies in psychology, and using Bayesian statistics to model whether replication studies are cited less than matched controls. Because citations are an important metric of scientific prestige and impact, citations are a commonly used proxy in the science of science literature (Hummer et al., 2017; D. Wang et al., 2013). While our

main focus is the overall difference in citations for replication studies and non-replication studies (RQ_1), the study will pose two sub-questions ($RQ_{1.1}$, $RQ_{1.2}$) aimed at understanding the mechanisms driving an eventual effect. As such, we pose three questions in the first part (S1) of the thesis.

1. RQ_1: Are replication studies cited less than non-replication studies?
2. $RQ_{1.1}$: Are replication studies receiving relatively more citations than non-replication studies over time?
3. $RQ_{1.2}$: Are replication studies receiving a relatively larger citation boost than non-replication studies for studies conducted by large teams?

The second study (S2) will attempt to answer our second question; ”is the ”Reform Psychology” movement addressing issues of diversity and inclusion?”. To accomplish this, we construct and analyze a large-scale longitudinal corpus of tweets from the ”Reform Psychology” community on Twitter. We analyze the semantics of the tweets in this corpus in an attempt to conceptually replicate and to further extend the findings reported in Murphy et al. (2020). We directly tackle our main question by focusing on the evolution of prosocial language over time. In addition, we create a network of Twitter accounts in the ”Reform Psychology” space to understand the centrality of various groups of reformers, informed by the historical thesis in Flis (2022). Finally, we use topic modeling and semantic linkage networks (Perry & DeDeo, 2021) to understand the key topics of discourse within the ”Reform Psychology” movement, and how the discourse and agenda have changed over time. In sum, the second part (S2) of the thesis poses three questions.

1. RQ_2: Is the ”Reform Psychology” discourse prosocial, and do we observe a growing focus on diversity and inclusion within the community?
2. $RQ_{2.1}$: Which topics dominate the ”Reform Psychology” discourse on Twitter, and do we observe changing agendas?
3. $RQ_{2.2}$: Are known ”Theory Reformers” marginalized to the periphery of the Twitter network of ”Reform Psychology”?

2 S1: Methods

2.1 Data curation

We use a fork of the Microsoft Academic Graph (MAG) from 2021-08-12 to address the three research questions (RQ_1, $RQ_{1.1}$, $RQ_{1.2}$) that we pose in this first part (S1) of the thesis. MAG has high coverage (K. Wang et al., 2020; Paszcza, 2016), and is widely used in the scientometrics and science of science literature (Murphy et al., 2020; C. Chen, 2020; Effendy & Yap, 2017; K. Wang et al., 2019). The backbone of the MAG is that it provides data about the references between studies (study x cites study y, and study y cites study z). The MAG also contains important classifications of documents and meta-data, which it uses AI to generate (see A.1.1). Of particular relevance for this thesis is the fact that the MAG contains a hierarchical classification of scientific documents into categories, which they call ”Field of Study” (FoS). For instance, ”Psychology” is a base-level FoS, while ”Replication” is a higher-level FoS. A document can be assigned to multiple FoS categories at various levels of this hierarchy. As such, a given scientific article can be classified as both ”Psychology” and ”Replication” at the same time. We can treat such a document as being a replication study conducted in the field of psychology. Each FoS also has an assigned probability which indicates the confidence that MAG has in the FoS assignment. Secondly, documents are classified into ”Document Types”, such as ”Conference”, ”Journal” and ”Patent”, and we have access to important meta-data such as publication date (see A.1.1).

We start by locating all publications that are categorized by MAG with ”Psychology” as their most probable base-level FoS. We only include articles that are categorized as either ”Conference” or ”Journal” by the MAG. This selection follows previous practice (Murphy et al., 2020) and the records contained in the ”Conference” and ”Journal” categories can be thought of as ”core peer-reviewed science”. Lastly, we only include papers which are published between 2005 and 2020 (including both full years). We call this sample ”Candidate Papers”.

From our ”Candidate Papers” data set we create two smaller samples of data which we proceed to model independently. In both cases, we further require that papers are published between 2005 and 2015 for inclusion. 2015 is the last year for inclusion because we use the number of citations five years after publication (c_5) as our outcome variable. Since 2020 is the last full year of data that we have access to, 2015 is the last year of publication for which we can compute c_5. The c_5 metric is inspired by Sinatra et al. (2016) who used c_{10} as their outcome variable in a study

which investigated the impact of scientific articles. The operationalization of scientific ”prestige” or ”impact” is discussed in section A.1.2.

In the first case, we select papers that are additionally categorized as ”Replication” studies by the MAG. This data set, which we will call R_{FOS}, matches $n = 620$ replication studies. We treat this as our main case throughout. For the second subset, we select papers from our ”Candidate Papers” data set which additionally contain ”replicat*” in their title (inspired by Makel et al., 2012). We call this data set R_{QUERY}, and we match $n = 1196$ replication studies with this approach. This data set acts as a robustness check of the results from the model that we condition on the R_{FOS} data. This is sensible because both data sets are likely to contain different biases (see A.1.5 and A.1.6), but if the inferences from models based on both data sets are congruent, then this should raise our confidence in the robustness of the results. We also attempt to construct two additional subsets of data, corresponding to the ”Open Science” and ”Reproducibility” categories in the MAG, which were used by Murphy et al. (2020). We do not model these data sets because we encounter issues of bias and low coverage (see A.1.4 and A.1.5), but we do report on the volume of papers within these sub-fields.

For both the R_{FOS} and the R_{QUERY} data sets, we match the obtained records with non-replication studies based on the exact year of publication ($YEAR$) and approximately on number of authors ($TEAMSIZE$). We are able to match records almost perfectly ($> 99\%$) on these metrics and in the few cases of imperfect matches, we retain the best possible match (see A.1.3). The matched control studies are sampled from our ”Candidate Papers” data set, and as such, they are automatically matched on the field of study (psychology) and document type (conference or journal). With matched controls, we thus have $n = 1240$ total data points in the R_{FOS} data set, and we have $n = 2392$ total data points in the R_{QUERY} data set.

This leaves us with two data sets for which we have the following variables;

1. c_5: Number of citations five years after initial publication.
2. $TEAMSIZE$: The number of listed authors on a publication.
3. $YEAR$: Year of publication.
4. $CONDITION$: Replication (experiment) article or a non-replication (control) article.
5. $GROUP$: shared group ID for each matched replication and non-replication pair.

2.2 Bayesian modeling

We proceed to create statistical models based on the R_{FOS} and R_{QUERY} data sets. We approach Bayesian modeling as an iterative process, and follow best-practice guidelines detailed in Gelman et al. (2020). Inference follows a long process of model building, model checking, model validation, model understanding, and troubleshooting of computational problems. I attempt to report and justify the central choices made in the modeling effort here, but some parts of the complex process must necessarily be omitted for clarity. For an overview of computation, see section A.1.15.

2.2.1 Likelihood functions

As papers can be cited 0 times, 1 time, or 100 times but not 1.5 times, the outcome represents count data. The two most common ways of modeling count data use either (i) binomial regression or (ii) Poisson regression (McElreath, 2020, pp. 336). The binomial is insufficient in our case because we are not aiming at a binary classification, for instance into ”cited” and ”not cited”. Instead, we will use the Poisson model which allows us to push the modeling of counts beyond binary classification. The Poisson is really a binomial in which the number of trials (N) is very large, and the probability of success on each trial (p) is very low. The expected value of the binomial distribution is Np and the variance is $Np(1-p)$. When N is very large and p is very small these quantities are roughly the same (McElreath, 2020, pp. 330–360). Assuming mean and variance to be identical, a Poisson distribution can be generated from just one parameter describing its shape (McElreath, 2020, pp. 360).

$$y_i \sim Poisson(\lambda)$$

The parameter λ is at the same time the expected value of the outcome (y) and the expected variance of counts (y_i). In order to ensure that the values of λ_i are always positive (as counts must be) we need to specify a log link function (McElreath, 2020, pp. 361).

$$y_i \sim Poisson(\lambda_i)$$
$$log(\lambda_i) = \alpha + \beta(x_i - \bar{x})$$

The Poisson model can model data without a known theoretical maximum (McElreath, 2020, pp. 336). This seems to be exactly what we need to model our outcome (c_5), since the number of citations a paper can receive during the first five years after publication is unknown. However, there are several issues that often emerge with real-world data. A typical issue is that the variance of counts is often not equal to the mean of counts. This can happen in cases where additional sources of variance lead to ”overdispersion”, which is cases where the variance is greater than the mean (McElreath, 2020, pp. 387). Another typical issue is cases of excess zeroes, which can happen when more than one process generates the observed outcome data. In this case, we will use more than one likelihood function (a mixture model) to model our outcome (McElreath, 2020, pp. 390). Before applying any models to our data, we already know that our outcome (c_5) is heavily overdispersed. For the R_{FOS} data set we observe $mean = 14.05$ and $variance = 1426.72$. We also observe that our data contains a lot of papers that do not receive any citations in the first five years following publication. Cases of $c_5 = 0$ make up 22.98% of the entire R_{FOS} data set. The statistics for the R_{QUERY} data set are qualitatively similar. In this case, we observe $mean = 18.67$, $variance = 3902.62$ and 17.73% of studies with $c_5 = 0$.

Motivated by the fact that we observe overdispersion and a high fraction of zeros in our data, we test three candidate likelihood functions in order to identify a good model fit to the underlying data.

Candidate Model 1 (CM1) is a zero-inflated Poisson (ZIP) model. This is a mixture model in which we model the data using a combination of a Poisson model and a zero-generating model. This should model our outcome data well if there is both a zero-generating process and a process that generates Poisson distributed counts. The likelihood is specified below and follows the notation of McElreath (2020, pp. 392).

$$
\begin{aligned}
y_i &\sim ZIP(p_i, \lambda_i) \\
logit(p_i) &= \alpha_p + \beta_p x_i \\
log(\lambda_i) &= \alpha_\lambda + \beta_\lambda x_i
\end{aligned}
$$

The λ parameter still describes the shape of the Poisson distribution. The argument p acts as a switch that controls whether we generate values from our Poisson distribution or whether we generate zeroes (not from the Poisson distribution). p is a probability (hence the logit link function) and controls the ”mixing” of the two likelihood functions. Notice that the Poisson process can also

generate zeroes, but we are here mixing the Poisson with an additional zero-generating process to inflate the number.

Candidate Model 2 (CM2) is a negative-binomial model (also called gamma-Poisson). This model extends on the Poisson model and assumes that each Poisson count has its own rate, described by a gamma distribution (McElreath, 2020, pp. 387). This should model our data well if the process generating our outcome (c_5) contains various sources of variance which lead to overdispersed counts, but no separate zero-generating process.

Candidate Model 3 (CM3) is a zero-inflated negative binomial model. This model combines the two former approaches, and should model the data well if there is both a separate zero-generating process and a process that generates overdispersed counts.

The gamma-Poisson (or negative binomial) model ($CM2$) samples well for models conditioned on both the R_{FOS} and the R_{QUERY} data sets. In both cases, we observe an absence of divergent transitions, all $\hat{R}$ diagnostics around 1, and a high number of effective samples for all parameters. These are the key sampling statistics used to evaluate Markov Chain Monte Carlo (MCMC) sampling (Gelman et al., 2020). The chains do not show auto-correlation (see figure 1) and the only potential issue is the presence of some high Pareto-k values, which indicate that we have influential data points. Both the Zero-Inflated Poisson ($CM1$) and the Zero-Inflated Negative Binomial ($CM3$) show various sampling issues across models conditioned on both the R_{FOS} and the R_{QUERY} data sets. Problems include some $\hat{R}$ diagnostics above 1.05, a low number of effective samples ($n < 200$) for some parameters, auto-correlation of MCMC chains, and a much greater number of high Pareto-k values (see section A.1.8 for more details). On this basis, we proceed to use the gamma-Poisson model ($CM2$) for the remainder of our modeling effort.

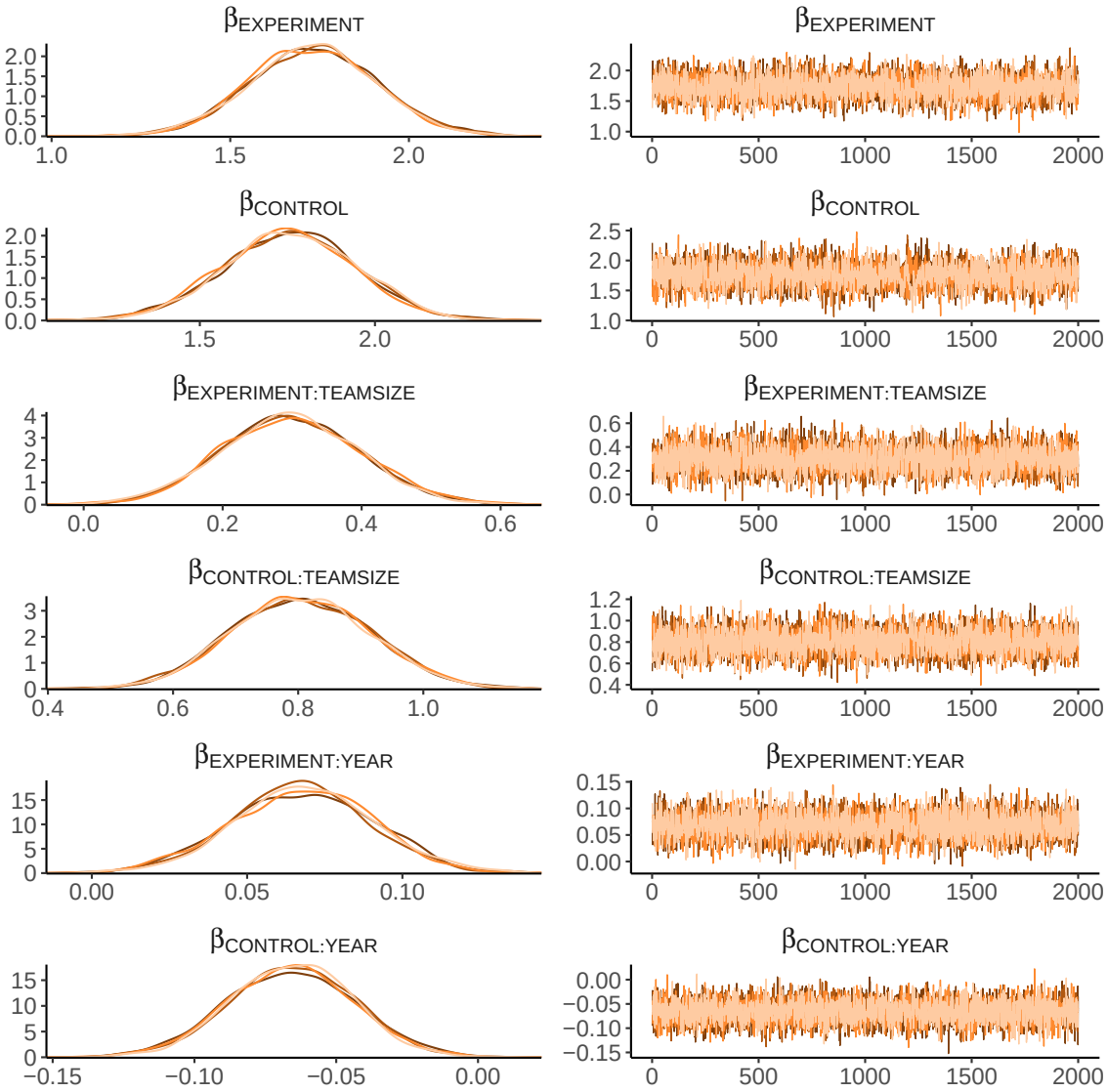

Figure 1: Density and trace plots for four sampled Markov Chain Monte Carlo (MCMC) chains. Sampling from gamma-Poisson model conditoned on R_{FOS} data. Only sampling of population effects (β) shown. Density plots (left) show consistent parameter estimation, and trace plots (right) show good mixing and exhibit no auto-correlation.

2.2.2 Model Specification

The model that we implement is specified below, along with the priors used for various parameters. The model is the same for both the R_{FOS} and the R_{QUERY} data sets. We use the $NB()$ notation, since the model is conceptualized as negative-binomial in `brms` (Bürkner, 2017).

$$
\begin{aligned}
c_{5i} &\sim NB(\mu_i, \phi_i) \\
log(\mu_i) &= \beta_{E:GROUP_i} + \beta_{C:GROUP_i} \\
&\quad + \beta_{E:GROUP_i:TEAMSIZE} + \beta_{C:GROUP_i:TEAMSIZE} \\
&\quad + \beta_{E:GROUP_i:YEAR} + \beta_{C:GROUP_i:YEAR} \\
\begin{bmatrix} \beta_{E:GROUP} \\ \beta_{C:GROUP} \end{bmatrix} &= MVNormal\left(\begin{bmatrix} \beta_E \\ \beta_C \end{bmatrix}, S\right) \\
S &= \begin{pmatrix} \sigma_{\beta_E} & 0 \\ 0 & \sigma_{\beta_C} \end{pmatrix} R \begin{pmatrix} \sigma_{\beta_E} & 0 \\ 0 & \sigma_C \end{pmatrix} \\
\beta_E &\sim Normal(log(10), 0.5) \\
\beta_C &\sim Normal(log(10), 0.5) \\
\beta_{E:TEAMSIZE} &\sim Normal(0.5, 0.5) \\
\beta_{C:TEAMSIZE} &\sim Normal(0.5, 0.5) \\
\beta_{E:YEAR} &\sim Normal(0, 0.5) \\
\beta_{C:YEAR} &\sim Normal(0, 0.5) \\
\sigma_{\beta_E} &\sim Exponential(1) \\
\sigma_{\beta_C} &\sim Exponential(1) \\
R &\sim LKJcorr(5) \\
\phi &\sim Exponential(0.5)
\end{aligned}
$$

First, let us unpack the data that we use to condition our models.

1. c_5: This is the response variable. It is the number of citations a scientific publication has received in the first five years following publication.

2. $TEAMSIZE$: This is the log of the number of authors on a scientific publication. This means that the minimum value for this predictor is 0, since a publication needs at least 1 author, and $log(1) = 0$.

3. $YEAR$: This is the year of publication after 2005. Since 2005 is the first year of inclusion into our sample, we set $2005 = 0, 2006 = 1, \ldots, 2015 = 10$.

4. $GROUP$: Each replication (experiment) study is matched with a non-replication (control) study. Each group is assigned a unique $GROUP$ ID.

5. $CONDITION$: There are two groups in our data. The experiment group (E) consists of replication studies and the control group (C) consists of matched non-replication studies.

2.2.3 Population-level (fixed) effects

We now discuss the parameters of our model and the priors that we place on those parameters. We have six population-level (or fixed) effects in the model defined above. These are all the parameters that we refer to as beta (β) values. First, we get an intercept for each group in our data, which is labeled as β_E and β_C above. The intercept for each group models the expected c_5 when predictors $YEAR$ and $TEAMSIZE$ are set to their lowest value ($YEAR = 0$ and $TEAMSIZE = 0$). This corresponds to a solo-authored paper published in 2005. We specify the prior as a normal distribution with $\mu = log(10)$ and $sd = 0.5$ for both groups (β_E and β_C). This means that we expect around $c_5 = 10$ for each group, but with considerable uncertainty. We also have β values for the interaction of each group with $TEAMSIZE$ ($\beta_{E:TEAMSIZE}$ and $\beta_{C:TEAMSIZE}$) and $YEAR$ ($\beta_{E:YEAR}$ and $\beta_{C:YEAR}$). For the interaction with $TEAMSIZE$, we specify the prior as a normal distribution with $\mu = 0.5$ and $sd = 0.5$ which only slightly nudges the model to believe that the number of authors of a paper is positively associated with the number of citations (c_5). We know that larger teams are associated with greater reference popularity (Fortunato et al., 2018; Wu et al., 2019) and we *should* incorporate this prior knowledge into the model. Importantly, we specify the same prior for each group (experiment, control), so even though we bias the model to expect an overall positive association between $TEAMSIZE$ and c_5 we do not bias it to expect the association to differ by group, prior to conditioning on data. For the interaction with $YEAR$, we specify the prior as a normal distribution with $\mu = 0$ and $sd = 0.5$. Since we have no strong reason to expect a clear positive or negative relation between $YEAR$ and c_5, our prior expectation is that there is no effect for either group.

2.2.4 Standard deviations of group-level (random) effects

In most cases, multilevel models deserve to be standard practice, since they provide a much better trade-off between underfitting and overfitting by partially pooling estimates between groups of observations (McElreath, 2020, pp. 15). This type of model is achieved through the specification and estimation of group-level (random) effects. We model a random intercept for each group (β_E and β_C) and thus allow studies to vary within the model. We must place a standard deviation on the random intercept, which indicates how much we think that the studies within each category vary. These are the parameters σ_{β_E} and σ_{β_C}. The prior for both of these parameters is an exponential distribution with $\lambda = 1$ which is a common choice that assigns most probability mass to small deviations and is very broad (McElreath, 2020, throughout). Additionally, we model these random intercepts as nested within our matched ID groups ($GROUP$). Recall that each replication study is matched with a non-replication study. If this matching captures any variance, e.g. that highly cited replication studies are generally matched with highly cited non-replication studies, then we should model this variance. This is what we do with the R parameter, which is modeled with an $LKJ()$ distribution for which we use $\eta = 5$. The default in `brms` is $LKJ(1)$ which is uniform and assigns equal probability to all correlations. For values of $\eta > 1$ extreme correlations become less likely, and for $0 < \eta < 1$ extreme correlations become more likely (Kurz, 2021, Chapter 14). Since we do not expect extreme correlations between $GROUP$ pairs, we set $\eta = 5$ which assigns more prior probability mass to small correlations and follows previous practice (Fusaroli et al., 2021). In sum, this means that we treat each condition β_E and β_C as distinct and pool estimates towards a mean within each condition rather than across conditions. However, we do exploit the fact that our matched ID groups should capture some covariance.

2.2.5 Family specific effects

There is one family-specific parameter in our model, which is related to the likelihood function used. The negative binomial has two parameters; a μ parameter, which models the mean of the data, and a shape (ϕ) parameter which is specific to this model family. The ϕ parameter controls the shape of the distribution, and thus allows us to properly model data with overdispersion. We show how the shape (ϕ) parameter affects the distribution in figure 2B. In the plot, we fix the mean of the distribution at $\mu = 10$, which is consistent with our $INTERCEPT$ prior, and we vary the shape (ϕ) parameter.

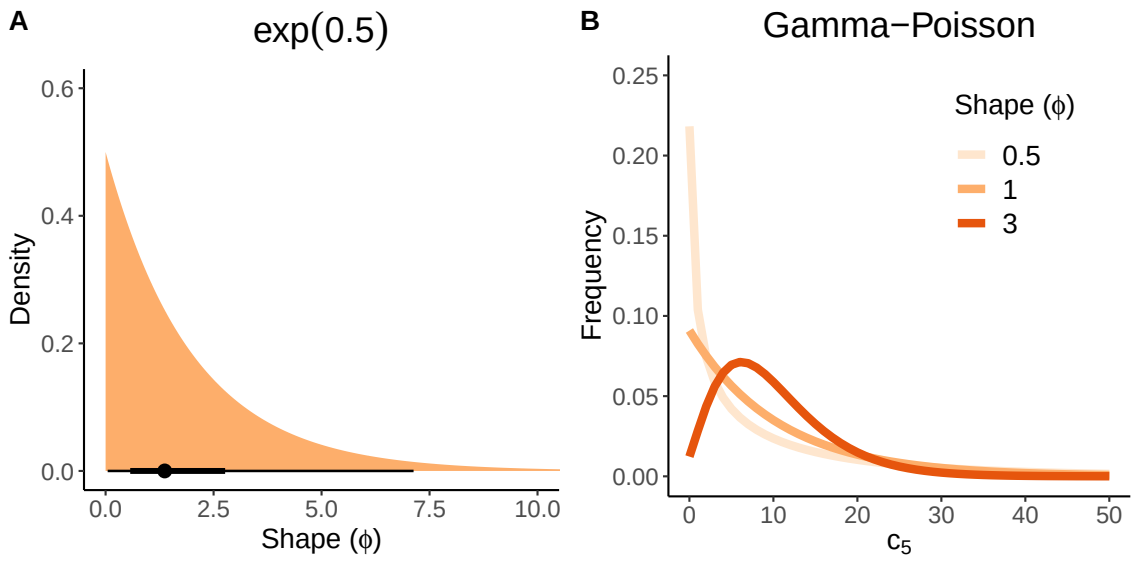

Figure 2: **A)** exponential prior with $\lambda = 0.5$ for the ϕ parameter of the negative-binomial distribution which is what we specify in our models. **B)** keeping $\mu = 10$ constant, we show how different ϕ values impact the distribution from the negative binomial likelihood. Low values of ϕ imply greater overdispersion.

The `brms` default prior for the shape (ϕ) parameter is a $\gamma(0.01, 0.01)$ distribution. This prior is rather extreme, and a prior distributed as $Exponential(0.5)$ was found to sample more effectively and converge better. Our choice is depicted in figure 2A. For a comparison of the $Exponential(0.5)$ and $\gamma(0.01, 0.01)$ prior for the shape (ϕ) parameter, see figure 21 in section A.1.9.

2.2.6 Prior Sensitivity

To test how sensitive our posterior inference is to the influence of our priors, we systematically vary the standard deviation on our population-level priors and monitor the effect on posterior parameter estimation. We want to have skeptical priors that are not so strong that the model cannot update following conditioning on data, but at the same time, we will slightly bias the model towards *not* finding an effect. As such, if the model does indeed estimate effects to be significantly different between groups, we can be more confident in the result.

As covered in the last section, the standard deviations on our prior distributions for all of our group-level effects are 0.5. To test the sensitivity of our posterior results to these priors we fit models with standard deviations on our population-level parameters ranging from 0.1 to 1.5 (step = 0.1) while keeping μ constant. The approach is the same as what has been advocated elsewhere (Cox, 2022). We generate estimates of group differences for population-level effects (β) and 95% credibility intervals (CI) using the "hypothesis" function from `brms` (Bürkner, 2017) and display this in figure 3.

Note that our priors are weakly informative and only slightly constrain the data. In the extent to which our priors do constrain the posterior estimation of parameters, they will constrain the posterior in the direction of finding no group difference. The same plot for the prior sensitivity of the model conditioned on the R_{QUERY} data set is provided in figure 22 in section A.1.10.

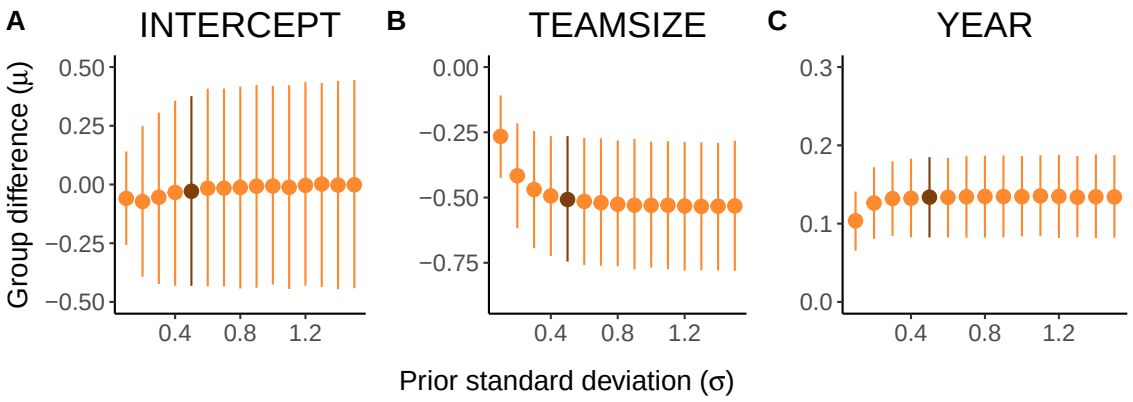

Figure 3: Prior sensitivity checks for the model conditioned on the R_{FOS} data set. The figure shows the posterior difference in population-level effects between groups (experiment, control) tested over a grid of prior standard deviation choices. In all cases, the point is the estimated difference between groups and lines are 95% credibility intervals. The dark dot and line highlights the standard deviation (σ) used in our main model and the light dots and lines show how posterior inference is affected by varying choices of standard deviation priors. This tests the sensitivity of our posterior inference to the choice of prior. **A)** $INTERCEPT$ difference between groups. **B)** $TEAMSIZE$ difference between groups (interaction effect). **C)** $YEAR$ difference between groups (interaction effect).

2.2.7 Prior and Posterior Predictive Checks

Prior and posterior predictive checks are important tools to understand (i) whether priors are unreasonable given our knowledge of the data and (ii) whether the model (posterior) has appropriately captured key patterns in the data (Gelman et al., 2020). The prior predictive check generates predictions before conditioning the model on data and the posterior predictive check generates predictions after conditioning the model on data. The prior predictive check happens before model fitting and the posterior predictive check is used to evaluate a fitted model (Gelman et al., 2020). We report them together in this section, although they inform the modeling process at different points. We mainly focus on posterior predictive checks here, as we validated that our priors do not unreasonably affect inference in the previous section. For the posterior, we want to ensure that (i) the distribution of posterior predictions visually matches the distribution of the observed outcome data and that (ii) the posterior predictions provide a reasonable estimate of key parts of the observed data, including the mean, the median, the fraction of zeroes, and the maximum value. For both prior- and posterior predictive plots, we follow the same strategy of generating 50 predictions (light orange) and overlaying the observed c_5 outcome data (dark orange). 50 is a reasonable number in this case because the check is not quantitative (or significance-based) but a visual inspection which is less effective with a large number of draws.

In figure 4 we show prior- and posterior predictive checks for the negative-binomial model conditioned on the R_{FOS} data set (for R_{QUERY}, see figure 23 in section A.1.11). Prior predictive checks (figure 4A) show that our priors are on a reasonable scale as compared to the observed data, but that our priors are also very broad. This is fine since we only want to constrain the model slightly. We observe that the posterior predictions generated by the model do not systematically diverge from the observed distribution of data, but they are also not identical to the observed distribution (figure 4B). Visually, the predicted distribution is a good fit to the data, and the model does not appear to have overfitted to the sample, in which case the predicted distributions should match the idiosyncrasies (i.e. small ”bumps”) in the observed distribution. Our model has captured the mean well (figure 4C), and our predictions for the median are also okay, although the predicted median tends slightly lower than the observed median (figure 4D). Most draws from the posterior generate slightly fewer zeroes than what is present in the actual outcome distribution, but again what we observe is only a slight bias which is not generally worrying (figure 4E). Most draws from our posterior predictive distribution predict a maximum value below the maximum c_5 observed in the real data, but we also observe some draws produce much higher maximum values (figure 4F). This is again reasonable,

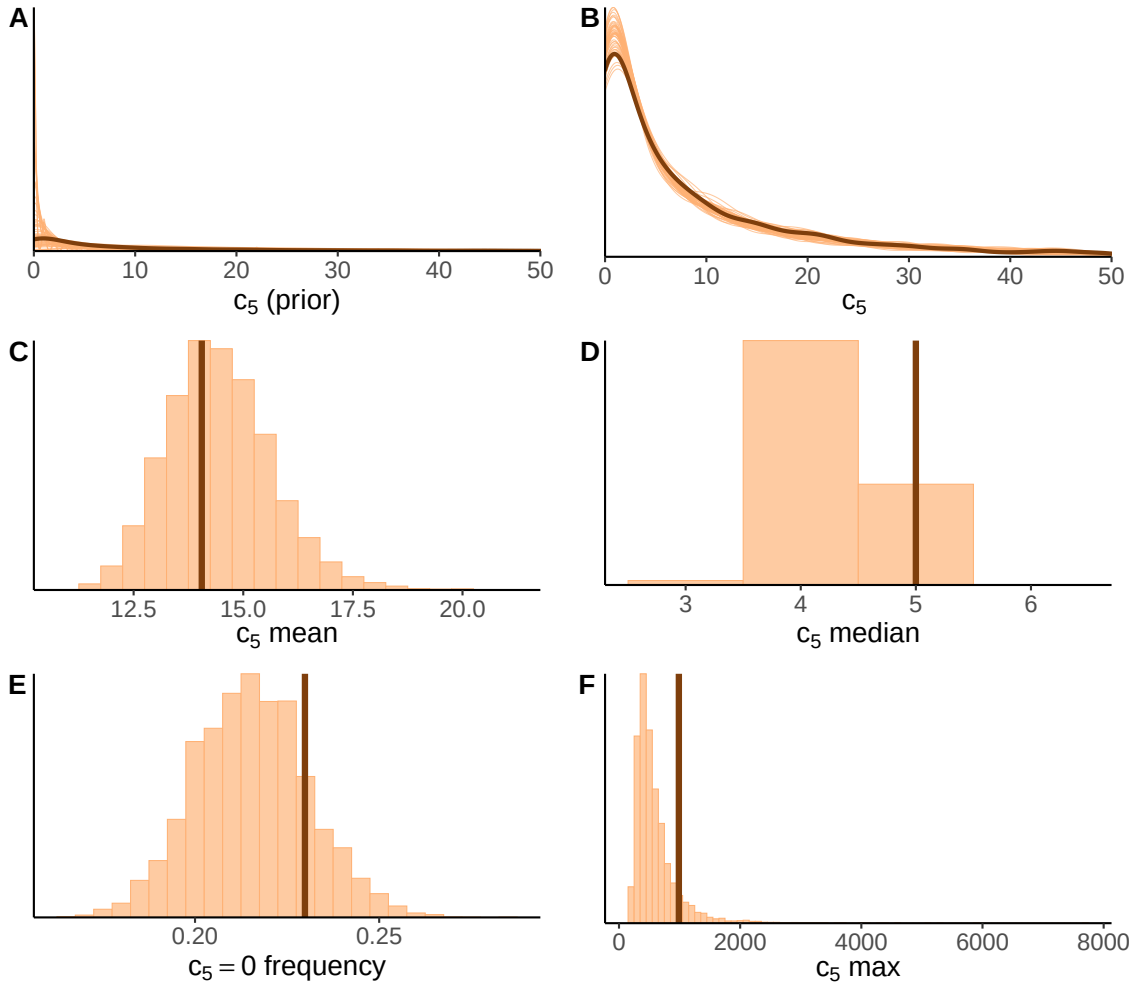

Figure 4: Prior and posterior predictive checks for the model conditioned on R_{FOS} data. In all cases, dark orange is the distribution (or value) of the observed data and light orange is samples from posterior (or prior) distributions. **A)** Prior predictive distributions with x axis cutoff at $c_5 = 50$ with observed distribution overlaid. **B)** Posterior predictive distributions with x axis cutoff at $c_5 = 50$ with observed distribution overlaid. **C)** Posterior predictive distribution of the mean with observed mean highlighted **D)** Posterior predictive distribution of the median with observed median highlighted. **E)** Posterior predictive distribution of the fraction of zeroes with observed fraction overlaid. **F)** Posterior predictive distribution of the maximum value with the observed maximum value overlaid.

and the highest maximum value generated ($c_5 \approx 6000$) is within a reasonable order of magnitude in terms of the number of citations a very successful paper can receive within five years. The model predictions should not match the actual maximum value observed in our data. The posterior predictions show that given the uncertainty in our parameters, $c_5 \approx 6000$ is around as high as we would expect any study in a similar sample to get, and we would expect that in most samples, the maximum c_5 will be slightly lower than what is observed in our particular sample.

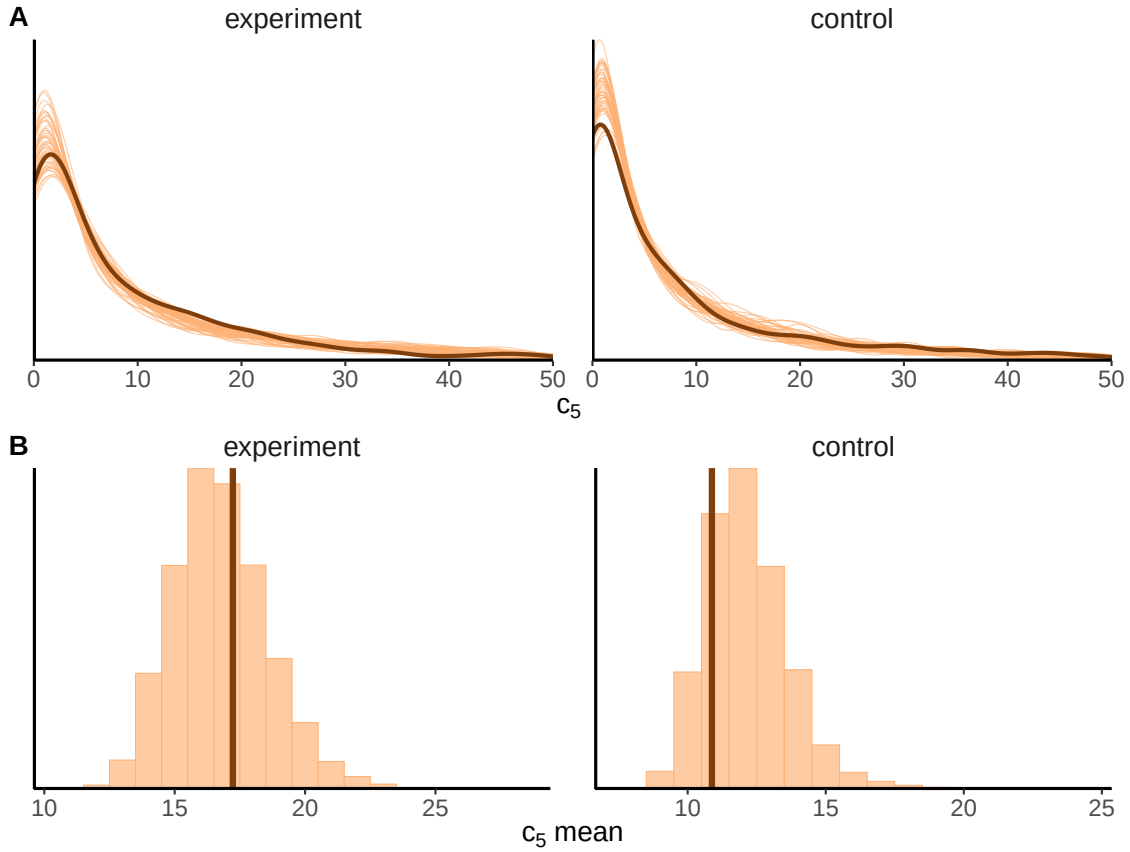

Figure 5: Grouped posterior predictive checks for the model condition on R_{FOS} data. **A)** Density plot of 50 posterior predictive draws (light orange) and the observed posterior distribution overlaid (dark orange). **B)** Histogram showing the mean of 50 posterior predictive draws (light orange) and line showing of mean for the observed data (dark orange).

Additionally, we conduct posterior predictive checks by group. These are shown for the model conditioned on the R_{FOS} data set in figure 5 (for R_{QUERY}, see figure 24 in section A.1.11). This

can reveal whether the model systematically misrepresents one group in the data. The distribution of c_5 is well recovered for both groups in our data (figure 5A). The mean is also well recovered for the replication (experiment) group, but the model appears to slightly overestimate the mean of the non-replication (control) group (figure 5B). The model inferences are conservative in the sense that we clearly regularize the expected replication (experiment) mean down and the non-replication (control) mean up, which will result in post-modeling inference suggesting that the groups are more similar than what the raw data suggests.

2.2.8 Updating Checks

A useful tool to assess whether priors are too restrictive is the prior-posterior updating check (Fusaroli et al., 2021; Cox, 2022), where we overlay the prior distributions for our parameters with the estimated posterior distributions. If our priors are reasonable, and there is enough signal in the data for the model to update properly, we should observe two things. First, we should observe that the posterior distributions have lower variance than the prior distributions (Fusaroli et al., 2021), which indicates that after conditioning on data the model has learned and become more certain. Second, we will not want the priors to be too restrictive, which means that we want our posterior distributions to be well within the range of values covered by our prior distributions (Fusaroli et al., 2021). We include updating checks for all parameters, including both our population-level (fixed) effects, our group-level (random) effects standard deviation (σ), the correlation of random effects, and the family-specific shape parameter (ϕ).

In figure 6 we show updating checks for our six population-level parameters (i.e. $INTERCEPT$, $TEAMSIZE$ and $YEAR$ for both conditions). In all cases, we note that the posterior distributions are much narrower (lower variance) than the prior distributions. We also note that the posterior distributions are not on the very edge of the prior distributions, although posterior density has shifted to the low end of our prior distribution for the $INTERCEPT$ parameter. A similar plot for the model conditioned on the R_{QUERY} data set is shown in section A.1.12, and updating checks for group-level effects and the family-specific parameter are also shown there.

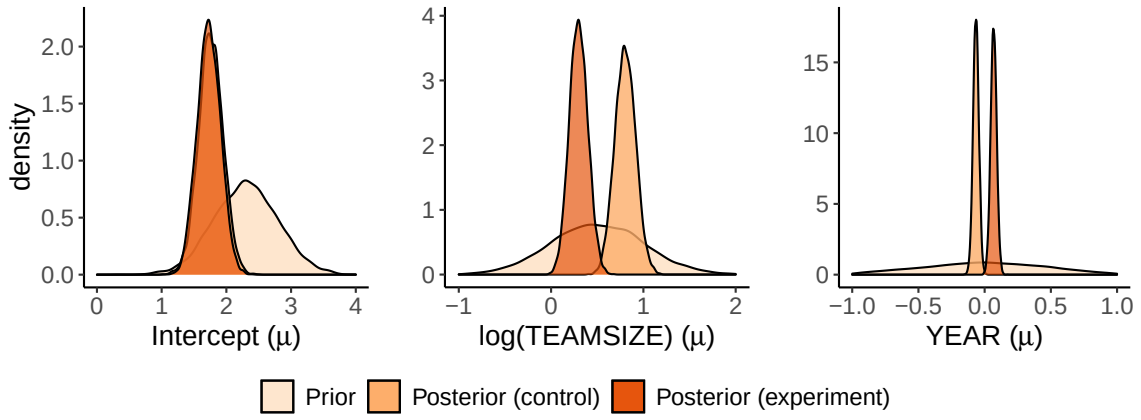

Figure 6: Prior-posterior updating checks for the population-level effects of the model conditioned on R_{FOS} data. Since the prior for each group is the same, we plot this only once. **A)** Updating check for the $INTERCEPT$ parameter. **B)** Updating check for $TEAMSIZE$ parameter. **C)** Updating check for the $YEAR$ parameter.

2.2.9 Research Question Evaluation

We evaluate our three research questions for this part of the thesis from two angles. Firstly, we test whether population-level effects are significantly different by group using the Evidence Ratio (ER) metric. ER is a ratio of evidence in favor of a hypothesis, such as $a > b$. If $ER = 1$ this indicates that there is an equal posterior probability for $a > b$ and $a < b$. If $ER > 1$ this indicates more posterior probability in favor of our hypothesis ($a > b$) and if $ER < 1$ then this indicates more posterior probability against our hypothesis (Bürkner, n.d.). Just as with the conventional $p < .05$ significance threshold, any Evidence Ratio (ER) significance threshold will be arbitrary. $ER > 1000$ has been used previously as providing very strong evidence in favor of a hypothesis (Fusaroli et al., 2021) and we follow this reporting convention here. Besides hypothesis testing at ER > 1000 threshold, we report point estimates and 95% credibility intervals (CIs) associated with our population-level effects.

Although we do report significance tests, our main goal is to assess whether the effects are meaningful. A challenge to interpretability is that the parameters of our model are not on the natural outcome scale of citations. Recall that we (i) specified a log-link function in our model and (ii) modeled $log(TEAMSIZE)$ rather than $TEAMSIZE$. This means that the direct point esti-

mates from our model will not be meaningful, and we cannot directly extrapolate interaction effects to the natural outcome scale. Critically, equally separated values of predictors (e.g. $YEAR$) can be associated with non-equal changes in the expected outcome of c_5 on the natural scale. For the $INTERCEPT$ the conversion back to the natural outcome scale is not problematic, and we can simply exponentiate the estimated $INTERCEPT$ parameter for each group. For the interaction effects, we create meaningful contrasts and hypothesis test these contrasts on the natural outcome scale of the data. For all hypothesis testing of population-level effects we rely on the "hypothesis" function from `brms` (Bürkner, 2017) which allows non-linear hypothesis testing based on custom contrasts (Bürkner, n.d.).

In addition to conducting hypothesis testing of population-level effects, we simulate data for unobserved levels in our model. This incorporates all levels of uncertainty (e.g. random effects variation) and allow us to investigate outlier behavior. Technically, we retain the observed combinations of $TEAMSIZE$, $YEAR$ and $CONDITION$ from our raw data but assign new ID values. As such we instruct the model to predict c_5 for unobserved groups of studies. We generate 8000 predictions for all levels to ensure robust estimation and compare the predicted distributions for each condition in our data.

3 S1: Results

In this section, we evaluate results pertaining to our main research question (RQ_1) and our two sub-questions ($RQ_{1.1}$ and $RQ_{1.2}$). Before we address the results from our modeling effort, we provide an overview of (i) the volume and growth of the "Reform Psychology" literature and (ii) the structure of our data through Exploratory Data Analysis (EDA).

3.1 "Reform Psychology" Volume

We do not end up reporting models conditioned on the "Open Science" (OS_{FOS}) or the "Reproducibility" ($R*_{FOS}$) data sets (see sections 2.1, A.1.4, and A.1.5). However, growth in the volume of publications that are associated with "Reform Psychology" practice is of interest to us. Thus, we report on the number of records that we match in the "Open Science" and "Reproducibility" data sets, as well as the two "Replication" data sets that we subsequently model (R_{FOS} and R_{QUERY}).

In figure 7A we plot the number of publications in each data set from 2005 until 2020 (including both full years). As we have already noted, the volume is largest for the two "Replication" data sets that we end up modeling (R_{FOS} and R_{QUERY}). We note that there is an increasing trend in volume for both of these. The volume of publications in the "Reproducibility" data set ($R*_{FOS}$) is rather stable throughout. The "Open Science" data set (OS_{FOS}) is increasing rapidly after 2015 but has low absolute volume throughout. In figure 7B we plot the percent-wise change relative to our base year 2005. It is apparent that the number of records that the MAG classifies as both "Open Science" and "Psychology" has expanded rapidly.

The overall scientific record is growing roughly exponentially (Szalay & Gray, 2006), and since the plots in figure 7A-B ignore this baseline expansion, they do not tell us anything about the relative growth of sub-literatures related to the "Reform Psychology" agenda. In 7C we address this, by plotting the fraction of papers for each of our sub-categories as a fraction of the overall psychology literature. This shows whether the sub-fields associated with "Open Science", "Reproducibility" and "Replication" are growing relative to the overall "Psychology" literature that they are embedded within. Both of the "Replication" data sets have been increasing relative to the overall volume of the "Psychology" literature (at least since 2010), while the "Reproducibility" literature appears to be diminishing relative to the overall "Psychology" literature.

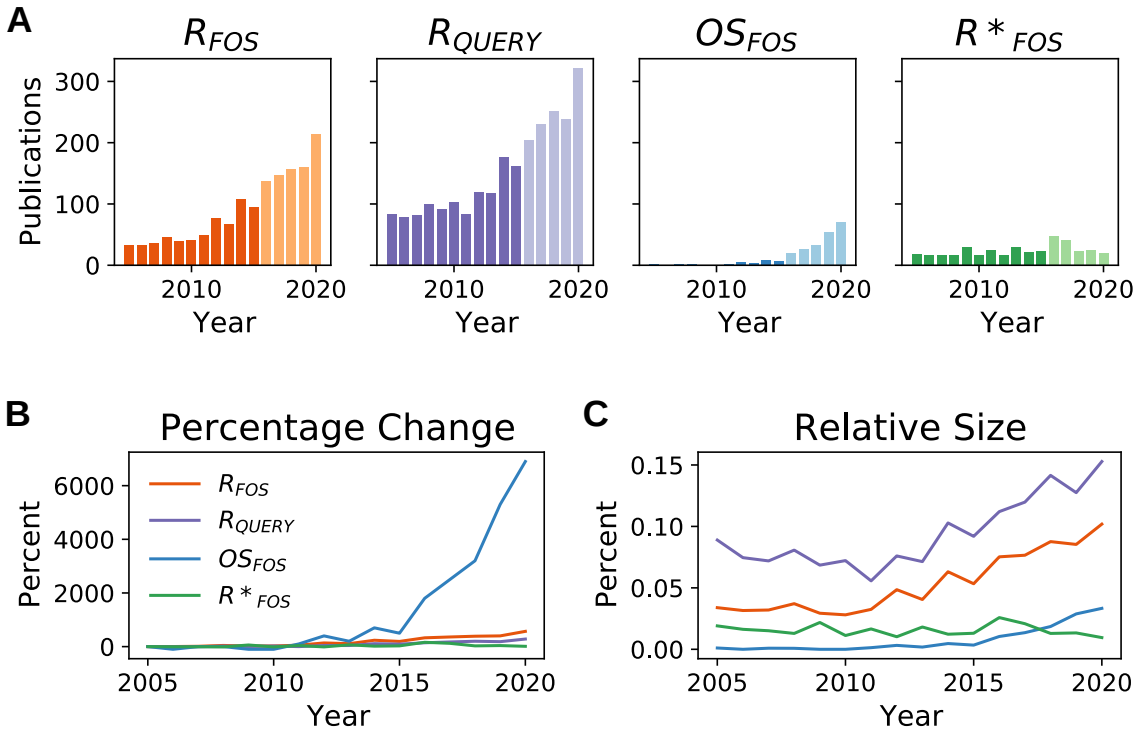

Figure 7: Investigating the volume of publications in the "Reform Psychology" sub-fields (R_{FOS}, R_{QUERY}, OS_{FOS} and $R*_{FOS}$) over time. **A)** plots the number of publication year-over-year for each category. Dark bars highlight the years of inclusion in our data sets for modeling (i.e. prior to 2016). **B)** plots the percentage-wise change relative to 2005 for each category. **C)** plots the relative size of each category as a percentage of the psychology literature (as categorized by the MAG).

3.2 EDA & Summary Statistics

We now focus on the two data sets for which we report models (R_{FOS} and R_{QUERY}). As discussed in the methods section, our outcome is the number of citations a paper accrues during the first five years following publication (c_5). Many scientific publications are never cited, and very few publications receive a disproportionate amount of citations (Piwowar et al., 2018). This is evident from figure 8 below, where we plot the outcome (c_5) distribution of the R_{FOS} data set grouped by category (experiment, control). The overall pattern is very similar for R_{QUERY} (see figure 28 in section A.1.13). For R_{FOS}, we observe a higher mean, median and maximum c_5 for replication studies (mean: 17.23, median: 5, max: 988) than for non-replication studies (mean: 10.88, median: 4, max: 153). However, we also observe a much higher standard deviation (σ) for the replication group ($sd = 50.11$) than for the non-replication group ($sd = 18.02$), suggesting that the much higher mean is driven at least partly by outliers. Finally, we observe a larger fraction of studies that are never cited ($c_5 = 0$) for non-replication studies (29.68%) than for replication studies (16.29%).

We plot the relationship between c_5 and $YEAR$ in figure 9A. The difference in slope is difficult to appreciate from the plot. However, the slope is slightly steeper for replication studies than for non-replication studies, showing that in our raw data, there is a slight tendency for replication studies to become relatively more cited over time as compared to non-replication studies.

We plot the relationship between c_5 and $TEAMSIZE$ in figure 9B and between c_5 and $log(TEAMSIZE)$ in figure 9C. In both cases, we observe a positive slope for each group, which is consistent with previous research on the relationship between team size and citation rates (Wu et al., 2019). The choice to model $log(TEAMSIZE)$ has clear implications for the association. For the plot with raw $TEAMSIZE$ (figure 9B), we observe a steeper slope for the replication group than for the matched controls, whereas we observe a slight association in the opposite direction for the plot with $log(TEAMSIZE)$ (figure 9C). The plot affirms the decision to log-transform the variable since the linear association in figure 9B is clearly driven by very few outlier studies, whereas the opposite (and much weaker) association in figure 9C is driven by the bulk of the data. Figure 29 in SI section $A.1.13$ shows a similar plot for the R_{QUERY} data.

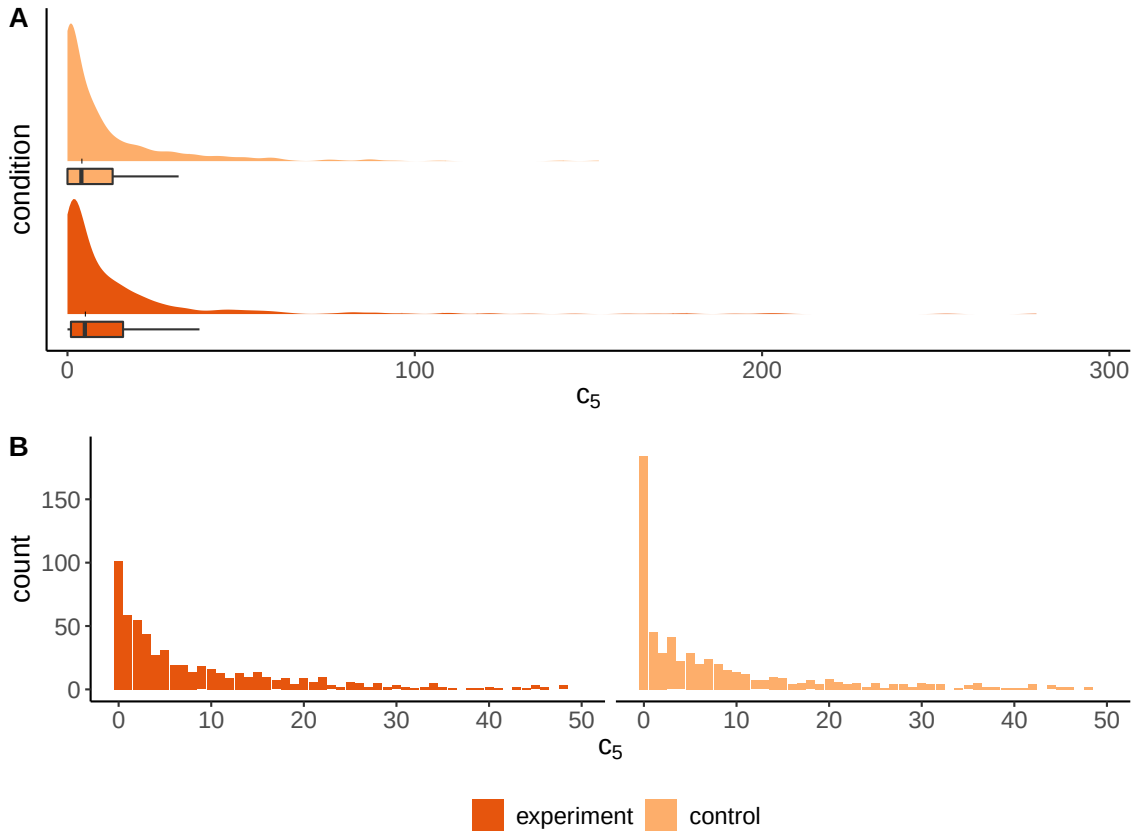

Figure 8: Visualization of the outcome distribution (c_5) for both replication studies (experiment) and non-replication studies (control). **A)** density plots and box-plots for each category. In order for the plot to be readable, we omit all studies with $c_5 > 300$ ($n = 1$). **B)** distributions for all studies in our data with $c_5 <= 50$ ($> 99\%$ of studies).

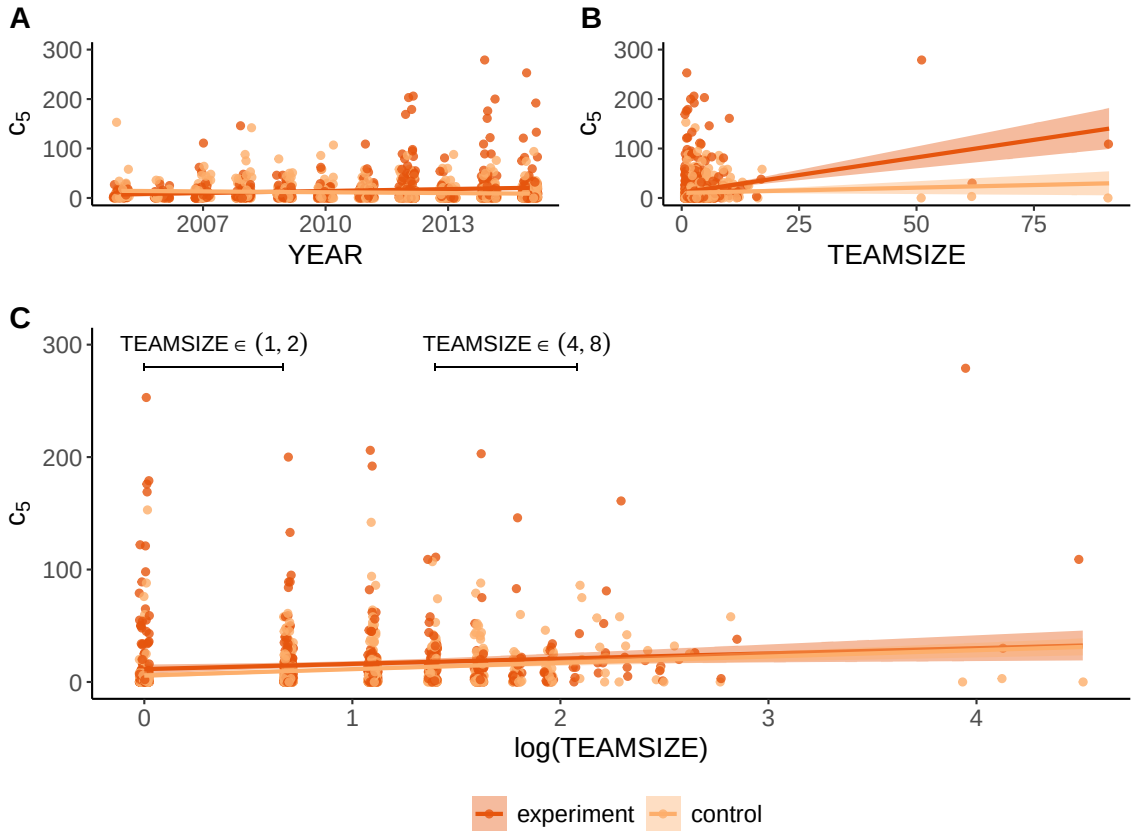

Figure 9: In all plots, we display raw data points and linear fit (including shaded 95% confidence intervals) for the association between our predictors and outcome. All plots use the R_{FOS} data and limit the y-axis at $c_5 = 300$, excluding $n = 1$ outlier study for visual purposes. **A)** scatterplot and linear fit between year of publication ($YEAR$) and outcome (c_5). **B)** scatterplot and linear fit between team size ($TEAMSIZE$) and c_5. **C)** scatterplot and linear fit between the $log(TEAMSIZE)$ and c_5. We emphasize the fact that for the log-transformed variable, the distance between a solo-authored paper and a duo-authored paper, and the distance between 4 and 8 authors is the same.

3.3 Modeling Citations

3.3.1 Citation Difference (RQ_1)

We now evaluate our main research question (RQ_1), which is whether replication studies are cited less than non-replication studies. The intercept might seem like a good candidate to approach this. Although the intercept has a meaningful interpretation (expected c_5 for a solo-authored paper in 2005), this does not directly address the research question that we pose. We address RQ_1 by constructing a more informative intercept and then proceed to simulate synthetic data from our model.

We compute the expected c_5 for each group when the predictors $YEAR$ and $TEAMSIZE$ are at their *mean* values and then compute the group difference. This is more informative with regards to average population-level group-differences than the original intercept which corresponds to *minimum* values of our predictors. For the R_{FOS} data set, the mean number of authors is 2.48 and the mean year is 2011.24. Based on our model, the expected population-level c_5 for a replication study conditional on mean values of our predictors is $11.35(9.79, 13.01)$ and the expected population-level c_5 for a non-replication study is $8.04(6.89, 9.28)$. We formulate the group difference as a non-linear hypothesis (Bürkner, n.d.) using `brms` (Bürkner, 2017). The estimated group difference is significant (ER > 1000) and the estimated group difference in citation gain is $3.31(1.61, 5.08)$. This means that given mean values for $TEAMSIZE$ and $YEAR$, the model expects replication studies to receive 3.31 citations more than non-replication studies. The corresponding c_5 distributions for both replication and non-replication studies are shown in figure 10A.

An issue with the evaluation above is that it does not tell us anything about outliers in the data because it only uses fitted population-level effects and as such does not incorporate the uncertainty in the random effects which accounts for outliers. In other words, the population-level comparison focus on the difference between average studies in each group. Citations to individual scientific works follow a heavy-tailed distribution (D. Wang et al., 2013), meaning that very few articles receive a large fraction of citations. Outliers in our case are impactful scientific articles, which means that an additional investigation is warranted. To address outlier behavior we generate predictions from our model which incorporate all levels of uncertainty, including random-effects variance. We instruct the model to compute predictions for the original studies in our sample, but we assign new ID-values, which instructs the model to predict c_5 values for ”unobserved” groups. In order to achieve robust estimates, we generate 8000 predictions for each original data point in our data set,

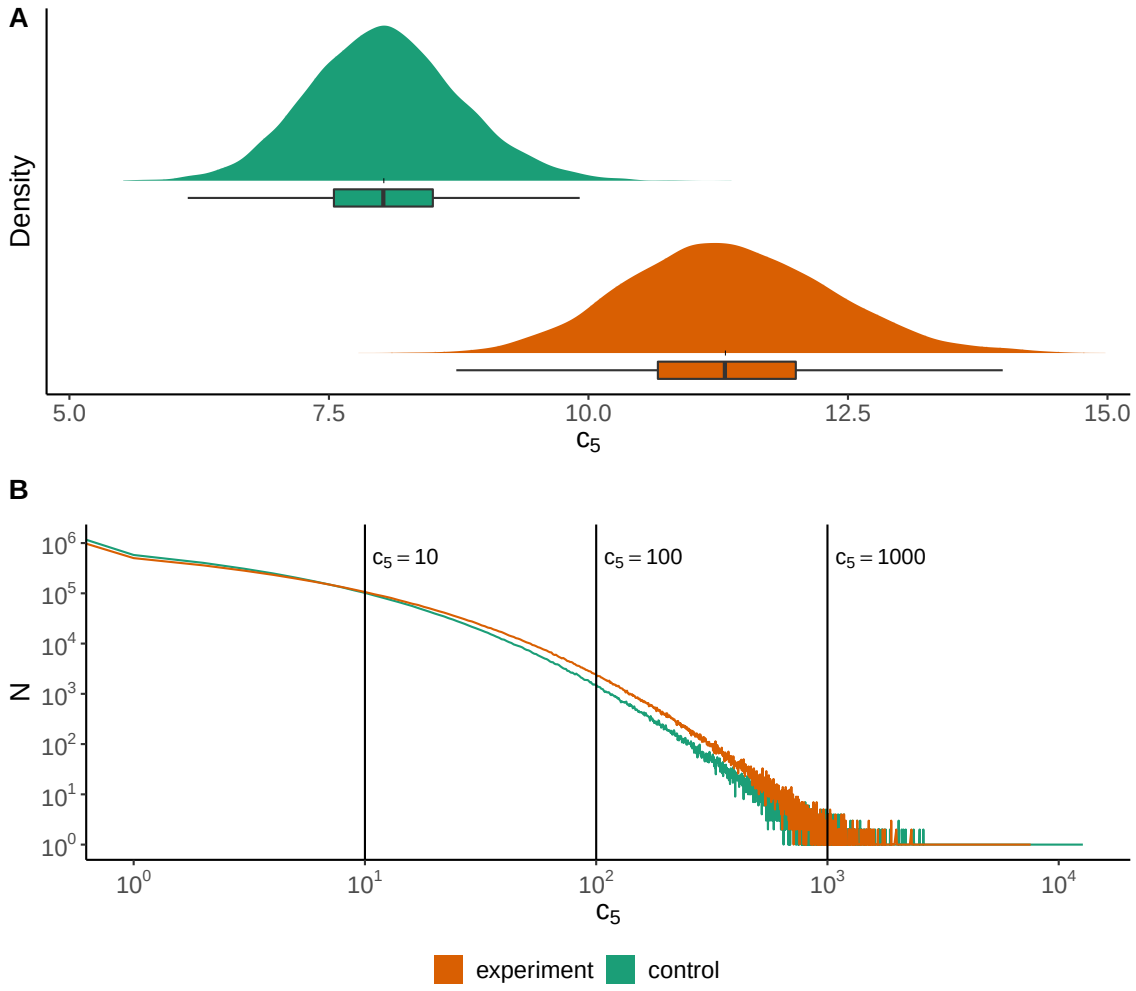

Figure 10: Conditional effects and predictions for the model conditioned on R_{FOS} data. **A)**: Density of posterior probability for each group (replication, non-replication) conditional on mean values of $TEAMSIZE$ (2.48) and $YEAR$ (2011.24). The distributions of expected c_5 values only overlap slightly, with significantly more probability mass consistent with the hypothesis that replication (experiment) studies are cited more than non-replication (control) studies. **B)** predictions of unobserved groups in our data plotted on a power-grid. In total 9.920.000 predictions. The distributions cross at $c_5 \approx 10$ showing that more non-replication (control) studies are predicted to have $c_5 < 10$ and more replication (experiment) studies are predicted to have $c_5 > 10$.

which results in a total of $9.920.000$ predictions for the model which is conditioned on R_{FOS} data. We do not conduct significance tests in this case, but rather visualize the distribution of simulated data in figure 10B. Because the citation distributions are highly skewed, with high density around low values of c_5 and very highly cited articles, we visualize the distributions on a power grid. The two distributions cross at around $c_5 = 10$ showing that non-replication studies are predicted to be denser below $c_5 = 10$ and replication studies denser above $c_5 = 10$. To make the implications more concrete, we define two desirable properties; (i) the probability that a scientific paper becomes "visible" ($c_5 > 0$) and (ii) the probability that a scientific paper becomes a "hit" ($c_5 > 100$). In both cases, we observe that replication studies outperform non-replication studies. The probability that a non-replication study becomes "visible" is 76.53% while the probability that a replication study becomes "visible" is 80.41%. The probability that a non-replication study becomes a "hit" is 1.55% while the probability that a replication study becomes a "hit" is 2.72%, almost twice as high.

3.3.2 Mechanism ($RQ_{1.1}$, $RQ_{1.2}$)

In an attempt to understand the mechanisms which is driving this effect, we investigate the interactions with $YEAR$ ($RQ_{1.1}$) and $TEAMSIZE$ ($RQ_{1.2}$). Since we have used a log-link in our model, we cannot directly back-transform interaction effects to the natural outcome scale of citations. This is because the interactions are non-linear on the natural outcome scale. This is shown in figure 11 which visualizes the conditional effect of $YEAR$ and $TEAMSIZE$ for both groups in our data.

Instead, we construct contrasts that we can evaluate on the natural outcome scale of the data. We compute the expected difference in citations between a study published in 2015 and a study published in 2005. Replication studies receive more citations in 2015 as compared to in 2005 (difference: 5.57, 95% CIs: $[2.46, 9.1]$, ER < 1000). The effect is not significant at $ER > 1000$ threshold, but the effect of $5.57(2.46, 9.1)$ additional expected citations is clearly meaningful and 95% CIs do not cross zero. Non-replication studies receive less citations in 2015 as compared to 2005 (difference: -2.86, 95% CIs $[-1.19, -4.76]$, ER > 1000). The effect is significant at ER > 1000 threshold, although the point estimate of the effect is smaller than the positive effect observed for replication studies. What we really care about is not the effects for each group, so much as the group difference. The group difference for the YEAR interaction is significant (difference: 8.42, 95% CIs: $[4.81, 12.5]$, ER > 1000) and the effect is rather large. The point estimate of

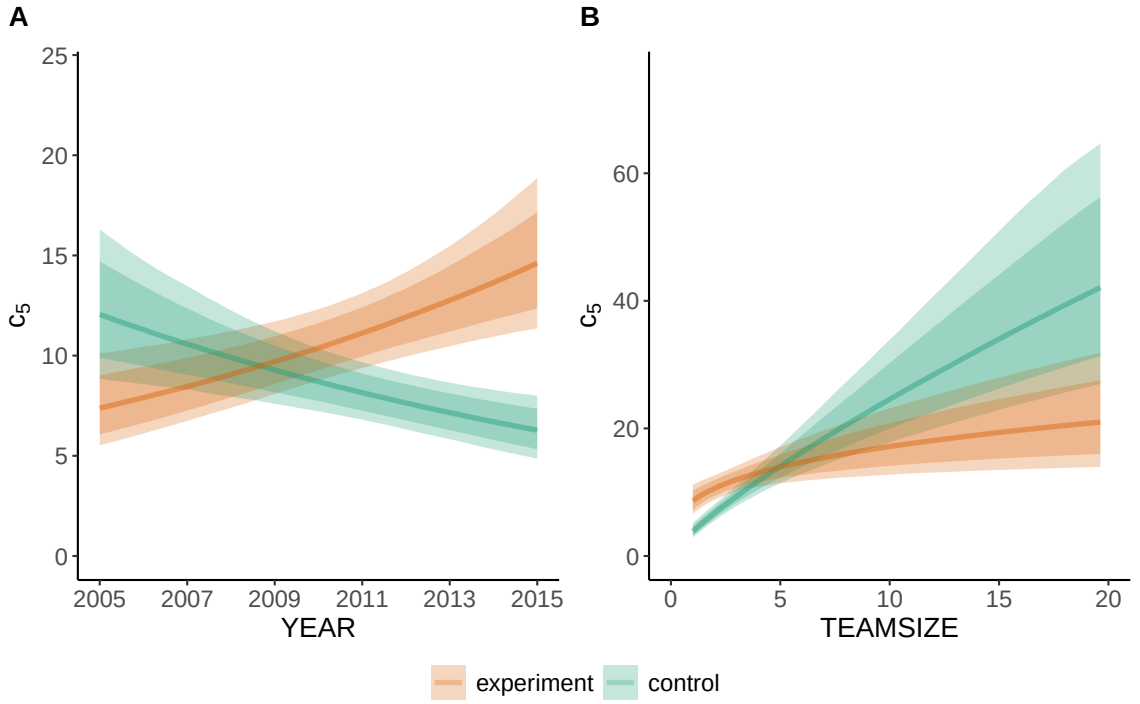

Figure 11: Conditional effects of the two interactions in our model for the model conditioned on R_{FOS} data. **A)** Keeping $TEAMSIZE$ constant at mean value (2.48) we generate expected population-level c_5 for both groups conditional on a grid of $YEAR$ values. **B)** Keeping $YEAR$ constant at mean value (2011.24) we generate expected population-level c_5 for both groups conditional on a grid of $TEAMSIZE$ values. Notice that in this case we only visualize expected population-level effects up to $TEAMSIZE = 20$. In both cases we observe clear interaction effects between groups.

$8.42(4.01, 12.5)$ means that the effect of 10 years (from 2005 to 2015) is associated with an expected 8.42 additional citations for replication studies as compared to non-replication studies. The $YEAR$ interaction is displayed in figure 11A.

For the interaction effect with $TEAMSIZE$, the relevant contrast is not as obvious as with the $YEAR$ interaction. Of course, we can compare the expected number of citations for the largest number of authors in our data set (91 authors) with the smallest number of authors in our data set (1 author). However, we match very few papers written by large scientific teams (see figure 9B-C), and less than 1% of papers in our data set have more than 20 authors. This motivates us to focus on $TEAMSIZE$ contrasts between smaller teams, and we evaluate two contrasts here; (i) we evaluate the expected citation difference for studies with 2 authors and 5 authors and (ii) we evaluate the expected citation difference for studies with 5 authors and 20 authors. The contrasts are intended to evaluate the difference between expected citations (c_5) for small, medium, and large teams in psychology. The $TEAMSIZE$ interaction is shown in figure 11B. The expected increase in citations associated with a medium-sized study (5 authors) as compared to a small study (2 authors) is significant for both replication studies (difference: 2.23, 95% CIs: [0.92, 3.84], ER $>$ 1000) and for non-replication studies (difference: 11.22, 95% CIs: [7.43, 15.83], ER $>$ 1000). This is not surprising given the EDA reported earlier (figure 9B-C) and the well-established finding that larger scientific teams are associated with higher citation rates (Wu et al., 2019). More interestingly, the group-difference for this contrast is also significant (difference: 8.99, 95% CIs: [4.89, 13.69], ER $>$ 1000). For the contrast between large teams (20 authors) and medium-size teams (5 authors) we find that large teams are associated with significantly higher c_5 than medium-sized teams for both replication studies (difference: 5.03, 95% CIs: [1.65, 9.85], ER $>$ 1000) and for non-replication studies (difference: 45.59, 95% CIs: [24.4, 75.24], $ER >$ 1000). The group difference for this contrast is also significant (difference: 40.56, 95% CIs: [18.784, 70.23], ER $>$ 1000). In sum, the citation advantage for larger teams is found to be more pronounced for non-replication studies than for replication studies.

3.3.3 Robustness

As a robustness check, we report all contrasts and predictions for the model conditioned on R_{QUERY} data. Conditional on mean values of predictors, the expected number of citations is higher for both conditions in this data set. For replication (experiment) studies, the expected c_5 is $14.24(13.05, 15.48)$ while for the matched control studies the expected c_5 is $9.66(8.69, 10.68)$.

The significant interaction is replicated in this case (difference: 4.58, 95% CIs: [3.25, 5.93], ER > 1000). In this model, replication studies are expected to receive $4.58(3.25, 5.93)$ citations (c_5) more than non-replication studies conditional on mean values of our predictors (see figure 12A). The mean values, in this case, are different, as they are based on a different data set. Specifically the mean number of authors is slightly higher (3.15) and the mean year is slightly lower (2010.77).

We also generate predictions based on the parameters estimated in the model conditioned on R_{QUERY} data. The power-axes plot of the distributions for each group is shown in figure 12B, which is visually very similar to the plot that we generated based on predictions from the model conditioned on the R_{FOS} data set (figure 10B). The two distributions still appear to cross at $c_5 \approx 10$ and the separation between the two distributions is even clearer in this case. We replicate the finding that the probability of being "visible" ($c_5 > 0$) is higher for replication studies (86.85%) than for non-replication studies (82.97%). We also replicate the finding that the probability of a scientific "hit" ($c_5 > 100$) is around twice as high for replication studies (3.89%) than for non-replication studies (2.06%). As such, we replicate all findings related to our main research question of whether replication studies are cited less than non-replication studies (RQ_1). In both cases, we find clear evidence that replication studies are cited *more* than non-replication studies.

We do not replicate the group differences for interactions with $TEAMSIZE$ and $YEAR$ with the model conditioned on R_{QUERY} data. In this data set, studies published in 2015 are associated with fewer citations than studies published in 2005 for both conditions. For replication studies, the effect is $-1.23(-3.02, 0.31)$ and for non-replication studies the effect is $-1.32(-2.39, -0.32)$ but none of these effects are significant at ER > 1000 threshold. The group contrast is also not significant (difference: 0, 95% CIs: $[-1.96, 1.9]$, ER: < 1000), and as such we do not replicate the interaction effect with $YEAR$.

Consistent with results from the model conditioned on R_{FOS} data, medium-sized teams (5 authors) are associated with more citations than small teams (2 authors) in the model conditioned on R_{QUERY} data. There is an expected citation gain for replication studies (difference: 9.35, 95% CIs: [7.34, 11.57], ER > 1000) as well as for non-replication studies (difference: 8.15, 95% CIs: [6.47, 10.04], ER > 1000). The interaction difference is not significant (difference: 1.19, 95% CIs: $[-1.53, 3.97]$, ER < 1000). Large teams (20 authors) are associated with more citations than medium-sized teams (5 authors). There is an expected citation gain for replication studies (difference: 30.14, CIs: [21.07, 40.79], ER > 1000) as well as for non-replication studies (difference:

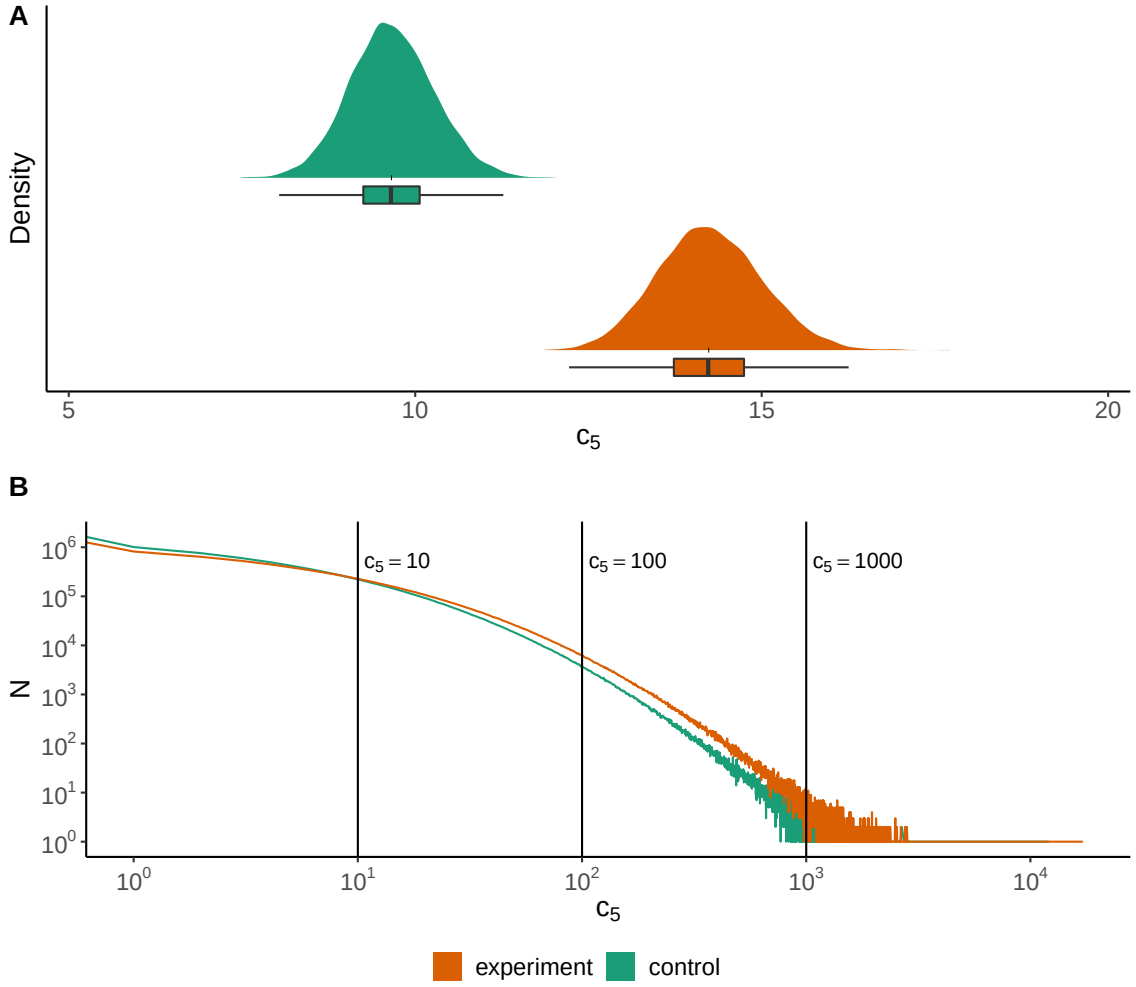

Figure 12: Conditional effects and predictions for the model conditioned on R_{QUERY} data. **A)**: Density of posterior probability for each group (replication, non-replication) conditional on mean values of $TEAMSIZE$ (3.15) and $YEAR$ (2010.77). The distribution of expected c_5 values is higher for replication (experiment) studies than for non-replication (control) studies. **B)** predictions of unobserved groups in our data plotted on a power-grid. In total 19.136.000 predictions. The distributions cross at $c_5 \approx 10$ showing that more non-replication (control) studies are predicted to have $c_5 < 10$ and more replication (experiment) studies are predicted to have $c_5 > 10$.

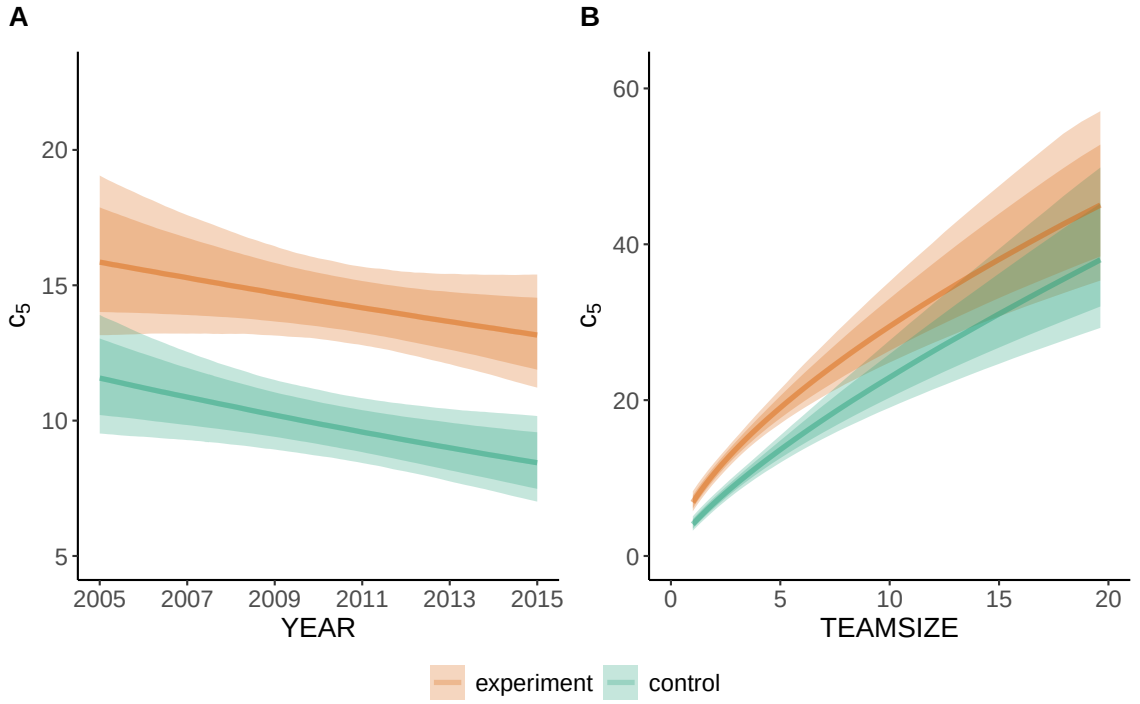

Figure 13: Conditional effects of the two interactions in our model for the model conditioned on R_{QUERY} data. **A)** Keeping $TEAMSIZE$ constant at mean value (3.15) we generate expected population-level c_5 for both groups conditional on various values of $YEAR$. **B)** Keeping $YEAR$ constant at mean value (2010.77) we generate expected population-level c_5 for both groups conditional on various values of $TEAMSIZE$. Notice that in this case we only visualize expected population-level effects up to $TEAMSIZE = 20$. We do not observe interaction effects in either case.

30.43, 95% CIs: [21.15, 41.74], ER > 1000). The interaction associated with this contrast is not significant (difference: 0.29, 95% CIs: $[-13.89, 14.65]$, ER < 1000). In sum, we find a strong positive association between $TEAMSIZE$ and c_5, which is expected, but we do not replicate the group difference.

4 S1: Discussion

Before we assess the evidence pertaining to our three research questions (RQ_1, $RQ_{1.1}$, $RQ_{1.2}$), we briefly comment on the volume of records in our data sets. We showed that the "Open Science" and "Replication" literature's have expanded at a rate greater than the overall "Psychology" literature. In particular, the "Replication" literature appears to accelerate relative to the overall "Psychology" literature from around 2012, and the "Open Science" literature appears to accelerate from around 2015. We do not observe growth in the "Reproducibility" literature relative to the overall growth in "Psychology", but as discussed elsewhere, the "Reproducibility" classification of the MAG skews heavily towards cognitive neuroscience, and as such does not quite capture the breadth of what we want "Reproducibility" to include (see section $A.1.5$ and specifically table 3). The results generally suggest an acceleration in the production of scientific work which can reasonably be associated with "Reform Psychology". The absolute values we report will be imprecise because neither the MAG labeling nor our query are exhaustive. Replication studies have been estimated to make up around 1% of the psychology literature (Makel et al., 2012) which is a much higher fraction than what we match. However, the fact that replication studies make up an increasing fraction of the total psychology literature is consistent with earlier findings (Makel et al., 2012), and the relative growth trends are likely to reflect actual developments.

4.1 Citation Difference (RQ_1)

Our main research question (RQ_1) asked whether replication studies are cited less than non-replication studies in psychology. We addressed this in two ways. First, we focused on the expected population-level effect conditional on mean values of predictors ($TEAMSIZE, YEAR$). For models conditioned on both R_{FOS} and R_{QUERY} data, we found a significant and meaningful difference between replication studies and matched controls (see figure 10A and figure 12A). For the model conditioned on R_{FOS} the expected citation difference was $3.31(1.61, 5.08)$ and for the model conditioned on R_{QUERY} the expected citation difference was $4.58(3.25, 5.93)$. In both cases, replication studies are expected to receive more citations than matched controls. This result focuses particularly on the population-level effect, and as such informs us about the expected citation difference for average studies in both conditions. Second, we generated predictions for unobserved groups in our data, incorporating all levels of uncertainty. We visualized the distributions of simulated c_5 values for the model conditioned on R_{FOS} data (figure 10B) and for the model con-

ditioned on R_{QUERY} data (figure 12B). In both cases, we found that non-replication studies were predicted to be denser for $c_5 < 10$ while replication studies were predicted to be denser for $c_5 > 10$. We reported the probability of a study being "visible" ($c_5 > 0$) and being a "hit" ($c_5 > 100$). For both models, we found that replication studies were more likely to become "visible" and almost twice as like to become "hits" as compared to non-replication studies. Taken together the results show that replication studies are cited *more* than non-replication studies, contrary to the direction in which we framed RQ_1, and seemingly contrary to the common notion that replication studies are disincentivized (Nosek et al., 2022). This is the case both for average studies (e.g. population-level effects) and for outliers (e.g. predictions). The consistency of this finding across two heterogeneous sets of data suggests that the effect is robust, and the observed differences are meaningful.

There are several potential mechanisms that might explain this difference, but we argue that the results might reflect that non-replication studies are more "risky" than replication studies. Replication studies often target recent original studies with high impact and high interest (e.g. Open Science Collaboration, 2015). In these cases, replication studies will be a relevant citation for the large number of scientists who engage with active and popular streams of research. We know that replication studies are often co-cited alongside original studies (Hardwicke et al., 2021), and when replication studies target high-impact research they might "latch on" to these lines of inquiry. In typical cases where studies are selected for replication based on impact this can lead to a situation in which replication studies primarily become associated with popular avenues of research, and as such receive more citations than the average original study. On the other hand, it will often be unclear prior to publication whether an original study will be perceived as relevant for future lines of inquiry of peers. For original work, it has been shown that studies that present novel ideas but fail to place them within the shared conceptual framework of peers are cited at low rates (Uzzi et al., 2013). In order for a study to become highly cited, it is not enough for the study to be great - it must also be grounded in shared knowledge which allows peers to incorporate it into their future work. In this sense, replication studies could be easier to digest and assimilate than at least some original work. Our operationalization of scientific impact (c_5) might exacerbate this effect since paradigm-changing discoveries are known to have limited early impact (D. Wang et al., 2013).

4.2 Interactions & Population Effects

Besides the proposed mechanisms related to our main research question (RQ_1) explored above, we also directly assess the effects of $YEAR$ ($RQ_{1.1}$) and $TEAMSIZE$ ($RQ_{1.2}$). We cannot draw strong conclusions based on the results in this part because we do not internally replicate our results. Based on the model conditioned on R_{FOS}, we found a significant interaction with $YEAR$, such that replication studies receive relatively more citations over time than non-replication studies. This was not replicated in the model conditioned on R_{QUERY} data. The same picture is true for the interaction with $TEAMSIZE$, in which case the model conditioned on the R_{FOS} data set showed a significant interaction effect, while the model conditioned on the R_{QUERY} data set failed to replicate this. Since the model conditioned on the R_{FOS} data set showed significant and meaningful interactions, we will briefly discuss plausible mechanisms.

Given the emphasis we have placed on large-scale replication studies, it is initially surprising that we find a stronger modulatory effect of $TEAMSIZE$ for non-replication studies than for replication studies. Recall that prior to modeling we observed an interaction in the opposite direction, such that the citation advantage for larger teams was more pronounced for replication studies than for non-replication studies (figure 9B for R_{FOS} and figure 29B for R_{QUERY}). There are several reasons why the model inferences diverge from the raw patterns observed prior to modeling. First, we log-transformed the variable following the logic in McElreath (2020, pp. 361–367). As we noted, this changes the association between $TEAMSIZE$ and our outcome. Additionally, we match very few large-scale studies, and the ones that we do match are partially-pooled towards the population mean, attenuating their effect on our inferences. As such, the outlier studies with a really large number of authors will not exhibit a strong influence on the parameters estimated by our model. Instead, the interaction difference is likely driven by differences between smaller studies ($TEAMSIZE < 20$) which dominate in our data sets. As such, our results do not suggest that large-scale replication studies are not appreciated by the scientific community, but rather that for smaller increments (e.g. 2 authors to 5 authors), there is a stronger citation boost for non-replication studies than for replication studies. We might speculate that there are different mechanisms which result in a large number of authors for replication studies as compared to original studies. Original studies are only likely to be conducted by large teams when the effect or paradigm necessitates this. In contrast, replication studies can have many co-authors even with simple paradigms, because replication studies often test effects with large sample sizes (higher effect size) and at multiple sites. Although this work is

important, it might not be perceived as equally ambitious compared to original work conducted by large teams of investigators.

The interaction between condition and $YEAR$, such that replication studies become relatively more cited than original studies over time, is less surprising. As we have discussed previously, there has been a great deal of interest in replication studies following the replication crisis, and it is not unreasonable that this could drive the effect. Of course, this still leaves different possibilities open. For instance, we do not know whether (i) psychologists have cited replication studies more in recent years because psychologists have better appreciated the importance of reproducibility and replicability, (ii) whether there are now just more (and better) replication studies to cite than there was previously, or (iii) whether the increase in citations for replication studies over time reflects pressure from the growing community of ”Reform Psychology”, and as such simply reflects psychologists paying ”lip-service”. In any case, the growing focus on replicability and reproducibility in psychological science following the replication crisis is likely to have resulted in more citations for replication studies.

4.3 Selection Bias

It is unlikely that our results are accidental or driven by database issues. Although both data sets suffer from idiosyncrasies and biases (see A.1.5 and A.1.6) this is unlikely to drive our main effect since (i) outliers are partially pooled and do not exert out-sized influence on our estimated parameters and (ii) the biases in the data sets are different and should not randomly lead to consistent results.

A better objection is that our effect could be due to the fact that we only model studies that actually got published. It has been suggested that editors apply lower standards to ”novel” research (Serra-Garcia & Gneezy, 2021), and they might apply particularly high standards for publishing non-novel research (Hummer et al., 2017) which replication studies are by definition. This could result in a selection bias where only outstanding replication studies are published. McElreath (2020, pp. 165–166) showed how a correlation between ”trustworthiness” (robustness) and ”newsworthiness” (novelty) can be induced by a selection in which reviewers care about both attributes, even though there is no correlation prior to conditioning on publication. Selection bias is a widespread issue for bibliometrics and scientometrics research which rely on citation databases (Piwowar et al., 2018). Each study should carefully consider whether it is indirectly conditioning on attributes such

as publication or availability.

We are able to partly address this challenge since we have access to preprints as well as to papers published at either conferences or in journals in the MAG. Nicely, if the same paper is published first as a preprint and then subsequently at either a conference or in a scientific journal, then these two documents are linked in the MAG database by an ID. This allows us to track the rate of studies that are initially published as preprints and subsequently as peer-reviewed publications. If there is a selection bias we would expect a lower rate of replication preprints to be published subsequently in peer-reviewed venues. We find that 28.57% of replication studies that are published as preprints are subsequently published in a journal or at a conference, while this is the case for 20.56% of non-replication studies. Although the number of replication preprints in psychology that we match is small ($n = 126$) the tentative result is in the opposite direction and does not suggest a selection effect.

4.4 Insights from an incomplete model

In both of our models, we observe high Pareto-k values, which indicate that we have influential observations (Vehtari et al., 2015). This is actually the same as saying that these observations are not well predicted by our model. The fact that our model is incomplete offers a way to learn something about the data-generating process that we are studying. The high Pareto-k observations are all studies with high c_5. It suggests that the predictors that are included in our model ($TEAMSIZE$, $YEAR$, $CONDITION$) are not sufficient to predict scientific "hits" ($c_5 > 100$). The model is not problematic in the sense that it is unable to produce outliers when we generate predictions for unobserved levels. However, the model cannot predict which individual observations will be impactful, and only accomplishes the generation of such outliers through substantial uncertainty in the random effects. It is intuitively reasonable that our model cannot predict impactful studies based on the included predictors, and I believe that any "static" model will be fundamentally limited in this regard. This is because each scientific publication has a life of its own, which is shaped by social dynamics, and where initial differences in citation rate can compound (D. Wang et al., 2013; Perc, 2014).

Consider the example of two of the most highly cited large-scale replication efforts in our data set, the "Many Labs" study by Klein et al. (2014), and the registered replication by Alogna et al. (2014). Both studies were published in 2014, and we plot their citation trajectories in the first five years

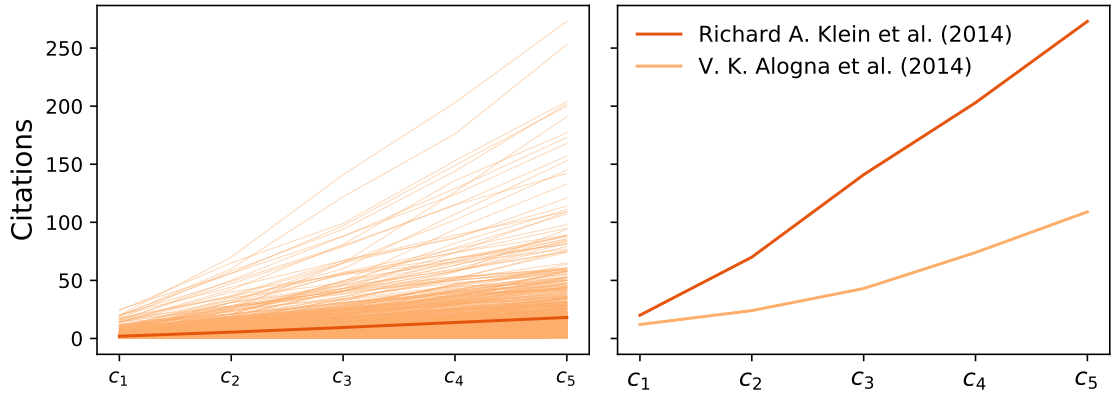

Figure 14: Left plot shows citation trajectory of all records in R_{FOS} that are cited ($n = 955$) excluding $n = 1$ outlier observation. The mean citation trajectory is highlighted. Right plot shows the citation trajectory of two large-scale replication studies which intuitively diverge over time.

following publication in figure 14 (right). The citation trajectories show that initial differences compound and this effect of "preferential attachment" for citations is well documented (D. Wang et al., 2013). It is reasonable that we should observe this pattern since more people are likely to hear about the more popular paper from a colleague or to encounter it in the list of references of other work. Platforms such as Google Scholar might further exacerbate this effect through algorithms which prioritize already popular research (Perc, 2014). In our example of this phenomenon, Alogna et al. (2014) receives $c_1 = 12$ citations while the Klein et al. (2014) paper receives $c_1 = 20$ citations. This small initial difference in traction compounds such that Klein et al. (2014) end up in our data set with $c_5 = 273$, while Alogna et al. (2014) end up in our data set with $c_5 = 109$. In efforts aiming to predict and explain the citation patterns of individual scientific works, dynamic models rather than static models could provide important insights. Although our model cannot reasonably predict the impact of individual scientific articles this does not fundamentally limit our ability to comment on the overall tendency that replication studies are cited at higher rates than non-replication studies.

5 S2: Methods

In this second part of the thesis, we set out to answer whether the ”Reform Psychology” movement is addressing issues of diversity and inclusion (RQ_2) as well as sub-questions related to dominant topics of discourse ($RQ_{2.1}$) and marginalization of important ”Theory Reformers” ($RQ_{2.2}$). In order to address these questions, we gather and analyze a large corpus of Twitter data.

The goal of the data curation is to collect a high-quality data set of tweets that is relevant to the ”Reform Psychology” agenda. This is not a trivial task, and there are multiple approaches one might have taken to accomplish this. Our approach was to identify two keyword searches (Open Science, Replication Crisis) that relate to the ”Reform Psychology” movement in different ways and match a non-trivial amount of tweets. The idea is then to use the intersection of accounts that appear in both ”Open Science” and ”Replication Crisis” as a plausible set of Twitter accounts that have a genuine stake in ”Reform Psychology”. For a discussion, see section A.2.5.

5.1 Scraping

We obtain access to the Twitter Academic API v2, which allows us to pull 10 million tweets each month, and to collect tweets all the way back to 2006 when the platform was launched. We use the `twarc2` command-line tool (“Twarc”, n.d.) to collect tweets from the Twitter API. We collect Tweets based on two queries. One query matches ”open science OR openscience” (Open Science) and one query matches ”replication crisis OR replicationcrisis” (Replication Crisis). In both cases, we retrieve tweets that match the queries regardless of capitalization. For each of these two queries, we collect all tweets from 2007 and until 2021, including both full years. This selection was made because 2007 is the first full year of data (Twitter was launched in 2006) and 2021 is the last full year of available data.

We gather all tweets that match our query terms. We gather both original tweets, retweets (RTs), quoted tweets (QTs), and replies. We extract all original tweets that match any of our search queries. For original tweets, replies, and QTs this means that one of our queries is matched in the text of the tweet. For replies and QTs, the tweet that is being replied to or quoted does not need to match our query term. We match the reply or the QT as long as their own text match one of our queries. The RT text consists of a handle (@xyz) followed by the text of the tweet that is being retweeted, so we match all RTs for which the original tweet uses one of our search terms. Because all RT text is

contained in our corpus of original tweets, replies, or QTs, we do not use RTs in semantic analysis, and only use this data to construct the network between users.

In most cases, the data that we collect from the Twitter API contains information about the relation between tweets. For instance, the data object might contain the information that tweet x is a retweet of tweet y. The only case in which we manually use mentions (i.e. the usage of @ to refer to another Twitter handle) is in the case of replies, since replies are often made to several other Twitter handles, and this information is not readily available in the data object that we obtain from the Twitter API. In this case, we manually create a connection between the handle that replies and each of the handles that it replies to. Besides this, we do not consider mentions and simply rely on the curated information from the Twitter API. This means that original tweets will never refer to any other tweet (or handle) in our data set, although they might mention another Twitter user (e.g. by using @ in their text).

5.2 Overview

As can be seen from the bottom plot in figure 15, the "Open Science" query matches more tweets than the "Replication Crisis" query. This is sensible because Open Science is a much broader (and older) community than the community around the current replication crisis. For the "Open Science" query, we match $2.003.614$ total tweets (618.134 tweets excluding RTs) from 343.688 unique accounts. For the "Replication Crisis" query we match 76.515 total tweets (34.756 tweets excluding RTs) from 40.350 unique accounts. As can be seen from figure 15, a substantial proportion of the tweets that appear in the "Open Science" and "Replication Crisis" data sets come from accounts which have tweets that appear in both of these data sets (the "overlap" category). Since "Replication Crisis" is the smaller category, a much larger fraction of accounts that have posted in "Replication Crisis" have also posted in "Open Science" ($\approx 40\%$) than the other way around ($\approx 5\%$). The fact that the set of accounts which appear in either corpus overlap to a large extent supports the notion that "Replication Crisis" and "Open Science" are strongly related. We argue that the intersection captures most accounts that are central to the "Reform Psychology" movement.

In the top-left part of figure 15, we plot the network of Twitter accounts that appear only in "Open Science" (green), only in "Replication Crisis" (purple), and users which appear in both data sets (orange). The network has been backboned (see 5.6) and additional weak edges have been removed for clarity. As is immediately obvious, the plot is completely dominated by accounts that only

appear in ”Open Science” and accounts that appear in both data sets. For further analysis, we will be working with only those accounts that overlap.

In order to give an impression of the types of accounts that we filter away, we highlight two of the densest non-overlap (only ”Open Science”) clusters in the network (figure 15, top right). One of these clusters is dominated by genomics and immonology (@Primary_Immune, @ResearchGenome, @AndGenomics, @CompChemBloBot) as well as accounts that focus on open source analysis tools (@PyData, @NumFOCUS). The other cluster is dominated by space exploration and observation, with the most important account being the European Space Agency (@EO_OPEN_SCIENCE). Both of these clusters do contain legitimate and coherent ”Open Science” communities, but since our focus here is on ”Reform Psychology”, it is reasonable to exclude tweets from these accounts from our corpus.

Limiting ourselves to accounts that have posted in both ”Open Science” and ”Replication Crisis” leaves us with 699.885 unique tweets (177.654 unique tweets excluding RTs) from 18.240 unique accounts. Since the data was found to be of varying quality still, we further restricted ourselves to only consider tweets from accounts with at least two original tweets in both the ”Open Science” data set and the ”Replication Crisis” data set. This leaves us with 118.932 unique tweets (64.621 unique tweets excluding RTs) from 989 unique handles. This is a much smaller data set, but the extra requirement does make it more plausible that a significant portion of the resulting tweets are from users who have a genuine stake in both ”Replication Crisis” and ”Open Science”, and as such in ”Reform Psychology”.

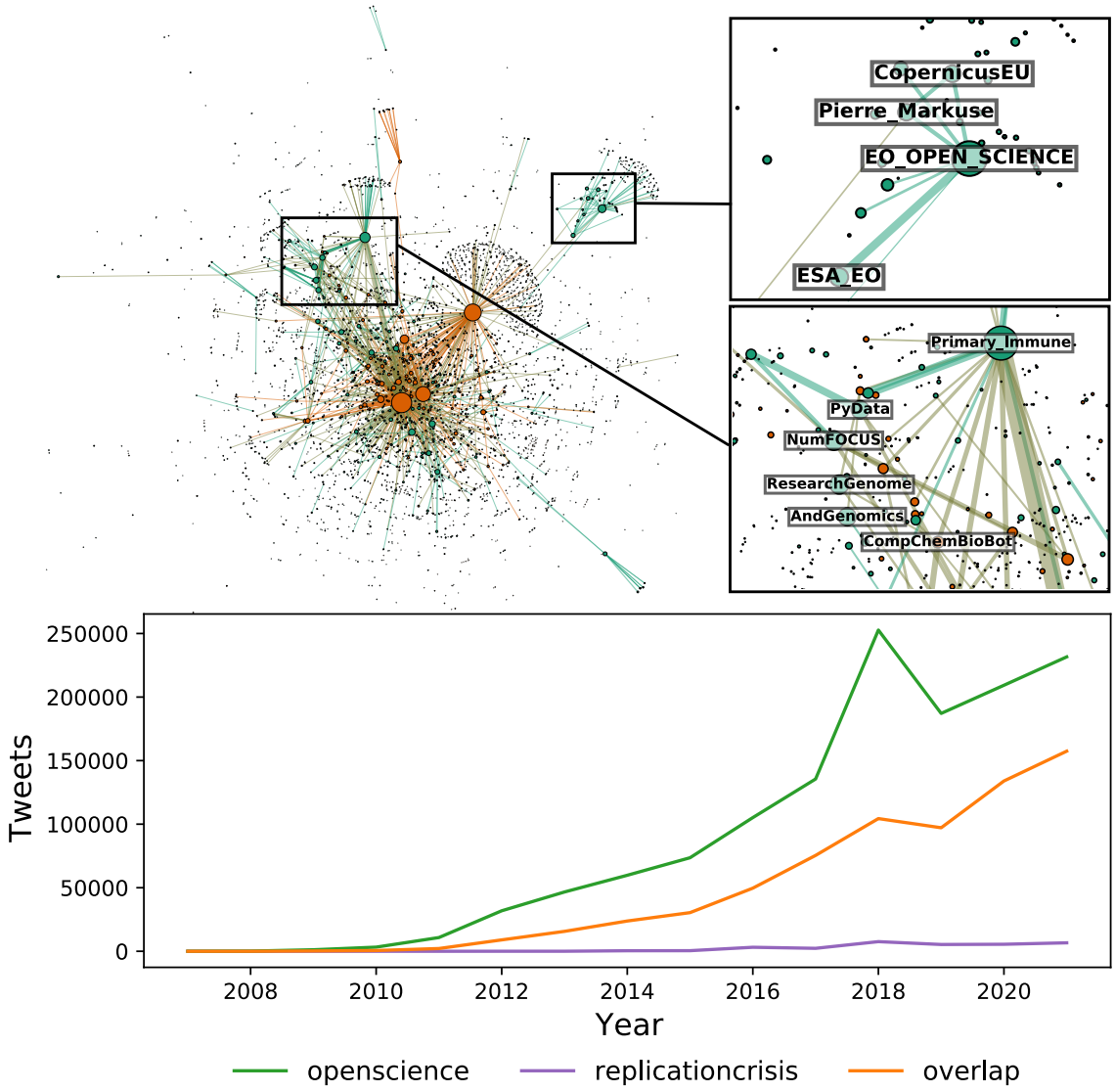

Figure 15: Top left: Network between users in both ”Open Science” and ”Replication Crisis” data sets. Node size scaled by weighted degree, and edge-weight based on number of interactions (QTs, RTs, replies). Network backboned using Disparity Filter (DF) algorithm ($c = 0.995$) and only the Giant Connected Component (GCC) displayed. Top right: Zoomed view of two dense clusters of mainly ”Open Science” discourse. Bottom: Volume of tweets from 2007 to 2021 (including both full years). In all plots, tweets are classified according to whether the author account appears only in ”Open Science”, only in ”Replication Crisis” or whether the author appears in both data sets.

5.3 Prosocial Language

To investigate the usage of prosocial language we use a list of 127 prosocial items (see A.2.1) which is adapted from Murphy et al. (2020). We are only concerned with whether a tweet matches at least one item in our list and not e.g. the rate of prosocial terms as compared to the total number of words used. Because of this, we do not care whether a tweet contains excess characters (e.g. hyperlink) and as such the only preprocessing conducted for this investigation is to lowercase everything.

In the first case, we compute the amount and fraction of tweets in the ”Open Science” and ”Replication Crisis” corpora which match at least one prosocial term. These are the full data sets before filtering, where we had 34.756 non-RT tweets in the ”Replication Crisis” corpus and 618.134 non-RT tweets in the ”Open Science” corpus. We further restrict this to only consider tweets that are labeled as containing English language. This results in 32.797 non-RT tweets in the ”Replication Crisis” corpus and 511.056 non-RT tweets in the ”Open Science” corpus. We also combine these two data sets, which results in 526.850 non-RT tweets. The number of tweets is not quite the sum of the ”Open Science” and ”Replication Crisis” corpora because a tweet can appear in both data sets if it matches both queries.

In the second case, we focus on our core ”Replication Crisis” corpora, which is the data set based on tweets from accounts that appear in both ”Replication Crisis” and ”Open Science” with at least two original tweets. Recall that this selection resulted in 64.621 unique non-RT tweets. Since our list of prosocial terms is English, we again restrict ourselves to tweets that contain English language. This results in a total of 59.369 tweets from 984 unique accounts. We investigate the evolution of prosocial language over time by plotting the proportion of tweets that contain prosocial language year-over-year.

5.4 Topic Modeling & Linkage Network

5.4.1 Preprocessing

We preprocess all English non-RTs (original tweets, QTs, replies) from the core ”Reform Psychology” data set described above. This is the data set where we limit ourselves to accounts that have at least two original tweets in both the ”Open Science” corpus and the ”Replication Crisis” corpus. Recall that this corpus contains 59.369 total English non-RT tweets from 984 unique accounts. The preprocessing follows a common natural language processing (NLP) cleaning pipeline. First, we

remove special characters including digits, emojis, hyperlinks, and mentions. In this first step, we also lowercase everything and ensure that all remaining words or characters are separated by a single whitespace. In the second step, we remove common English stopwords using the `nltk` library (Bird et al., 2009). The nltk list of stopwords is just a simple list with 179 common English words in lowercase. The list includes common words such as "i, me, you, he, she, so" as well as common contractions such as "don't, isn't, haven't" (full list in section A.2.2). These words and contractions will appear relatively uniformly in any English corpus, and as such, they are not helpful in terms of separating semantic topics in our corpus. We then tokenize the remaining words, using the encoding of the `SpaCy` library (Honnibal & Montani, 2017). Finally, we use the `SpaCy` library to lemmatize the tokens. Lemmatization groups together the inflected form of words, such that "dogs" becomes "dog" for instance. This is important since the topic model considers each unique token.

5.4.2 Topic Modeling

The two most popular approaches for topic modeling are Latent Dirichlet Allocation (LDA) and Non-negative Matrix Factorization (NMF) (Purpura, 2018). LDA is a probabilistic method, while NMF uses matrix decomposition. In this paper, we use NMF for topic modeling. This is motivated by empirical investigations, which have shown that NMF tends to produce better topics than LDA for short and noisy text, such as Twitter (Y. Chen et al., 2019; Habbat et al., 2020). We can conceptualize our documents (tweets) and lemmas (words) as a matrix $X = words \times documents$. The job of NMF is to factor this matrix into two matrices $W = words \times topics$ and $H = topics \times documents$ which approximate the original matrix X (Lee & Seung, 1999). This means that for each tweet (document) we get a distribution of topics that this tweet is associated with and for each topic we get a distribution of words that this topic is associated with (see figure 16).

NMF typically relies on a term-document matrix (X) in which terms are weighted by Term Frequency Inverse Document Frequency (TF-IDF) (Y. Chen et al., 2019). We perform this feature extraction with the `scikit-learn` package (Pedregosa et al., 2011). TF-IDF is an efficient and simple method for conducting feature extraction, which makes it a popular choice (Ramos et al., 2003). TF-IDF determines the relative frequency of a word in a document (here a tweet) compared to the inverse frequency of the same word across the entire corpus of documents (all the tweets) (Ramos et al., 2003). As such, TF-IDF measures the relevance, or importance, of a word to a document. Basically, a word that is uncommon across the corpus, but which appears in a specific document (tweet) is more informative with regards to the content of the document than the pres-

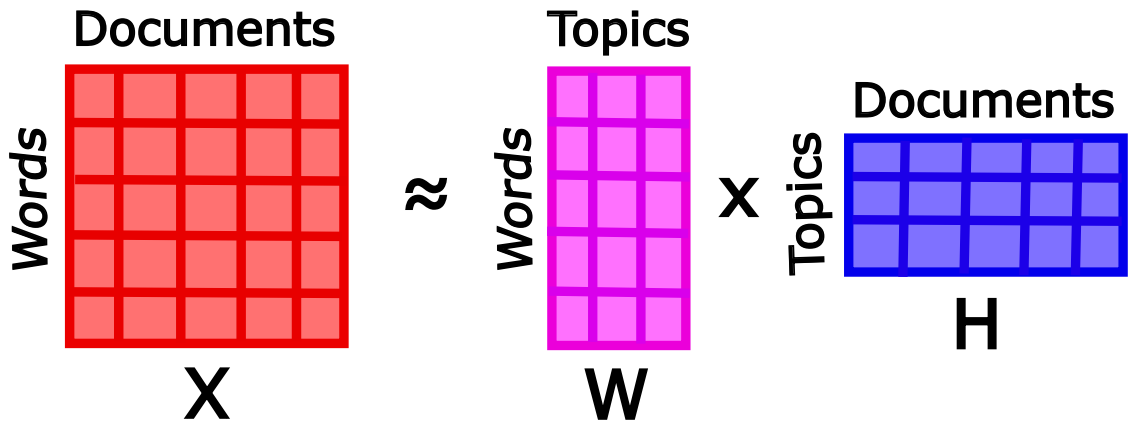

Figure 16: Schematic of the matrix decomposition of the matrix X into W and H that we accomplish using Nonnegative Matrix Factorization (NMF).

ence of a word that is common across the whole corpus. We limit ourselves to uni-grams, which means that we do not consider phrases of lemmas but only single lemmas. We only consider the top 2.500 most common lemmas in our corpus, and we remove lemmas that appear in more than 25% of documents. This is done for efficiency, but should also remove common and uninformative words which are not already filtered away by our list of stopwords.

For the NMF decomposition, we rely on the implementation by Kim et al. (2014) which is available on Github (Kim, n.d.). We use the default algorithm which is based on alternating nonnegative least squares (ANLS) with block principal pivoting. We vary the number of topics ($n = 10, 20, 50, 100$) to explore performance and topic coherency, using a grid-search approach similar to Allen et al. (2017). Although methods for finding "optimal" resolutions exist (Allen et al., 2017), manual inspection and validation are commonly used in applications similar to ours (Allen et al., 2017; Perry & DeDeo, 2021) and the approach follows best-practice guidelines (Roberts et al., 2016). Similar to the work by Perry and DeDeo (2021) we find that the best topic coherency is achieved with the maximum number of topics tested ($n = 100$). Our total data set $X = words \times documents$ has dimensions (2.500×59.369) since we limit ourselves to the 2.500 most important terms and our corpus consists of a total of 59.369 documents (English non-RT tweets). Using NMF, this is factored into $W = words \times topics$ which then has dimension (2.500×100), and $H = topics \times documents$ which then has dimensions (100×59.369).

5.4.3 Topic Selection

Although many of the resulting topics are coherent, a meaningful proportion of the topics suffer from at least one of two issues. The first problematic case is topics comprised of a bunch of common words that have not been excluded by the stopword removal or the limit that they must not appear in more than 25% of documents. We call these ”low coherence” topics. For instance, topic 87 has ”people, lot, thing” as the most strongly weighted words, and topic 68 has ”like, really, feel” as the most strongly weighted words. The second problematic case is topics that might be coherent but revolve around themes that are not meaningful for our purposes. We call these ”low interest” topics. For instance, topic 51 has ”talk, slide, podcast” as the most strongly weighted words, and topic 24 has ”join, come, discussion” as the most strongly weighted words. By removing topics with either low coherence or low interest (or both), we end up with 50 meaningful topics out of 100 total topics. The approach of manually removing meaningless or incoherent topics follows previous practice (Perry & DeDeo, 2021) and here we prioritize the relevance and coherence of topics rather than the amount of data. In table 4 (section A.2.3) we show the 10 most strongly weighted words for all 100 topics, along with the communities that we find for the 50 topics that we include in the semantic analysis.

5.4.4 Linkage Network

Inspired by Perry and DeDeo (2021), we construct a semantic linkage network between the 50 selected topics. We have approximated a matrix $H = topics \times documents$ in which all topics are distributions over documents. Some of these topic distributions will overlap more than others. We isolate this by computing the pairwise distances between topics in the H matrix using cosine distance as the metric and relying on the `scikit-learn` implementation (Pedregosa et al., 2011). By doing so, we obtain a matrix with dimensions (100×100) that contains the pairwise distances between our topics. Naturally then, smaller values are associated with proximity, and we build a weighted semantic linkage network based on this notion. Technically, we simply compute $proximity = 1 - distance$ since $distance \in (0, 1)$. Proximity then is the strength of association between topics, and we use this as the edge weight in the resulting network. Since all topics are related to all other topics with some strength (i.e. no topics are completely separated), it is useful to backbone this network (see 5.6). After backboning the semantic linkage network, we run a community detection algorithm on the network. We use the Louvain community detection algorithm (Blondel et al., 2008) which finds the best partition to be 5 communities, following the general

approach in Perry and DeDeo (2021). Before visualization, additional edges are removed to clarify the large-scale structure (top plot in figure 18). We track the volume of tweets which fall primarily in each of our topics over time. As each topic is associated with a community, we can aggregate this to track the number of tweets that are associated with each of our 5 communities over time (bottom plot in figure 18).

5.5 User Network

We construct a network between core accounts (those that have at least two original posts in both "Open Science" and "Replication Crisis") in which accounts are labeled based on the semantic community that they are most strongly associated with across their tweets. Recall that in this data set we have 984 unique accounts. The natural resolution of our topic modeling is that $tweets$ (H) and $words$ (X) are associated with $topics$. In order to arrive at an association between users and communities, we must do a bit of work. We assign each document (tweet) to the most related topic, and retain information about the contribution of this maximum topic for the tweet (e.g. $t_{max} = 0.18$). For all of the topics that we use in the semantic analysis ($n = 50$), we count the weight of this topic for each unique Twitter user in our sample. We then group the topics into communities and we assign the community with the highest score to the author (Twitter account). There are only 249 accounts displayed in this network, which is a result of several things. First, some accounts ($n = 135$) have no interactions (RT, QT, reply) with any other accounts in this sub-network. In addition, some accounts do not have any posts which appear in either of the 50 selected communities, the backboning removes some additional nodes, and we only plot the Giant Connected Component (GCC).

We investigate the centrality of users in the network of core accounts using weighted degree and eigenvector centrality. Weighted degree is the sum of in-degree (number of references to node x) and out-degree (number of references by node x). Weighted degree might be criticized as a naïve measure of centrality, or as a metric that simply fails to capture some of the key properties of what we mean by centrality (Ruhnau, 2000). This is because the measure does not consider how a node x is embedded in the larger network structure, but only considers the number (and strength of) connections with immediate neighbors. It is intuitive that if a node x has a strong connection with a central node y, this should make us more confident that node x is itself central than if x has a strong connection with a peripheral node z. It is this property that the weighted degree metric

fails to consider. We address this by also computing eigenvector centrality for all of the accounts ($n = 249$) in our backboned network of central actors. Eigenvector centrality (or "rank prestige") is a suitable metric for the comparison of "node-centrality" within a graph (Ruhnau, 2000) and it is arguably a better indication of centrality than weighted degree on social networks such as Twitter (Maharani, Gozali, et al., 2014). The algorithm recursively calculates the centrality of nodes based on the centrality of neighbors (Ruhnau, 2000), and as such captures the key intuition that how a node x is embedded in the larger network structure matters for centrality, in addition to the raw number of connections.

5.6 Backboning

Backboning is often applied to networks which are densely connected because the underlying structure of such networks can be hard to parse (Coscia & Neffke, 2017). In particular, we conduct network backboning to achieve two things. First, we backbone our networks for visual purposes. We want to declutter the networks to allow the overall structure of the network to emerge more clearly (Coscia & Neffke, 2017). Second, we conduct network backboning because some algorithms (e.g. community detection) can achieve better performance on a backboned network (Coscia & Neffke, 2017). A naïve approach is to simply remove edges with a weight that is lower than some arbitrary threshold (Coscia & Neffke, 2017). However, more sophisticated structural and (edge) significance-based approaches perform better in most cases (Coscia & Neffke, 2017). We rely on the python implementation of the Disparity Filter (DF) algorithm by Coscia and Neffke (2017) which is available online (Coscia, n.d.).

We use backboning at several points in our analysis, and we use different thresholds across cases. We mainly use backboning in order for our visualizations to be meaningful and as such we fine-tune the thresholds with that aim in mind. In the first case, we backbone the network which is visualized in figure 15 and consists of all "Open Science" and "Replication Crisis" connections. In this case, we use a cutoff that filters away most edges ($c = 0.995$), and additionally, we only display edges with weight > 150 to further declutter the visualization. Second, we backbone the semantic linkage network between topics which is displayed in figure 18. In this case, we backbone the network prior to running a community detection algorithm, and we use a softer filter ($c = 0.8$). Third, we backbone the user network before visualization (see figure 19). Here we use a threshold between the two previous cases ($c = 0.9$).

6 S2: Results

6.1 Prosocial Language

First, we address our main research question RQ_2 for this section (S2) of the thesis. Recall that RQ_2 asks whether the "Reform Psychology" movement is addressing issues of diversity and inclusion. We use the same list of prosocial terms that was used in Murphy et al. (2020) to assess the degree to which the "Open Science" and "Replication Crisis" groups on Twitter use prosocial language. For the "Replication Crisis" corpus we find that 8.115 out of 32.797 non-RT tweets (25%) match at least one prosocial term. For the "Open Science" corpus we find that 191.731 out of 511.056 non-RT tweets (38%) match at least one prosocial term. In the combined data set of 528.040 tweets, we find that 195.790 (37%) of tweets match at least one prosocial term. As such, the "Open Science" discourse contains more prosocial language than the "Replication Crisis" discourse. This conceptually replicates the findings of Murphy et al. (2020), where it was reported that the "Open Science" literature had particularly high rates of prosocial language.

We extend on the analysis by focusing on our core "Reform Psychology" corpus, which only contains tweets from accounts that have posted at least two original tweets in both "Open Science" and "Replication Crisis". Here we find that 19.745 out of 59.369 non-RT tweets (31%) contain at least one prosocial term. This places "Reform Psychology" between "Replication Crisis" and "Open Science" in terms of the amount of prosocial language. We break down the corpus by year and find that the relative frequency of tweets that contain prosocial language has increased substantially from 2009 to 2021. The proportion of tweets that contain prosocial language increases from around 10% in 2009 to around 40% in 2021 (see figure 17A). The total volume of tweets in 2009 is very low, and as such we might with more confidence remark on the increase in tweets that use prosocial terms from around 15% in 2014 to around 40% in 2021. Additionally, we find that in some semantic communities the fraction of tweets that match prosocial terms is much greater than in others (see figure 17B). "Culture & Training" is by far the most prosocial community with 3.112 out of 69.83 tweets matching prosocial terms (44.57%). For the "Publication" community 2.727 out of 7.535 tweets match prosocial terms (36.19%), which is similar to the "Data & Policy" community in which 1.717 out of 4.757 tweets match prosocial terms (36.09%). The "OSF" community contains less prosocial language, with 693 out of 2.457 tweets matching prosocial terms (28.21%), while the "Reform Psychology" community is the least prosocial with 1.941 out of 7.377

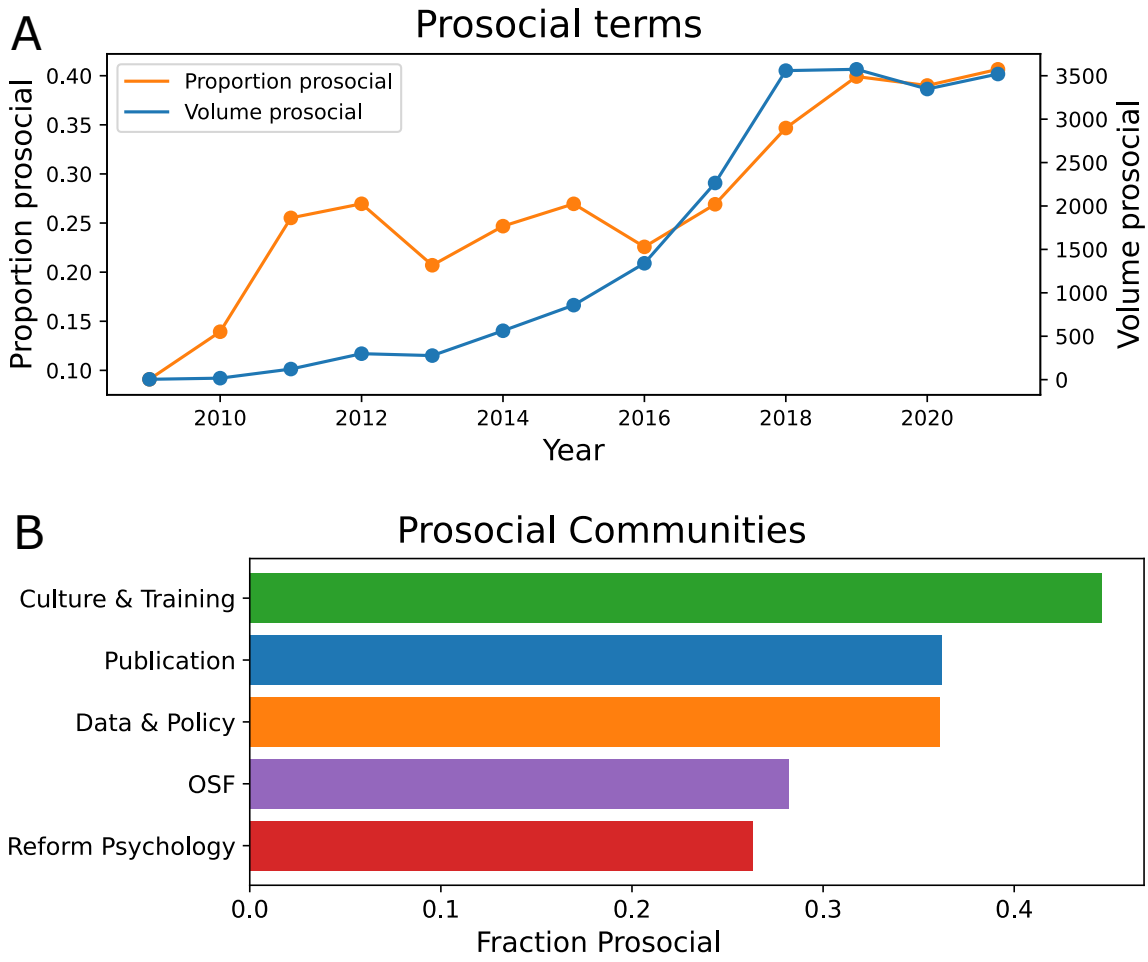

Figure 17: **A)**: Proportion and number of tweets which include prosocial terms from 2009 to 2021, including both full years. Shows prosocial evolution of ”Reform Psychology” discourse. **B)** Proportion of tweets which include prosocial terms grouped by community. Some communities (e.g. ”Culture & Training”) contain much more prosocial language than other communities (e.g. ”Reform Psychology”). Both figures based on the corpus which contains tweets only from users with at least two original tweets in both the ”Open Science” and ”Replication Crisis” corpora.

tweets matching prosocial terms (26.31%). In sum, we find that the ”Open Science” community is prosocial, that some ”Reform Psychology” communities use more prosocial language than others, and most importantly, there has been a strong positive trend toward prosocial language within the ”Reform Psychology” movement.

6.2 Topic Modeling & Linkage Network

Second, we address sub-question $RQ_{2.1}$ which aimed at understanding the dominant topics of discourse for the ”Reform Psychology” movement on Twitter. We address this by identifying 50 meaningful topics of discourse through topic modeling. Since topics overlap in semantic content it is natural to conceptualize this as a semantic linkage network (Perry & DeDeo, 2021). The semantic linkage network between topics is visualized in the top plot of figure 18. The linkage network shows how users link ideas (topics) and provide an intuitive visualization of the co-occurrences of various discourses. At the basic level, each node is a topic, and the connections (edges) between topics encode semantic overlap. For example, topic 35 and topic 75 (both in the ”Reform Psychology” community) are closely related. This makes sense since the top 10 most strongly weighted words in topic 35 are ”psychology, resolve, explain, forward, bad, theory, preregistration, specify, badly, personality”, and the top 10 most strongly weighted words in topic 75 are ”crisis, replication, fix, statistical, solution, theory, measurement, error, failure, bad”. Clearly, these two topics are strongly related semantically, although only a few terms directly overlap.

We use the Louvain algorithm (Blondel et al., 2008) to classify topics (nodes) into communities. Before applying the algorithm, we backbone the network (see 5.6). The algorithm finds that 5 communities represent the ”best partition” of the semantic linkage network. The communities reveal the large-scale structure of the semantic space and provide an intuitive overview of the central discourses within our corpus. The individual topics are colored by the communities they are assigned to in the top plot of figure 18.

The communities are generally well delineated in the sense that the topics (nodes) that are contained within each community do share common word-usage patterns and topic focus. We manually label the communities as focusing mainly on (1) Publication, (2) Culture & Training, (3) Data & Policy, (4) Reform Psychology and (5) Open Science Framework (OSF). I will briefly present examples of topics that fall within the 5 communities that we find, in order of how prosocial the communities were found to be (figure 17B).

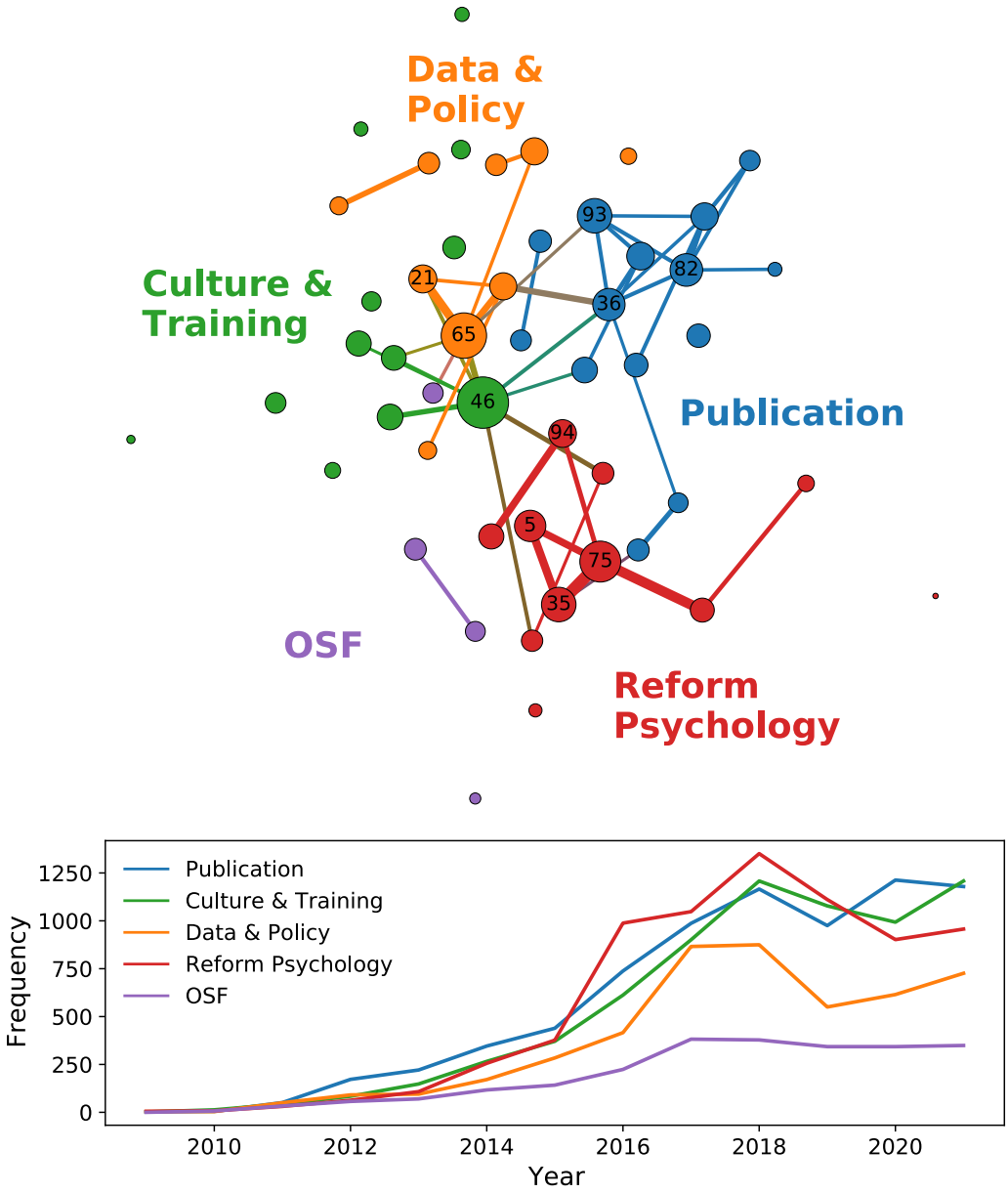

Figure 18: Results of the semantic modeling of the core "Reform Psychology" corpus from Twitter. Top plot shows the semantic linkage network between 50 topics of discourse. Bottom plot shows the volume of discourse in each community from 2009 to 2021 (incl. both full years).

In the most prosocial community (Culture & Training) we find topics with ”culture”, ”change” and ”incentives” as highly weighted terms (topic 20) as well as topics dealing with community building (topic 41) and ”training”, ”course” and ”handbook” (topic 37). As such, this topic appears to capture cultural changes both in terms of publishing, training, and community building. In the ”Publication” community we find topics that deal with peer review (topic 2), publication services such as the OSF (topic 42), discussion of publish-perish culture (topic 48), and several topics which discuss ”badges” (e.g. topics 82 and 60). As such, this community appears to contain both critique of publishing culture (e.g. ”publish-perish”) as well as discussion about attempts to improve the incentive structure of publishing (e.g. ”badges”). The ”Data & Policy” community contains topics which focus on ”opendata”, ”openresearch”, ”opensource” and ”openaccess” (topic 92), ”sharing” of ”code” (topic 21) as well as ”datascience”, ”bigdata” and ”ai” (topic 59). As such the topic captures technological improvements and a range of practices related to the open sharing of data and code. The ”OSF” community is small ($n = 4$ topics) and while it does primarily focus on this preregistration platform, several topics within the ”OSF” community also contain ”student” as a strongly weighted term (e.g. topic 98 and 45). The least prosocial community, ”Reform Psychology”, covers topics that are related to methodology, such as statistical understanding and low power (e.g. topics 15 and 75) as well as ”reproducibility” and ”replicability” (topic 34). Both of these discourses are most strongly associated with the early ”Mainstream Reform” response to the replication crisis as discussed previously. The community also contains topics which focus on solutions to issues of reproducibility in psychological science. For instance ”preregistration” is a highly weighted term in several topics (e.g topic 1 and 35). Discussion of ”theory” is also present in several topics (e.g. topics 35 and 75) showing that the recent ”Theory Reform” agenda is also represented here.

The semantic linkage network (figure 18A) allows us to inspect the underlying semantic landscape of the ”Reform Psychology” discourse. It appears that the ”Culture & Training” and ”Publication” communities form a center in the network. These communities are also associated with the most central actors in our network of core accounts (figure 19). Connected to this center we have the ”Data & Policy” community on the one hand, and the closely related ”OSF” and ”Reform Psychology” communities on the other hand.

We visualize the prevalence of topics in either of the five communities over time in the bottom plot in figure 18. Firstly, this provides an overview of our corpus, and it makes it obvious that we are not tracking each year since 2007 equally. The volume of tweets increases sharply from

2015 to 2016 and remains high throughout. The volume associated with the "Reform Psychology" community increases sharply from 2015 to 2016, which could indicate that this topic captures the reproducibility and methodology discussion which pervaded psychological science following the Open Science Collaboration (2015) article. This provides some empirical backing for the notion that psychological science only entered a phase of widespread lack of confidence following this central publication. The volume in this community drops after 2018 and we observe that both the "Culture & Training" and "Publication" communities gradually make up a larger fraction of the discourse. Since these two communities are also found to be the two most prosocial communities, this could explain at least some of the overall trend towards a more prosocial discourse. We cannot say that one agenda dominates the semantic landscape of the "Reform Psychology" movement, but it does appear as if discussions related to cultural change and the publication system are increasingly dominant, while the early focus on methodology is gradually receiving less attention.

6.3 Central Accounts

Third, we address our last sub-question ($RQ_{2.2}$) which was whether "Theory Reformers" are marginalized to the periphery of the network of "Reform Psychology". We attempt to locate the Twitter accounts (people, organizations) that are central for the "Reform Psychology" agenda, and to understand how they relate to each other internally. We use the data set of core accounts that have posted at least two original tweets in both "Open Science" and "Replication Crisis". This is the same subset of accounts that formed the basis for our topic modeling and semantic analysis in the previous section. We create a weighted network in which the strength of connections (edges) are determined by the number of interactions between accounts (RT, QT, reply). The network is backboned following the same methods as for the backboning of the semantic network described previously (5.6).

The full backboned network can be seen in the top of figure 19. The network is surprisingly dense, in the sense that almost all of the most well-connected (central) accounts are situated in a small core. We highlight this core in the bottom-left plot in figure 19. 8 of the top 10 most central accounts, as measured by weighted degree, are present in this core. We label these accounts in the zoomed plot.

In the bottom-right plot in figure 19, we plot the top handles based on weighted degree. Brian Nosek (@BrianNosek) is the most popular account with some margin, which is unsurprising given our treatment of him as one of the most important "Mainstream Reformers" in the introduction. He

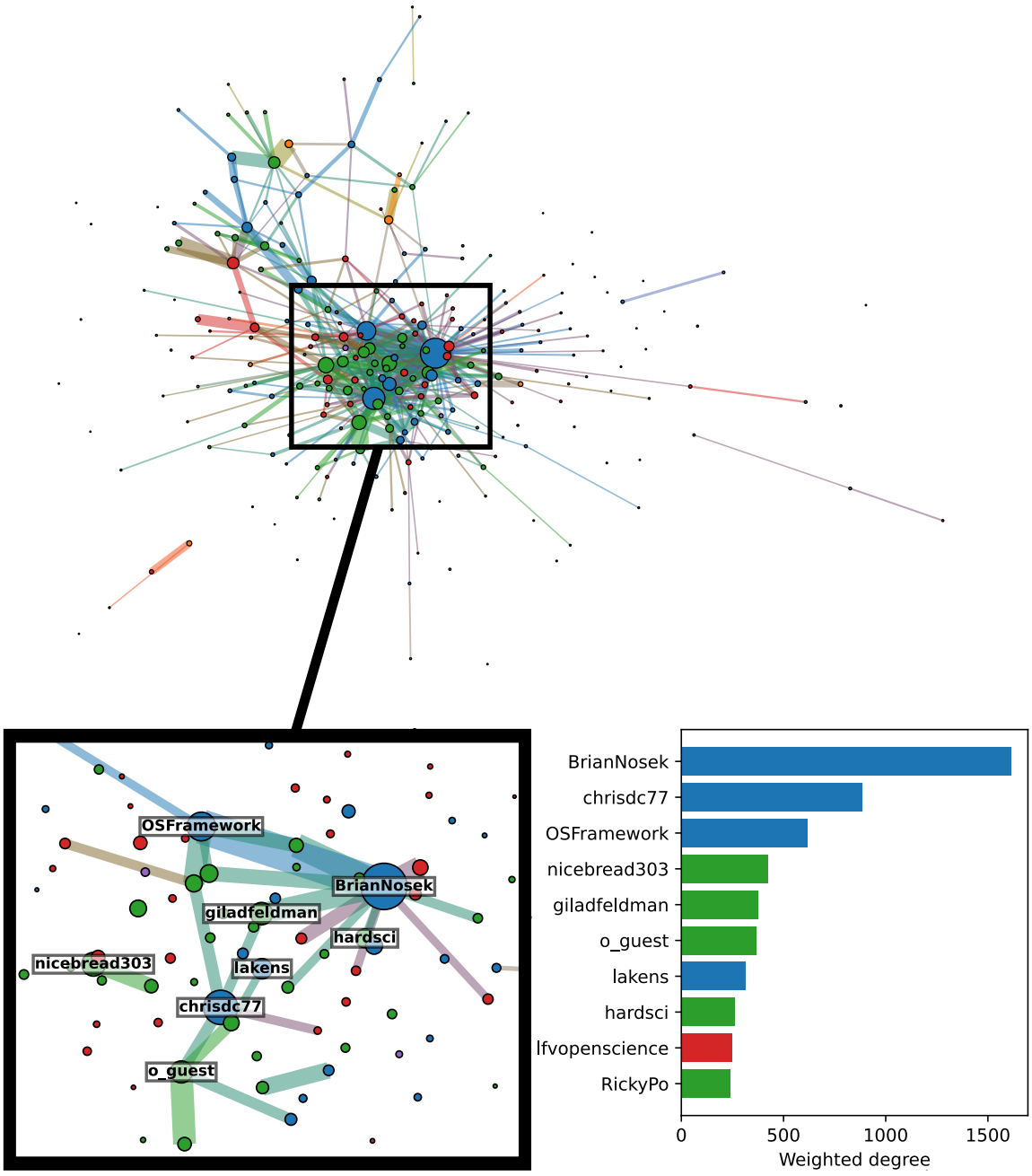

Figure 19: The weighted connections (edges) in the plot are determined by actual interactions (RTs, QTs, replies) between accounts and nodes are scaled in size based on their weighted degree. Following previous plots we have ”Publication” (blue), ”Culture & Training” (green), ”Data & Policy” (orange), ”Reform Psychology” (red), ”OSF” (purple).

is followed by Chris Chambers (@chrisdc77), the OSF (@OSFramework), and Felix Schönbrodt (@nicebread303). All of these accounts can also reasonably be classified as belonging to ”Mainstream Reformers” or organizations that support their agenda (see 7.3). We also observe that Olivia Guest (@o_guest), who we have already discussed as a ”Theory Reformer”, is present in the core of the network. The fact that we capture both central ”Mainstream Reformers” and known ”Theory Reform” critics suggests that our data selection has accurately captured the community of interest.

As described in section 5.5, weighted degree is a problematic measure of ”node centrality”, and because of this we also rank accounts based on eigenvector centrality. We plot the top 10 accounts ranked by eigenvector centrality in figure 30 (section A.2.4). Although this list partly overlaps with the top 10 most central accounts by weighted degree, we note that Olivia Guest (@o_guest) drops out of the top 10. In fact, we observe that all ”Theory Reformers” drop in relative position of centrality ranking when we rank accounts based on eigenvector centrality rather than weighted degree. This is the case for Olivia Guest (@o_guest) who drops from position 6 to position 14 (-8), for Eiko Fried (@EikoFried) who drops from position 25 to position 37 (-12), for Berna Devezer (@zerdeve) who drops from position 32 to 52 (-20) and for Iris Van Rooij who drops from position 22 to 147 (-125). Danielle Navarrio (@djnavarro) and Paul Smaldino (@psmaldino) are the only two ”Theory Reformers” highlighted by Flis (2022) that are not present in the set of 249 key actors which are displayed in figure 19. Overall, our investigation shows that ”Theory Reformers” are generally popular (in terms of activity) but they are less well connected to other central accounts in the network. This supports the thesis that the ”Theory Reformers” (both male and female) are marginalized to some extent.

7 S2: Discussion

In this second part of the thesis (S2) we set out to answer one main question (RQ_2) which was whether the "Reform Psychology" movement is addressing issues of diversity and inclusion. We posed two additional sub-questions ($RQ_{2.1}$ and $RQ_{2.2}$). $RQ_{2.1}$ was aimed at understanding the key topics of discussion within the movement, and how the reform agenda has changed over time. $RQ_{2.2}$ posed the question of whether the "Theory Reformers" identified in Flis (2022) are marginalized in the broader "Reform Psychology" community. We will evaluate these in turn on the basis of the findings presented in the results section.

7.1 Prosocial Language (RQ_2)

We tested our main research question (RQ_2) for this part (S2) by classifying tweets as either prosocial or not using a previously applied (Murphy et al., 2020) list of prosocial terms (see A.2.1). We found that for the "Open Science" corpus, 38% of tweets were classified as prosocial, whereas for the "Replication Crisis" corpus this was just 25%. More importantly, in our core "Reform Psychology" corpus, we tracked the evolution of prosocial language from 2009 to 2021 and found a strongly increasing trend in the fraction of tweets containing prosocial language (10% in 2009 to 40% in 2021). This suggests that the "Reform Psychology" movement is addressing issues of diversity and inclusion, and thus provides an affirmative answer to RQ_2. We further observed that the "Culture & Training" community was the most prosocial with 44.57% of tweets matching prosocial terms, whereas the "Reform Psychology" community was found to be the least prosocial community, with 26.31% of tweets matching prosocial terms. We also observe that the "Culture & Training" community contained most volume in 2021, whereas the "Reform Psychology" community dominated from 2016 to 2018. As such, the overall trend towards more prosocial language might be related to a shift in focus from identifying issues of methodology and questionable research practices (QRPs) and to a growing focus on questionable scientific culture (QSC).

I recognize two challenges to the evaluation that the "Reform Psychology" movement is increasingly addressing topics of diversity and inclusion. First, the list of prosocial terms that we use might not adequately capture what we mean by prosocial, and might fail to capture the promotion of "diversity" and "inclusion". In particular, some words on our list of prosocial terms might reasonably be suspected to be tied not exclusively to prosocial discourse. For instance, "concern" is something

that is often expressed for the *field* of psychological science, and this expression of concern does not necessarily bear a prosocial connotation. In addition, although a word such as ”team” might suggest a prosocial concern, it should also figure prominently in debates about big-team science (Coles et al., 2022) which might focus on the higher statistical power associated with large investigations, rather than on the prosocial aspect of big-team science.

A second challenge is claiming that the reform movement is addressing issues of inclusion of diversity by showing that the reform movement is increasingly using ”prosocial language”, which is a move that we ideally want to make. There are two possible mechanisms that could render this move invalid. First, it is possible that some actors in the ”Reform Psychology” movement are not actually prosocially minded, but that they nonetheless want to appear prosocial, and thus use prosocial language. Some tweets might even argue that a prosocial focus is misguided for the ”Reform Movement” and still be labeled as ”prosocial”. This is possible since we do not model the context in which we match prosocial terms. Second, it is possible that although there is a genuine prosocial concern, this does not translate into behavioral change or improved outcomes.

We can only address the second challenge with reference to other lines of evidence, and we save this for the overall discussion (section 8). However, I believe that we can defend the challenge regarding our operationalization of ”prosocial” language here. First, we should recognize that it will always be the case that we either capture too much (i.e. classify some tweets as prosocial which are not) or that we capture too little (i.e. fail to classify some tweets as prosocial that in fact are prosocial). We should not disregard this objection, but I believe that several factors should strengthen our belief that we actually do reasonably capture prosocial discourse. First, the results are consistent with observations by Murphy et al. (2020), who used the same list of prosocial terms to show that the ”Open Science” discourse was especially prosocial. We conceptually replicate this result in a new domain here, which should increase our confidence that the list of prosocial terms does consistently capture something. That this something is indeed prosocial language is supported by the fact that the ”Culture & Training” community is found to be the most prosocial community with some margin. Clearly, this semantic community deals with the most explicitly prosocial themes in our corpus, and as such we have consistency across methods. Lastly, the results year-over-year are consistent and large, such that only around 10% of tweets match prosocial terms in 2009 while the fraction reaches roughly 40% in 2021. In sum, we can reasonably claim that the ”Reform Psychology” agenda is increasingly dominated by a prosocial focus.

7.2 Topics & Semantics

Sub-question $RQ_{2.1}$ aimed at understanding the key topics of discourse in the ”Reform Psychology” movement and how these have evolved over time. In the semantic linkage network between topics (top plot in figure 18) we observed that topics within the communities ”Publication” and ”Culture & Training” form a center. We supported this notion by showing that in tweets from the most central actors in the movement, these topics dominate almost completely. In particular, the three most central actors, Brian Nosek (@BrianNosek), Chris Chambers (@chrisdc77), and the Open Science Framework (@OSFramework) are all primarily tied to the ”Publication” community. Many other central actors, such as Felix Schönbrodt (@nicebread303), Gilad Feldman (@giladfeldman), and Olivia Guest (@o_guest), are primarily tied to the ”Culture & Training” community.

To understand what this implies it is helpful to consider the volume of tweets that are primarily tied to either of our five communities over time (bottom plot in figure 18). We observed a spike in ”Reform Psychology” volume in 2016, following the Open Science Collaboration (2015) publication. This community is generally associated with what we might call early ”Mainstream Reform” discourse, although not exclusively. Topics focusing on low power, methodology, and statistics dominate many of the individual topics which are grouped together in this community. Interestingly, the volume of tweets tied to this community reaches a maximum in 2018 while the volume of tweets in the ”Publication” and ”Culture & Training” communities has continued to grow. This might suggest a shift of focus in the reform movement such that diagnosing issues (e.g. QRPs) has become a less important discourse, potentially because some consensus has been reached. The data suggest that two debates have eclipsed this previously dominant discourse. First, while the first wave of response to the replication crisis focused on research culture, it appears as if concerns with regard to the problematic culture in academia more broadly construed have become central issues of debate. This is supported by the fact that the ”Culture & Training” community becomes the largest in terms of volume for the first time in 2021. This community is by far the community that has the strongest prosocial focus, and the topics in this community are also those which most directly seem to address this broader concern with the culture in psychological science. This is evident by the fact that highly weighted terms in topics that are grouped into this community include ”change”, ”culture”, ”climate” and ”behavior” (topic 20). Second, there has been a sustained focus on ”Publication”, and we observe that the most central accounts (Nosek, Chambers, OSF) in our system are primarily tied to this community. This suggests that attempts to change the publication

system, and as such to change the incentive structure in psychological science, has become the main agenda for the group of ”Mainstream Reformers”.

The only technical caveat here is that we are weighting the number of individual tweets that are tied to each community, and not the amount of traction (RTs, QTs, replies) that the tweets have received. If there are systematic differences in the popularity of tweets from the semantic communities that we find, then this could complicate the presented narrative. We do not have reason to suspect this, but it is technically a limitation on what we can conclude.

7.3 Central Accounts

Sub-question $RQ_{2.2}$ aimed at understanding whether the central ”Theory Reformers” identified in Flis (2022) are marginalized in the ”Reform Psychology” movement. We located the most central accounts based on weighted degree and displayed the top 10 accounts in Figure 19 (bottom right). We already highlighted some of the central accounts in the results section. Unsurprisingly, the list is dominated by what we might call ”Mainstream Reformers”. In order of centrality ranking by weighted degree they are Brian Nosek (@BrianNosek), Chris Chambers (@chrisdc77), OSF (@OSFramework), Felix Schönbrodt (@nicebread303), Gilad Feldman (@giladfeldman), Daniël Lakensn (@lakens), Sanjay Srivastava (@hardsci). These accounts all belong to psychologists (except for the organizational OSF account) and some of the most important accounts are directly involved with the ”Open Science” movement (Brian Nosek, Chris Chambers, Felix Scönbrodt). The others are related through work on mass-replications (Gilad Feldman) or are actively involved in the discussion about methods, statistics, and QRPs (Daniël Lakens and Sanjay Srivastava).

At the heart of the network is Brian Nosek, who chairs the Center for Open Science (COS). The two subsequent accounts in terms of ranking (Chris Chambers and OSF) are also directly related to COS activity. Chris Chambers currently chairs the Registered Reports (RR) committee which is supported by the COS (“OSF Registered Reports”, 2022), and as we mentioned previously RRs is one of the central reform proposals from the ”Mainstream Reformers” (Chambers, 2019). OSF is an important preregistration platform (“OSF”, 2022) which was created by (and is maintained by) the COS. Semantically, all of these accounts are tied primarily to the ”Publication” community, and this strongly suggests that the core of the ”Mainstream Reform” movement is now primarily tied to the COS which in turn promotes institutional reform within psychological science.

The representation of the ”Theory Reformers” as central voices is less clear. Given the notion of

this group as outsiders (Flis, 2022) it is surprising to find Olivia Guest in the core of the network and as one of the most central accounts. She continues to be a strong critic of both the culture of the ”Open Science” community (#bropenscience) and of the preregistration format which is so central to the ”Mainstream Reformers”. As we noted, all the ”Theory Reformers” that are identified by Flis (2022), and which we observe in our network of central actors, drop in relative centrality position when we re-weight the network based on eigenvector centrality. This suggests that although the ”Theory Reformers” have strong platforms, they are less well connected to most central accounts in the network (e.g. Brian Nosek). In fact, we observe that Olivia Guest is one of the central accounts that is most distantly connected to Brian Nosek (figure 19 bottom left). As such, while it does not appear to be true that the ”Theory Reformers” (e.g. Olivia Guest, Berna Devezer and Iris Van Rooij) are ignored, they are clearly not part of the inner circle of ”Mainstream Reformers”.

A key limitation of the analysis is that we do not attempt to disentangle the nature of the communication between accounts, for instance with sentiment analysis. In most cases, it is reasonable to assume that a strong connection signals a continued positive interaction. As an example, the interaction between Brian Nosek and the OSFramework is almost certainly positive, since Brian Nosek is the president of the Center for Open Science (COS) which maintains the OSF. Although I believe that it is true that most strong associations between accounts are positive, it is possible that interactions between some accounts are antagonistic. In particular, we are limited in the conclusions we can draw with regards to the nature of the discourse between ”Mainstream Reformers” and ”Theory Reformers”. It is possible that either (i) the ”Mainstream Reform” community is embracing the perspective and critique from the ”Theory Reformers”, or that (ii) the ”Mainstream Reform” community is criticizing or marginalizing the viewpoints of the ”Theory Reformers”. All we can say is that our analysis shows that most of the ”Theory Reformers” that Flis (2022) highlights are present in both the core network of ”Reform Psychology” and that several topics of discussion (topic 5, 35 and 75) suggest that discussions of ”theory” are important in the reform space.

8 Discussion Overall

We have addressed research questions related to replication studies and incentives in psychological science (S1) and related to the prosocial culture, topics of discussion, and social dynamics of the ”Reform Psychology” movement (S2). In the discussion sections for each part of the thesis, we have evaluated the strength of our findings. In this overall discussion, we evaluate the findings in light of other research, and we suggest a coherent picture of the state and direction of the ”Reform Psychology” movement.

8.1 Incentives of Reform

It has been argued in various places that incentive structures are not well aligned to support the production of reproducible and replicable research. Incentives are important because the wrong incentives can lead to harmful consequences on research quality even when they are unconscious or implicit (Munafò et al., 2022), and misaligned incentives can naturally select for bad science without any conscious strategizing by scientists (Smaldino & McElreath, 2016). Even positive accounts, which argue that incentives are changing, acknowledge that the right incentive structure ”isn’t in place yet” (Allen & Mehler, 2019). A common notion is that novelty is rewarded more than rigor (Uhlmann et al., 2019) with some indications suggesting that the emphasis on ”novelty” and ”innovation” has increased substantially over time (Smaldino & McElreath, 2016). There might also be unique challenges associated with the production of particular types of scientific documents. For instance, replication studies have been argued to be particularly disincentivized, because replications have historically been perceived as ”attacks” on the authors of the original effects (Nosek et al., 2022). Smaldino and McElreath (2016) modeled the ”prestige” conferred onto a scientist from publishing a replication study as only half of the prestige conferred onto a scientist who publishes a novel finding, describing this as ”highly favorable” to replications.

In our case, we have focused specifically on citation rate as our proxy for ”prestige”. Citations are useful because (i) they are accessible, (ii) they are a basic measure of the diffusion of ideas through the scientific record, and (iii) there is prestige associated with high citation rates and related measures such as the h index (Gruber et al., 2021; Smaldino, 2020). We believe that higher citation rates to certain kinds of research production should provide a positive incentive to engage with these lines of research. Science is a complex system where policies, norms, incentives, and funding all

affect the kind of research that gets published (Nosek et al., 2022; Smaldino & McElreath, 2016). As such, citation rates are only a limited proxy, but citation rates and h indices are important metrics of impact (Gruber et al., 2021; Smaldino & McElreath, 2016) and as such, they matter for hiring and scientific advancement. Citation rates also matter because they are related to the impact factor of journals, and if citation rates are low for reproducible research then journals might hesitate to publish reproducible research (Hummer et al., 2017). Some journals have refused to publish failed replications of novel findings published in their venues (Smaldino & McElreath, 2016).

In this thesis, we have shown that replication studies in psychological science are cited at significantly higher rates than matched records of non-replication studies. This result holds across two heterogeneous sets of data (R_{FOS}, R_{QUERY}) and the effect in both cases is meaningful ($3.31 - 4.58$ additional expected citations for replication studies as compared to matched controls). Additionally, the model conditioned on the R_{FOS} data set suggests that this trend might be increasing over time, supporting other indications (Nosek et al., 2022) that psychological scientists are increasingly valuing reproducible work. Analysis of citation rates in other domains that can reasonably be associated with "Reform Psychology" similarly suggests that citation rates are high for reproducible and open science research. We have previously mentioned the initial finding that Registered Reports (RRs) are cited at rates comparable to, or even higher, than conventional articles (Hummer et al., 2017). Another line of work has focused on Open Access (OA) papers. This is slightly further removed from the present scope of the thesis, but OA is another important goal of the "Open Science" and "Reform Psychology" movements. There are mixed results from this line of research, but in most cases, OA papers have been associated with a citation advantage, and they appear to receive more media coverage than papers that are not OA (McKiernan et al., 2016).

Although the complementary lines of research are generally consistent with our findings, the results presented in this thesis considerably strengthen the overall evidence of a general citation advantage for "Reform Psychology" research. The evidence for a citation advantage of RRs as compared to conventional articles is limited since the RR format is new, and the Hummer et al. (2017) study consequently had a small effective sample size ($n = 97$) of RR articles. In comparison, we show a consistent effect across two much larger samples of data, with $n = 620$ replication studies matched in the R_{FOS} data set and $n = 1196$ replication studies matched in the R_{QUERY} data set. In addition to improving on the sample size and the statistical modeling sophistication of previous investigations, we also extend on previous findings by exploring a new domain. I am not aware of previous investigations which have compared the rate of citations for replication studies to non-

replication studies. This broadens the scope of "Reform Psychology" formats which have now been associated with high citation rates from OA and RR, to also include replication studies.

The role of replications has been questioned by some (Szollosi & Donkin, 2021; Irvine, 2021) and it has been shown that psychologists do not update their beliefs about effects in the literature optimally following negative replications (McDiarmid et al., 2021). In particular, it has been found that findings can persist in the literature even when subsequent replication studies with much higher power have failed to replicate effects (Hardwicke et al., 2021; Tatsioni et al., 2007). Even if replication studies are useful, they might not be the panacea that some reformers have hoped (Smaldino & McElreath, 2016). Still, the "Mainstream Reformers" continue to emphasize the self-correcting role that replication studies can play (Nosek et al., 2022), and efforts outside of psychology (e.g. Errington et al., 2021) are increasingly adopting the large-scale replication formats that we have seen in psychological science (Open Science Collaboration, 2015). Without taking an explicit stance on the effectiveness of replication studies, and the role that they should play, I believe that this is enough to defend the importance of understanding citation rates to articles of this format. Simply put, citation rates are related to both author and journal incentives, and they are also a metric of the influence and impact that articles have on the scientific record. If citation rates are high, we can generally assume that this is a positive incentive and that the replication efforts do have an effect on subsequent knowledge production, which would justify resource investment. Overall, the emerging picture is that research that promotes robustness, replicability, and openness is broadly embraced by the psychological science community and that this trend might be accelerating. We believe that the combined findings relating to the high citation rates of "Reform Psychology" formats challenge the prevailing notion that reproducible and open research is disincentivized.

8.2 Cultural Change

The previous section suggests that incentive structures are changing, and that reform practice is increasingly popular. An important further question that we address in this thesis is whether the reform movement will also address issues with the broader scientific culture, and work for a stronger focus on prosocial behavior, inclusive practice, and promotion of diversity. As we have argued previously, diversity is desirable from both a political and an epistemic point of view (Devezer et al., 2019; Hofstra et al., 2020; Nielsen et al., 2017; Sulik et al., 2021). The question is how we can achieve a more diverse psychological science, with a stronger representation of traditionally

marginalized groups. In terms of obtaining tenure track positions and securing grants, women appear to be underrepresented because of lower submission rates rather than lower acceptance rates, and women are more likely to report that they do not feel that they belong in their organization (Gruber et al., 2021). This might suggest that the primary issue is not with institutional bias (e.g. hiring committees) and that the strongest barrier could be the interpersonal culture, including whether the scientific community promotes prosocial norms.

We have already noted that there are conflicting lines of evidence with regards to whether the ”Mainstream Reform” community is prosocial. A group of (especially female) ”Theory Reformers” have criticized the ”Mainstream Reformers” under the #bropenscience hashtag, highlighting issues of exclusion and narrow demographics (Whitaker & Guest, 2020). This group of ”Theory Reform” outsiders have also remarked that criticizing ”Mainstream Reform” agendas (e.g. preregistration) comes at the cost of unproductive and insulting responses (D. Navarro, 2020). These experiences do not suggest a prosocial culture within the reform movement, and might lead to female drop-out to the extent that ”bullying” and ”incivility” are cultural barriers for female involvement (Gruber et al., 2021). This echoes some of the early criticism from the ”Old Guard”, which largely focused on the perceived ”incivility” from the ”Mainstream Reformers”, and the ”tone” of the debate (Derksen & Field, 2021; S. T. Fiske, 2016). On the other hand, empirical evidence (Murphy et al., 2020) has suggested the the ”Open Science” community uses prosocial language and is communal, and suggested a causal relation to an increasing representation of women in this literature.

In this thesis, we show that the ”Reform Psychology” community on Twitter is increasingly using prosocial language. Using the same list of prosocial terms as Murphy et al. (2020), we find that the fraction of tweets that match at least one of these terms has increased from around 10% in 2009 to around 40% in 2021 (figure 17). Given the size of the effect, and the complementary evidence that Murphy et al. (2020) present, we believe that this does reflect a genuinely improving prosocial tendency within the movement. We show that some semantic communities are more prosocial than others, and in particular that the ”Culture & Training” and ”Publication” communities use much more prosocial language than for instance the ”Reform Psychology” community. We also show that the ”Culture & Training” and ”Publication” communities have continued to grow in volume and that discussion around these topics is now more prevalent than discussion related to the ”Reform Psychology” community. It seems natural that the ”Reform Psychology” community employs relatively little explicitly prosocial language. This community expands quickly in 2016, following the Open Science Collaboration (2015) publication, and I believe that this community

roughly captures the ”first wave” of reform, in which methodology was the dominant topic. Topics such as methodology, statistics, and research practice are technical and lack a natural connection to a prosocial agenda. The subsequent and sustained growth in the ”Culture & Training” and the ”Publication” communities might reflect a ”second wave” of reform, which focus on institutional change at various levels and calls for a broader reform of scientific culture. The mechanism which has driven this change is unclear, but science is not a silo, and external developments such as the #MeToo movement, as well as criticism from ”Theory Reformers”, might both play a role. At least the critique of a broken culture that the #MeToo movement has highlighted across many arenas seem echoed to some degree in the #bropenscience debate which Olivia Guest has started in the ”Reform Psychology” movement. As such, the increased focus on diversity is likely due to both inside critics (e.g. ”Theory Reformers”) and a broader cultural trend. Whether this focus will be reflected in stronger female (and minority) representation is still unclear. If there is a causal connection between prosocial language and minority representation as suggested in Murphy et al. (2020), then our results should reinforce the positive outlook that they suggest.

It is reasonable to assume that a shift in agendas can account for the increase in the fraction of prosocial language that we observe. We have argued that language in itself is important and that culture and discourse could be more important than explicit institutional change. As such, the results are positive and important in their own right. However, we might additionally be interested in whether the growing prosocial focus has resulted in concrete action and better outcomes. The ”Mainstream Reform” community has recognized issues of representation (Coles et al., 2022) and made the intention to address issues of representation explicit (Moshontz et al., 2018; Nosek, 2017). However, concrete actions and interventions appear to be rare, and we still observe that traditionally marginalized groups, such as females and non-western labs, continue to be underrepresented in the reform movement (Coles et al., 2022; Moshontz et al., 2018; Nosek, 2017). If we broaden the scope and look at whether the ”Theory Reformers” have been successful in challenging the ”preregistration” agenda, there are tentative but positive developments. A recent review of the state of the ”Reform Psychology” movement explicitly emphasized the importance of better understanding ”measurement” and ”theory” moving forward (Nosek et al., 2022). This suggests that the ”Theory Reformers” have achieved some success in shaping the outlook of the reform movement, although intentions will have to be followed up by action (e.g. funding and promotion). It is possible that outcomes and institutional change will catch up with the bottom-up focus on a more prosocial culture that we observe on Twitter and in other domains (Murphy et al., 2020). The early ”Mainstream

Reform" movement has achieved significant change with regards to methods and statistical practice (Nosek et al., 2022) and reforming curricula (Smaldino, 2020), and it might take some time before more recent agendas achieve similar wide-spread recognition and adoption. Overall, we suggest that the "Reform Psychology" movement consists of a core of "Mainstream Reformers" (e.g. Brian Nosek, Chris Chambers) who are the primary forces in pushing agendas of methodology and reproducibility, but that the outsider group of "Theory Reformers" (e.g. Danielle Navarro, Olivia Guest) have achieved some success in influencing the direction of the movement.

8.3 Outlook and future directions

Recent developments suggest that incentives are increasingly favoring the production of "Mainstream Reform" research. Rates of citations are a limited measure, but we observe positive developments in several domains. To name a few, we observe growth in funding and fellowships awarded for "Open Science" research (McKiernan et al., 2016), we observe that universities increasingly emphasize adherence to "Open Science" practices and principles in hiring decisions (Nosek et al., 2022), we observe strong growth in the number of journals that now accept submissions in the "Registered Report" format (Chambers, 2019) and who award "badges" which distinguish open and reproducible research (McKiernan et al., 2016; Science, n.d.). There are reasons to believe that adoption is following. We showed that research categorized as both "Open Science" and "Replication" has grown at higher rates than the overall "Psychology" literature. This finding is congruent with other observations, for instance, the non-linear growth in the user-base on the OSF, which recently crossed 400.000 users (Rice & Alexis, 2021)

Looking forward, we might expect the trend to accelerate further. In particular, the observation that early career researchers are disproportionally invested in reform practices suggests a positive outlook. It has been found that PhD students and postdocs first-authored 78% ($n = 82$) of the Registered Reports (RRs) in *Cortex* compared with 67% ($n = 57$) in a control sample (Chambers, 2019). In general, there is a strong emphasis on education and training in the reform community, aiming at equipping psychologists with the technical skills to adopt reform practice (Jekel et al., 2020), and more fundamentally, aiming at reforming curricula such that necessary statistical and theoretical skills are increasingly taught (Smaldino, 2020). Through our semantic modeling of the "Reform Movement" discourse, we support the notion that this is an increasingly central agenda within the reform movement. We observe that words and phrases such as "student", "young",

”early career”, ”generation”, ”training” and ”course” (see table 4) feature prominently in many topics of discourse. This suggests that institutionalization of reform practice is affecting psychology programs, which could lead to a more methodologically sophisticated, and culturally reformed, generation of psychologists.

If the trends highlighted above continue in this direction, then incentives will increasingly favor the production of research streams that are proposed by the ”Mainstream Reform” community. Some have voiced concern that the reform movement might run out of steam (Inzlicht, 2020) but the evidence presented here supports the opposite conclusion. Of course, there is a genuine discussion to be had about whether the ”Mainstream Reform” that appears to be succeeding, does sufficiently address the issues of psychological science. Some have argued that the reform movement is perpetuating reliance on the Null Hypothesis Significance Testing (NHST) paradigm with the focus on replication efforts and that this is fundamentally unproductive. The methodological focus on higher power, reproducible research, and the replication of NHST effects might be seen as treating symptoms rather than the root cause of issues (Szollosi & Donkin, 2021). While this is an important discussion, I believe that it is too early to assess whether more radical reform thinking can gain wide-spread traction. In particular, developments on Twitter take time to disseminate into the published record, and this means that some developments are still to early to gauge. I believe that a parsimonious reading of the data at hand is that the methods-dominated reform agenda is accelerating, and that reform is likely to also affect the broader culture of psychological science. The reform direction is currently dictated by the ”Mainstream Reform” community, but we observe that the group is not insular, and that ”Theory Reformers” have achieved some success in setting the agenda in this actively evolving community of reformers.

9 Conclusion

By investigating the ”Reform Psychology” movement across Twitter as well as the published literature, this thesis shows that the ”Reform Psychology” movement is transforming the culture and institutions of psychology. We show that replication studies are cited at higher rates than non-replication studies. This trend might increase over time, which would support other indications that the psychological science community is increasingly favoring reproducible research. Together with research on the citation rates of Registered Reports (Hummer et al., 2017) and Open Access papers (McKiernan et al., 2016), this suggests that reproducible and open science is generally cited at high rates. Assuming that high citation rates are a positive incentive, the findings suggest that in this area incentives favor reform practice. Developments with regards to funding and hiring, as well as the adoption of reform practice by early career researchers, suggests that reform practice might continue to accelerate. We show that the ”Reform Psychology” agenda is increasingly dominated by topics related to publication and culture, while discussion of methodology is less dominant now than it was during the initial response to the replication crisis. This signals a shift in the reform agenda, such that addressing institutional issues (e.g. publication incentives) and broader cultural issues (e.g. inclusiveness) have emerged as the most important agendas for reformers. We find that the ”Reform Psychology” community is increasingly prosocial, with a growth in the fraction of tweets that match prosocial terms from around 10% in 2009 to around 40% in 2021. We also observe that discussion of ”theory” is represented in several semantic topics, and mainstream reformers have recognized ”theory” and ”measurement” as key focus areas moving forward (Nosek et al., 2022). This jointly suggests that the group of ”Theory Reformers” have achieved some success in shaping the direction of the reform agenda. We show that critics of the culture and the proposals of the ”Mainstream Reform” movement are not ignored, but our findings also suggest that while the voices of critics are being heard, they are not well connected to the most central ”Mainstream Reformers” which are driving institutional change. In sum, our findings suggest that the reform movement is gradually becoming the new mainstream in psychology, and importantly, that this transition is likely to address not only methodological problems, but also deeper theoretical issues and broader cultural problems.

Bibliography

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *PLoS Medicine*, *2*(8), e124. https://doi.org/10.1371/journal.pmed.0020124

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

Flis, I. (2022). The Function of Literature in Psychological Science. *Review of General Psychology*, 108926802110664. https://doi.org/10.1177/10892680211066466

Gruber, J., Mendle, J., Lindquist, K. A., Schmader, T., Clark, L. A., Bliss-Moreau, E., Akinola, M., Atlas, L., Barch, D. M., Barrett, L. F., Borelli, J. L., Brannon, T. N., Bunge, S. A., Campos, B., Cantlon, J., Carter, R., Carter-Sowell, A. R., Chen, S., Craske, M. G., … Williams, L. A. (2021). The Future of Women in Psychological Science. *Perspectives on Psychological Science*, *16*(3), 483–516. https://doi.org/10.1177/1745691620952789

Devezer, B., Nardin, L. G., Baumgaertner, B., & Buzbas, E. O. (2019). Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one*, *14*(5), e0216125.

Scotchmer, S. (1991). Standing on the shoulders of giants: Cumulative research and the patent law. *Journal of economic perspectives*, *5*(1), 29–41.

Zeigler, D. (2012). Evolution and the cumulative nature of science. *Evolution: Education and Outreach*, *5*(4), 585–588.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M. B., et al. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual review of psychology*, *73*, 719–748.

Munafò, M. R., Chambers, C., Collins, A., Fortunato, L., & Macleod, M. (2022). The reproducibility debate is an opportunity, not a crisis. *BMC Research Notes*, *15*(1), 43. https://doi.org/10.1186/s13104-022-05942-3

Murphy, M. C., Mejia, A. F., Mejia, J., Yan, X., Cheryan, S., Dasgupta, N., Destin, M., Fryberg, S. A., Garcia, J. A., Haines, E. L., Harackiewicz, J. M., Ledgerwood, A.,

Moss-Racusin, C. A., Park, L. E., Perry, S. P., Ratliff, K. A., Rattan, A., Sanchez, D. T., Savani, K., … Pestilli, F. (2020). Open science, communal culture, and women's participation in the movement to improve science. *Proceedings of the National Academy of Sciences*, *117*(39), 24154–24164. https://doi.org/10.1073/pnas.1921320117

Rodgers, J. L., & Shrout, P. E. (2018). Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights from the Behavioral and Brain Sciences*, *5*(1), 134–141.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Allen, C., LUOc, H., Murdock, J., PUa, J., Wang, X., Zhai, Y., & Zhao, K. (2017). Topic modeling the hàn diǎn ancient classics (□□□□). *Journal ISSN*, *2371*, 4549.

Allen, C., & Mehler, D. M. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS biology*, *17*(5), e3000246.

Hummer, L., Thorn, F. S., Nosek, B. A., & Errington, T. (2017). Evaluating registered reports: A naturalistic comparative study of article impact.

Chambers, C. (2019). What's next for registered reports?

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Heering, P. (2017). Science museums and science education. *Isis*, *108*(2), 399–406.

Jamieson, V. (2018). Why are there so few women in physics? *New Scientist*, *240*(3203), 32–37.

Hofstra, B., Kulkarni, V. V., Galvez, S. M.-N., He, B., Jurafsky, D., & McFarland, D. A. (2020). The diversity–innovation paradox in science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291.

Nielsen, M. W., Alegria, S., Börjeson, L., Etzkowitz, H., Falk-Krzesinski, H. J., Joshi, A., Leahey, E., Smith-Doerr, L., Woolley, A. W., & Schiebinger, L. (2017). Opinion: Gender diversity leads to better science. *Proceedings of the National Academy of Sciences*, *114*(8), 1740–1742.

Whitaker, K., & Guest, O. (2020). # bropenscience is broken science: Kirstie whitaker and olivia guest ask how open 'open science'really is. *The Psychologist*, *33*, 34–37.

Coles, N. A., Hamlin, J. K., Sullivan, L. L., Parker, T. H., & Altschul, D. (2022). Build up big-team science. *Nature*, *601*(7894), 505–507. https://doi.org/10.1038/d41586-022-00150-2

Bandiera_abtest: a Cg_type: Comment Subject_term: Research management, Institutions

Derksen, M., & Field, S. M. (2021). *The tone debate: Knowledge, self, and social order* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/g8nhu

Gelman, A. (2016). *What has happened down here is the winds have changed*. Retrieved May 23, 2022, from https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Voodoo correlations in social neuroscience. *Perspectives on psychological Science*, *4*(3), 274–290.

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.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Engber, D. (2017). Daryl bem proved esp is real: Which means science is broken. *Slate, May*, *17*.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & Van Der Maas, H. L. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on bem (2011).

Szollosi, A., & Donkin, C. (2021). Arrested Theory Development: The Misguided Distinction Between Exploratory and Confirmatory Research. *Perspectives on Psychological Science*, *16*(4), 717–724. https://doi.org/10.1177/1745691620966796

Stevens, J. R. (2017). Replicability and Reproducibility in Comparative Psychology. *Frontiers in Psychology*, *8*, 862. https://doi.org/10.3389/fpsyg.2017.00862

Malich, L., & Munafò, M. R. (2022). Introduction: Replication of Crises: Interdisciplinary Reflections on the Phenomenon of the Replication Crisis in Psychology. *Review of General Psychology*, 108926802210779. https://doi.org/10.1177/10892680221077997

Hesse, B. W. (2018). Can psychology walk the walk of open science? *American Psychologist*, *73*(2), 126.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., McDougall, D., Nosek, B. A., Ram, K., Soderberg, C. K., et al. (2016). Point of view: How open science helps researchers succeed. *elife*, *5*, e16800.

Williamson, E. W. (2022, May 23). *After 10 Years, 'Many Labs' Comes to an End – But Its Success Is Replicable*. Retrieved May 25, 2022, from https://news.virginia.edu/content/after-10-years-many-labs-comes-end-its-success-replicable

OSF Registered Reports. (2022). Retrieved May 23, 2022, from https://osf.io/rr/

Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., Grahe, J. E., McCarthy, R. J., Musser, E. D., Antfolk, J., Castille, C. M., Evans, T. R., Fiedler, S., Flake, J. K., Forero, D. A., Janssen, S. M. J., Keene, J. R., Protzko, J., Aczel, B., … Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing Psychology Through a Distributed Collaborative Network. *Advances in Methods and Practices in Psychological Science*, *1*(4), 501–515. https://doi.org/10.1177/2515245918797607

OSF. (2022). Retrieved May 23, 2022, from https://osf.io/

Fiske, S. T., Schacter, D. L., & Taylor, S. E. (2016). Introduction. *Annual Review of Psychology*, *67*(1), annurev-ps-67-121415–100001. https://doi.org/10.1146/annurev-ps-67-121415-100001

Barrett, L. F. (2015). Psychology is not in crisis. *The New York Times*, *1*.

Derksen, M., & Morawski, J. (2022). Kinds of Replication: Examining the Meanings of "Conceptual Replication" and "Direct Replication". *Perspectives on Psychological Science*, 174569162110411. https://doi.org/10.1177/17456916211041116

Fiske, S. (2016). *Fiske presidential guest column_APS Observer_copy-edited.pdf*. Retrieved May 23, 2022, from https://www.dropbox.com/s/9zubbn9fyi1xjcu/Fiske%

5C%20presidential%5C%20guest%5C%20column_APS%5C%20Observer_copy-edited.pdf

Fiske, S. T. (2016). A Call to Change Science's Culture of Shaming. *APS Observer*, *29*. Retrieved May 23, 2022, from https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Guest, O., & Martin, A. E. (2021). How Computational Modeling Can Force Theory Building in Psychological Science. *Perspectives on Psychological Science*, *16*(4), 789–802. https://doi.org/10.1177/1745691620970585

Fried, E. I. (2020). Theories and Models: What They Are, What They Are for, and What They Are About. *Psychological Inquiry*, *31*(4), 336–344. https://doi.org/10.1080/1047840X.2020.1854011

van Rooij, I., & Baggio, G. (2020). Theory Development Requires an Epistemological Sea Change. *Psychological Inquiry*, *31*(4), 321–325. https://doi.org/10.1080/1047840X.2020.1853477

van Rooij, I. (2022). Psychological models and their distractors. *Nature Reviews Psychology*, 1–2. https://doi.org/10.1038/s44159-022-00031-5

Navarro, D. J. (2021). If Mathematical Psychology Did Not Exist We Might Need to Invent It: A Comment on Theory Building in Psychology. *Perspectives on Psychological Science*, *16*(4), 707–716. https://doi.org/10.1177/1745691620974769

Smaldino, P. E. (2020). How to Build a Strong Theoretical Foundation. *Psychological Inquiry*, *31*(4), 297–301. https://doi.org/10.1080/1047840X.2020.1853463

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2021). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, *16*(4), 744–755.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahník, Š., Bernstein, M. J., Bocian, K., Brandt, M. J., Brooks, B., Brumbaugh, C. C., Cemalcilar, Z., Chandler, J., Cheong, W., Davis, W. E., Devos, T., Eisner, M., Frankowska, N., Furrow, D., Galliani, E. M., … Nosek, B. A. (2014). Investigating Variation in Replicability:

A “Many Labs” Replication Project. *Social Psychology*, *45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, Š., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., Busching, R., … Nosek, B. A. (2018). Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, *1*(4), 443–490. https://doi.org/10.1177/2515245918810225

Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Chartier, C. R., Christopherson, C. D., Clay, S., Collisson, B., Crawford, J., Cromar, R., Vidamuerte, D., Gardiner, G., Gosnell, C., Grahe, J., Hall, C., Joy-Gaba, J., Legg, A. M., Levitan, C., … Ratliff, K. (2019). Many Labs 4: Failure to Replicate Mortality Salience Effect With and Without Original Author Involvement. https://doi.org/10.31234/osf.io/vef2c

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., Baranski, E., Bernstein, M. J., Bonfiglio, D. B., Boucher, L., Brown, E. R., Budiman, N. I., Cairo, A. H., Capaldi, C. A., Chartier, C. R., Chung, J. M., Cicero, D. C., Coleman, J. A., Conway, J. G., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, *67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corley, M., Hartshorne, J. K., IJzerman, H., et al. (2020). Many labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, *3*(3), 309–331.

Soderberg, C. K., Errington, T., Schiavone, S. R., Bottesini, J. G., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2020). Initial evidence of research quality of registered reports compared to the traditional publishing model.

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying long-term scientific impact. *Science*, *342*(6154), 127–132.

Perry, C., & DeDeo, S. (2021). The cognitive science of extremist ideologies online. *arXiv preprint arXiv:2110.00626*.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Dong, Y., & Kanakia, A. (2020). Microsoft academic graph: When experts are not enough. *Quantitative Science Studies*, *1*(1), 396–413.

Paszcza, B. (2016). *Comparison of microsoft academic (graph) with web of science, scopus and google scholar* (Doctoral dissertation). University of Southampton.

Chen, C. (2020). A glimpse of the first eight months of the covid-19 literature on microsoft academic graph: Themes, citation contexts, and uncertainties. *Frontiers in research metrics and analytics*, *5*, 24.

Effendy, S., & Yap, R. H. (2017). Analysing trends in computer science research: A preliminary study using the microsoft academic graph. *Proceedings of the 26th international conference on world wide web companion*, 1245–1250.

Wang, K., Shen, Z., Huang, C., Wu, C.-H., Eide, D., Dong, Y., Qian, J., Kanakia, A., Chen, A., & Rogahn, R. (2019). A review of microsoft academic services for science of science studies. *Frontiers in Big Data*, *2*, 45.

Sinatra, R., Wang, D., Deville, P., Song, C., & Barabási, A.-L. (2016). Quantifying the evolution of individual scientific impact. *Science*, *354*(6312), aaf5239.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., & Modrák, M. (2020). Bayesian Workflow. *arXiv:2011.01808 [stat]*.

McElreath, R. (2020). *Statistical rethinking: A bayesian course with examples in r and stan*. Chapman; Hall/CRC.

Bürkner, P.-C. (2017). Brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, *80*, 1–28.

Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., Petersen, A. M., Radicchi, F., Sinatra, R., Uzzi, B., et al. (2018). Science of science. *Science*, *359*(6379), eaao0185.

Wu, L., Wang, D., & Evans, J. A. (2019). Large teams develop and small teams disrupt science and technology. *Nature*, *566*(7744), 378–382.

Kurz, A. S. (2021). *Statistical rethinking with brms, ggplot2, and the tidyverse: Second edition* (version 0.2.0). https://bookdown.org/content/4857/

Fusaroli, R., Weed, E., Fein, D., & Naigles, L. (2021). Caregiver linguistic alignment to autistic and typically developing children.

Cox, R. F.

bibinitperiod C. (2022). *Workshop on Bayesian Inference: Priors and workflow*. Retrieved May 25, 2022, from https://4ccoxau.github.io/PriorsWorkshop/index.html

Bürkner, P.-C. (n.d.). *Non-Linear Hypothesis Testing — hypothesis.brmsfit*. Retrieved May 23, 2022, from https://paul-buerkner.github.io/brms/reference/hypothesis.html

Szalay, A., & Gray, J. (2006). Science in an exponential world. *Nature*, *440*(7083), 413–414.

Piwowar, H., Priem, J., Larivière, V., Alperin, J. P., Matthias, L., Norlander, B., Farley, A., West, J., & Haustein, S. (2018). The state of oa: A large-scale analysis of the prevalence and impact of open access articles. *PeerJ*, *6*, e4375.

Hardwicke, T. E., Szűcs, D., Thibault, R. T., Crüwell, S., van den Akker, O. R., Nuijten, M. B., & Ioannidis, J. P. A. (2021). Citation Patterns Following a Strongly Contradictory Replication Result: Four Case Studies From Psychology. *Advances in Methods and Practices in Psychological Science*, *4*(3), 251524592110408. https://doi.org/10.1177/25152459211040837

Uzzi, B., Mukherjee, S., Stringer, M., & Jones, B. (2013). Atypical combinations and scientific impact. *Science*, *342*(6157), 468–472.

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances*, *7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., & Gabry, J. (2015). Pareto smoothed importance sampling. *arXiv preprint arXiv:1507.02646*.

Perc, M. (2014). The matthew effect in empirical data. *Journal of The Royal Society Interface*, *11*(98), 20140378.

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., Bornstein, B. H., Bouwmeester, S., Brandimonte, M. A., Brown, C., Buswell, K., Carlson, C.,

Carlson, M., Chu, S., Cislak, A., Colarusso, M., Colloff, M. F., Dellapaolera, K. S., Delvenne, J.-F., … Zwaan, R. A. (2014). Registered Replication Report: Schooler and Engstler-Schooler (1990). *Perspectives on Psychological Science*, *9*(5), 556–578. https://doi.org/10.1177/1745691614545653

Twarc. (n.d.). Retrieved May 23, 2022, from https://twarc-project.readthedocs.io/en/latest/

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with python: Analyzing text with the natural language toolkit*. " O'Reilly Media, Inc."

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [To appear].

Purpura, A. (2018). Non-negative matrix factorization for topic modeling. *DESIRES*, 102.

Chen, Y., Zhang, H., Liu, R., Ye, Z., & Lin, J. (2019). Experimental explorations on short text topic mining between lda and nmf based schemes. *Knowledge-Based Systems*, *163*, 1–13.

Habbat, N., Anoun, H., & Hassouni, L. (2020). Topic modeling and sentiment analysis with lda and nmf on moroccan tweets. *The Proceedings of the Third International Conference on Smart City Applications*, 147–161.

Lee, D. D., & Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Ramos, J. et al. (2003). Using tf-idf to determine word relevance in document queries. *Proceedings of the first instructional conference on machine learning*, *242*(1), 29–48.

Kim, J., He, Y., & Park, H. (2014). Algorithms for nonnegative matrix and tensor factorizations: A unified view based on block coordinate descent framework. *Journal of Global Optimization*, *58*(2), 285–319.

Kim, J. (n.d.). *Kimjingu/nonnegfac-python: Python toolbox for nonnegative matrix factorization*. Retrieved May 23, 2022, from https://github.com/kimjingu/nonnegfac-python

Roberts, M., Stewart, B., & Tingley, D. (2016). Navigating the local modes of big data: The case of topic models. comput. *Soc. Sci*, *4*.

Blondel, V. D., Guillaume, J.-L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, *2008*(10), P10008.

Ruhnau, B. (2000). Eigenvector-centrality—a node-centrality? *Social networks*, *22*(4), 357–365.

Maharani, W., Gozali, A. A. et al. (2014). Degree centrality and eigenvector centrality in twitter. *2014 8th international conference on telecommunication systems services and applications (TSSA)*, 1–5.

Coscia, M., & Neffke, F. (2017). Network backboning with noisy data. *International Conference on Data Engineering (ICDE)*.

Coscia, M. (n.d.). *Michele Coscia | Network Backboning*. Retrieved May 23, 2022, from https://www.michelecoscia.com/?page_id=287

Uhlmann, E. L., Ebersole, C. R., Chartier, C. R., Errington, T. M., Kidwell, M. C., Lai, C. K., McCarthy, R. J., Riegelman, A., Silberzahn, R., & Nosek, B. A. (2019). Scientific Utopia III: Crowdsourcing Science. *Perspectives on Psychological Science*, *14*(5), 711–733. https://doi.org/10.1177/1745691619850561

Irvine, E. (2021). The Role of Replication Studies in Theory Building. *Perspectives on Psychological Science*, *16*(4), 844–853. https://doi.org/10.1177/1745691620970558

McDiarmid, A. D., Tullett, A. M., Whitt, C. M., Vazire, S., Smaldino, P. E., & Stephens, J. E. (2021). Psychologists update their beliefs about effect sizes after replication studies. *Nature Human Behaviour*, *5*(12), 1663–1673. https://doi.org/10.1038/s41562-021-01220-7

Tatsioni, A., Bonitsis, N. G., & Ioannidis, J. P. A. (2007). Persistence of Contradicted Claims in the Literature. *JAMA*, *298*(21), 2517. https://doi.org/10.1001/jama.298.21.2517

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.

Sulik, J., Bahrami, B., & Deroy, O. (2021). The Diversity Gap: When Diversity Matters for Knowledge. *Perspectives on Psychological Science*, 174569162110060. https://doi.org/10.1177/17456916211006070

Navarro, D. (2020). *Paths in strange spaces: A comment on preregistration* (Preprint). PsyArXiv. https://doi.org/10.31234/osf.io/wxn58

Nosek, B. A. (2017). *How can we improve diversity and inclusion in the open science movement?* Retrieved May 23, 2022, from https://www.cos.io/blog/how-can-we-improve-diversity-and-inclusion-open-science-movement

Science, C. f. O. (n.d.). *Open Science Badges*. Retrieved May 23, 2022, from https://www.cos.io/initiatives/badges

Rice, B. A., & Alexis, N. P. (2021). *OSF reached a pinnacle of 400,000 registered users*. Retrieved May 23, 2022, from https://www.cos.io/blog/osf-reaches-400000-users

Jekel, M., Fiedler, S., Allstadt Torras, R., Mischkowski, D., Dorrough, A. R., & Glöckner, A. (2020). How to teach open science principles in the undergraduate curriculum—the hagen cumulative science project. *Psychology Learning & Teaching*, *19*(1), 91–106.

Inzlicht, M. (2020). *The Replication Crisis Is Not Over*. Retrieved May 30, 2022, from http://michaelinzlicht.com/getting-better/2020/6/26/the-replication-crisis-is-not-over

Priem, J., Piwowar, H., & Orr, R. (2022). Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*.

DeDeo, S. (2022). Using big data to track major shifts in human cognition. *Proceedings of the National Academy of Sciences*, *119*(4).

Cumming, G. (2014). The new statistics: Why and how. *Psychological science*, *25*(1), 7–29.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does “failure to replicate” really mean? *American Psychologist*, *70*(6), 487.

A Supplementary Information

A.1 S1: Supplementary Information

A.1.1 MAG data base

The Microsoft Academic Graph (MAG) has been discontinued as of the 2021-12-31 and will be superseded by Open Alex (Priem et al., 2022). We use a fork of the MAG from 2021-08-12 at which point the MAG was still being maintained. The MAG is a large-scale and free publication database which has been widely used in previous research (Murphy et al., 2020; C. Chen, 2020; Effendy & Yap, 2017), especially in the science of science domain (K. Wang et al., 2019). The MAG uses web-crawl to achieve high coverage (K. Wang et al., 2020). MAG coverage is comparable to Google Scholar and it has much higher coverage than other commonly used data bases such as Scopus and Web of Science (Paszcza, 2016)

An issue with the MAG database is that we do not know how the MAG classifies documents, besides that it uses AI (K. Wang et al., 2020). Critically in our case, we do not know how it classifies documents into document types, such as ”Journal” and ”Conference”, and we do not know how it classifies documents into fields of study (FoS) such as ”Psychology” and ”Replication”. This leaves potential issues of ”unknown unknowns”, which is a problem that is not limited to the MAG but is common across many data bases (DeDeo, 2022). We attempt to address this issue by not relying on a single data set, but we curate multiple sets of data, which should increase the robustness of our findings.

The ”Replication” category is assigned at multiple levels, but we only include studies that are assigned at the appropriate level (level = 2). The ”Open Science” and ”Reproducibility” categories which were used in a related study (Murphy et al., 2020) are only assigned at this level in the hierarchy, and we take this to indicate that this is the appropriate level.

In the MAG, each paper is assigned to a document type. The largest document types are ”None”, ”Journal” and ”Patent”. As mentioned in the methods section, the MAG uses web-crawl to achieve high coverage, but this also means that it includes more than we would traditionally label as being actual scientific papers. In our main data subset (”Candidate Papers”) we only allow ”Conference” and ”Journal” document types which follows previous work (Murphy et al., 2020).

A.1.2 Variables

As detailed in the METHODS section, we construct two main sets of data which we model (R_{FOS} and R_{QUERY}). We extract the number of authors ($TEAMSIZE$), the year of publication ($YEAR$) and we calculate the number of citations five years after initial publication (c_5). We should specify, that although our main data sets (R_{FOS}, R_{QUERY}) only consist of papers that are labeled with the "Psychology" field of study and either "Conference" or "Journal" as document type, we do not restrict the citing papers to this selection. As such papers with "Medicine" as their field of study, or papers with "Patent" as document type can count towards the c_5 of the papers in our sample.

The outcome we are modeling is the number of citations a scientific publication has received in the first five years following publication (c_5). Clearly, a longer citation-delay will result in a more accurate estimate of the long-term impact of scientific articles. In particular, it has been found that paradigm-changing articles have limited early impact (D. Wang et al., 2013), and our operationalization will underestimate the impact of these articles. However, as D. Wang et al. (2013) note, the early citation rates for most articles is a good indication of long-term scientific impact. Since this paper is concerned with recent developments there is a trade-off between the length of citation delay we chose and how recent papers we can include in our sample. Since the last complete year of data in our sample is 2020, we would only be able to investigate papers published up until 2010 (including 2010) if we worked with c_{10}. In view of the discussion in the introduction, 2010 is probably too early to capture any replication crisis response, including the recent emphasis on large-scale replication efforts (Coles et al., 2022). By using c_5 we are able to include papers up until 2015 (including 2015) in our sample which I believe represents a reasonable trade-off.

A.1.3 Matching

Both of our main data sets (R_{FOS} and R_{QUERY}) consists not only of replication (experiment) studies, but also of an equal number of matched non-replication (control) studies. The matched control studies are sampled from the "Candidate Papers" data set, and as such they are also "Psychology" papers published as either "Conference" or "Journal" between 2005 and 2015. We match exactly on number of authors ($TEAMSIZE$) and approximately on year of publication ($YEAR$). For R_{FOS} we match 618 of the 620 total replication (experiment) records exactly on $YEAR$ (99.68%). For R_{QUERY} we match 1194 of the 1196 total replication records exactly on $YEAR$ (99.83%).

A.1.4 Amount of Data

As mentioned earlier we are able to match $n = 620$ replication papers for the R_{FOS} data set ($n = 1240$ including controls) and we are able to match $n = 1196$ replication papers for the R_{QUERY} data set ($n = 2392$ including controls). We match 228 papers for the "Reproducibility" ($R*_{FOS}$ data set ($n = 456$ including controls) which is also a decent sample size. However, we are only able to match 27 papers for the "Open Science" (OS_{FOS}) data set ($n = 54$ including controls). For this reason, we decide not to model or further investigate the OS_{FOS} data set. Note that this also means that the results reported in (Murphy et al., 2020) regarding the prosocial discourse in the "Open Science" literature does not intersect to a large degree with psychological science.

A.1.5 Data Quality

We conduct manual checks of our data sets to ensure that they actually match the categories specififed by the MAG. For the sample which we collect by matching on the appearance of "replicat*" in titles, we also must check data quality, since papers can contain "replicat*" and not actually be replication studies. We will want to have some idea about the fraction of studies that are (i) not related to replication at all, and (ii) the fraction that are conceptual or meta-science articles rather than actual replications.

Since the amount of data is generally to large to manually check, we select a random subset ($n = 20$) of records from each data set (R_{FOS}, R_{QUERY} and $R*_{FOS}$) excluding the Open Science (OS_{FOS}) data set which we already reported to be too limited to warrant modeling. We manually check whether the selected records actually fall within the desired categories. Generally, we observe that R_{FOS} and R_{QUERY} do mainly contain actual replication studies, which is what we want. For both, we see many titles include phrases such as "a replication study", "replication project", "a replication and extension", etc. The random sample of papers from $R*_{FOS}$ seems to be heavily skewed towards cognitive neuroscience. We have a disproportionate amount of articles in this sample which contain references to brain areas (e.g. "hippocampus", "amygdala") as well as neuroscience methodology (e.g. "mri", "transcranial magnetic stimulation"). We can speculate that this might be because the MAG uses NLP to extract document categories (as well as e.g. publication venue) and that "reproducibility" is a more prominent term within this line of inquiry. However, what we really want to match is studies from psychology broadly that have e.g. reproducible data and code. We do not appear to have accurately matched the desired sample in this case. For this reason, we do

not report on the model based on this data sample. We report the random subset ($n = 20$) of data for all four data samples below with titles and year of publication (see tables, 1, 2, 3 below).

Year	PaperTitle
2006	experimental analysis and treatment of multiply controlled problem behavior a systematic replication and extension
2007	identifying subtypes of criminal psychopaths a replication and extension
2008	a computational model of spatial visualization capacity
2008	prediction of clinical outcomes from rtms in depressed patients with lateral visual field stimulation a replication
2008	assessing l2 reading texts at the intermediate level an approximate replication of crossley louwerse mccarthy mcnamara 2007
2008	the tripartite influence model of body image and eating disturbance a replication with a japanese sample
2011	measuring change during behavioral parent training using the parent instruction giving game with youngsters piggy a clinical replication
2011	the intellectual disability version of the very short form of the physical self inventory psi vs id cross validation and measurement invariance across gender weight age and intellectual disability level
2011	comprehensive application of behavior analysis to schooling in italy the pilot project
2012	immigrant groups vocational training and employment
2012	application of research experiment teaching in the course of basic clinical laboratory medicine
2012	an inter rater reliability study of a self assessment for the multiple intelligences
2014	change starts with journal editors in response to makel 2014

2014	e satisfaction a moroccan replication and extension
2014	learning from failed replications cognitive load manipulations and charitable giving
2014	schizotypal personality questionnaire brief revised psychometric replication and extension
2015	is there a connection between electrosensitivity and electrosensibility a replication study
2015	applicant reactions are similar across countries a refined replication with assessment center data from the european union
2015	a replication of a factor analysis of motivations for trapping
2015	the dispositional basis of attitudes a replication and extension of hepler and albarracin 2013

Table 1: R_{FOS}: ($n = 20$) sampled papers

Year	PaperTitle
2005	the strange stories test a replication study of children and adolescents with asperger syndrome
2005	cognitive abilities chunk strength and frequency effects in implicit artificial grammar and incidental l2 learning replications of reber walkenfeld and hernstadt 1991 and knowlton and squire 1996 and their relevance for sla
2007	cognitive therapy for bulimia nervosa an a b replication series
2007	social approach behaviors are similar on conventional versus reverse lighting cycles and in replications across cohorts in btbr t tf j c57bl 6j and vasopressin receptor 1b mutant mice
2007	identifying subtypes of criminal psychopaths a replication and extension

2009	sra winner a replication of choice control change c3 obesity prevention curriculum for middle school students using a lead teacher model
2010	does the dsm iv clinical significance criterion for major depression reduce false positives evidence from the national comorbidity survey replication
2011	are social workers ignoring the cornerstone of science by failing to replicate their research
2011	confirmatory analyses of the school refusal assessment scale revised replication and extension to a truancy sample
2011	population level right handedness for a coordinated bimanual task in naturalistic housed chimpanzees replication and extension in 114 animals from zambia and spain
2011	mammoth cloning reminds us of jurassic park but storm replication does not naturalistic settings do not aid the retrieval of distant analogs
2012	comorbidity of partial and subthreshold ptsd among men and women with eating disorders in the national comorbidity survey replication study
2012	professional nursing burnout and irrational thinking a replication study
2013	is there an association between temperament and apolipoprotein e a replication of a 1993 young finns study
2013	issues involving the use of significant sameness in testing replications and generating knowledge
2014	investigating variation in replicability a many labs replication project
2015	the dual pathway model of bulimia replication and extension with anorexia
2015	reduction in ventral striatal activity when anticipating a reward in depression and schizophrenia a replicated cross diagnostic finding
2015	replication of the training program in nonverbal communication in gerontology

2015	replication and external validation of a bi factor parameterization of attention deficit hyperactivity symptomatology

Table 2: R_{QUERY} ($n = 20$) sampled papers

Year	PaperTitle
2005	accuracy of stereotaxic positioning of transcranial magnetic stimulation
2005	reproducibility of the spectral components of the electroencephalogram during driver fatigue
2006	reproducibility of activation in four motor paradigms an fmri study
2007	screening for refractive errors in preschool children with the vision screener
2008	development and validation of the japanese version of the constitution in chinese medicine questionnaire ccmq
2009	reproducibility of placebo analgesia effect of dispositional optimism
2009	reproducibility of bold signal change induced by breath holding
2009	electromyographic standardized indices in healthy brazilian young adults and data reproducibility
2010	reproducibility of brainscan studies questioned
2010	functional magnetic resonance imaging fmri reproducibility and variance components across visits and scanning sites with a finger tapping task
2011	multi parametric neuroimaging reproducibility a 3 t resource study
2011	the reproducibility and convergent validity of the walking index for spinal cord injury wisci in chronic spinal cord injury

2012	is there a correlation between hippocampus and amygdala volume and olfactory function in healthy subjects
2013	reproducibility of functional network metrics and network structure a comparison of task related bold resting asl with bold contrast and resting cerebral blood flow
2013	eyes open eyes closed dataset sharing for reproducibility evaluation of resting state fmri data analysis methods
2013	reproducibility of somatosensory spatial perceptual maps
2014	retest imaging of 11c nop 1a binding to nociceptin orphanin fq peptide nop receptors in the brain of healthy humans
2014	regional reproducibility of calibrated bold functional mri implications for the study of cognition and plasticity
2015	accuracy and reliability of the pfeffer questionnaire for the brazilian elderly population
2015	validity and reproducibility of measures of oropharyngeal dysphagia in preschool children with cerebral palsy

Table 3: $R{*}_{FOS}$ ($n = 20$) sampled papers

A.1.6 Top Cited Articles

In addition to checking a random sample of $n = 20$ papers for our four samples of data, we also check the top-cited articles ($c_5 > 100$) to see whether the top cited articles are skewed in any way. We only do so for R_{FOS}, $R{*}_{FOS}$ and R_{QUERY} since the amount of data for OS_{FOS} is too small to warrant modeling in any case.

For R_{FOS} we observe that many of the top cited articles are review or meta-science pieces. One example, is Cumming (2014) which is the article with the highest number of citations in this data set ($c_5 = 988$). Other examples include Makel et al. (2012) which is has $c_5 = 206$ and Maxwell et al. (2015) which has $c_5 = 192$. In this sample of data there are also many actual replications that

are highly cited. For instance, the second most cited article ($c_5 = 279$) in the sample is Klein et al. (2014).

For R_{QUERY} we observe some of the same trends. We have a mixture of review articles, discussing the role of replication in psychology along with actual replication studies. In addition however, we also have several instances from the "national comorbidity survey replication" (https://www.hcp.med.harvard.edu/ncs/). There are several of these because they are conducted yearly and for various effects. The fact that these studies have a large presence in the top part of the c_5 distribution prompts the question of whether these studies can legitimately be considered psychology replication studies.

For $R*_{FOS}$ we do not observe any apparent issues in the top c_5 articles. The top cited articles are still skewed towards cognitive neuroscience, but we already noted this in the sample of $n = 20$ random articles from this data set.

A.1.7 Conclusion on data curation

In sum, we do not have enough data in one sample (OS_{FOS}) to warrant a modelling effort ($n = 27$ matches). For $R*_{FOS}$, R_{FOS} and R_{QUERY} we have enough data for a modeling effort to yield robust results. The random samples of $n = 20$ papers from each of these data sets reveal that $R*_{FOS}$ is skewed towards cognitive neuroscience. For both R_{FOS} and R_{QUERY} the random sample suggest that the bulk of data are actual replication studies. However, a manual check of the top cited articles ($c_5 > 100$) suggests that we have some bias in our data for both R_{FOS} and R_{QUERY}. Although the idiosyncrasies overlap to some extend, we observe different biases in the two sets of data, and we believe that results that are consistent across both should be robust.

A.1.8 Sampling Statistics

This section reports sampling statistics for the models conditioned on both the R_{FOS} and the R_{QUERY} data sets with three different candidate models.

1. CM1: zero-inflated Poisson (ZIP)
2. CM2: gamma-Poisson (negative binomial)
3. CM3: zero-inflated gamma-Poisson

For models conditioned on both R_{FOS} and R_{QUERY} data sets, the model using the gamma-Poisson (or negative binomial) show better sampling and convergence statistics. For the model conditioned on the R_{FOS} all $\hat{R}$ values are effectively 1 and way below typical thresholds (1.05) used in the literature (Fusaroli et al., 2021). Both the number of effective samples for the tail and the bulk of data are high, and the minimum number of effective samples for any parameter is 963.45 which is way above typical thresholds (200) used in the literature (Fusaroli et al., 2021). In the methods section we showed that the chains do not exhibit auto-correlation. We do observe some high Pareto-k values ($n = 89$) indicating that we have influential observations.

For the model conditioned on the R_{QUERY} data all $\hat{R}$ values are effectively 1 and the minimum number of effective samples for any parameter is 997.89. The MCMC chains do not show auto-correlation, but as in the model conditioned on R_{FOS} data we do observe a number of high Pareto-k values ($n = 191$).

For the model conditioned on R_{FOS} data, both $CM1$ and $CM3$ sample worse than $CM2$. For $CM1$ we observe a lower number of effective samples (minimum is $n = 439.37$) which is still above recommendations of at least $n = 200$. In addition however, this model has a much larger number of high Pareto-k values ($n = 859$) indicating that the model is unable to predict outlier values. For $CM3$ the number of effective samples for the worst-sampled parameter is $n = 191.64$ which is below recommendations.

For the model conditioned on R_{QUERY} data, we also observe that $CM1$ and $CM3$ sample worse than $CM2$. For $CM1$ we observe slightly higher $\hat{R}$ values (1.03), the minimum number of effective samples for any parameter is lower (241.89) and the number of high Pareto-k values is much greater ($n = 1790$). For $CM3$ we the minimum number of effective samples for any parameter is 52.65 which is really low, and we observe $\hat{R}$ values above 1.05. In figure 20 we show the MCMC chains for the ZIP model ($CM1$) conditioned on the R_{QUERY} data to show an example of poor sampling and auto-correlation. This is in contrast to the good MCMC sampling observed for both of the models based on $CM2$.

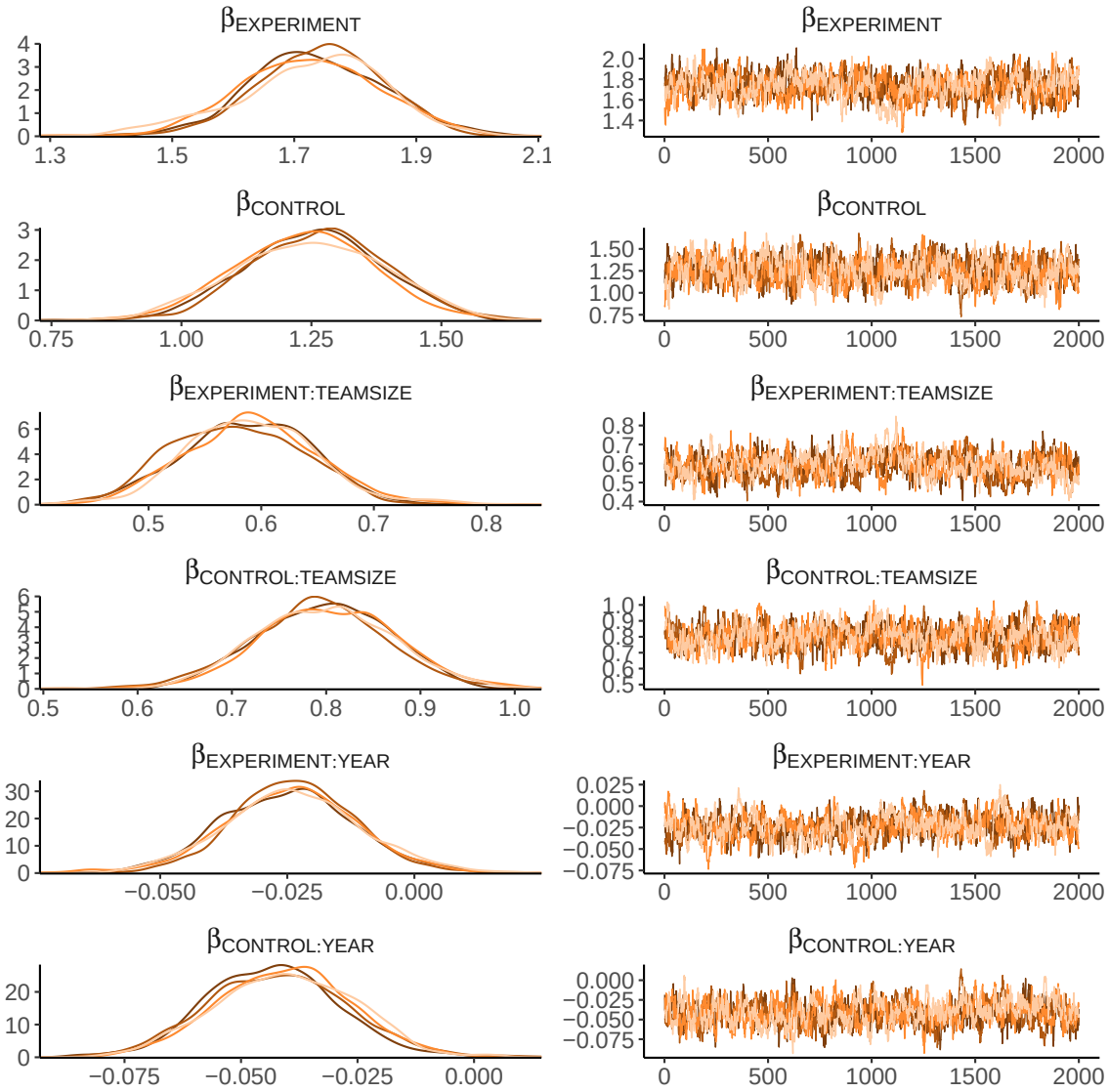

Figure 20: Plots of parameters for β values for Markov Chain Monte Carlo (MCMC) draws for the Zero-Inflated Poisson (ZIP) model conditioned on data from R_{QUERY} data set. Density plot for the four chains on the left and trace plot for four chains on the right. We observe that the four density plots (one for each chain) exhibit clear auto-correlation.

A.1.9 Family Specific Effects

We use an $Exponential(0.5)$ distribution as our prior for the shape (ϕ) of the negative binomial. This is in contrast to the `brms` default which is a gamma $\gamma(0.01, 0.01)$ distribution. These are depicted in Figure 21. We experienced sampling and convergence issues with the $\gamma(0.01, 0.01)$ prior, which is more extreme than the $Exponential(0.5)$ which we ended up using.

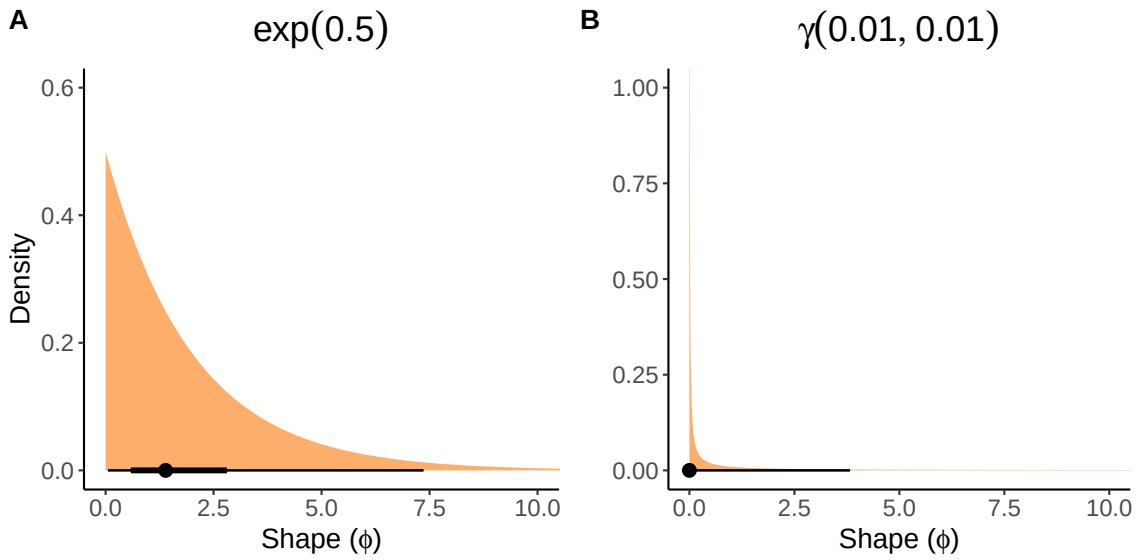

Figure 21: Prior distributions for the shape ϕ parameter. **A)** The prior that we use in our model. **B)** The default `brms` prior. In both cases the x-axis is limited at a maximum of $x = 10$ for visual purposes.

A.1.10 Prior Sensitivity

We conduct prior sensitivity checks for models conditioned on both the R_{FOS} and the R_{QUERY} data. Both have the same parameters and priors. We report prior sensitivity checks for the model conditioned on the R_{FOS} data in the methods section for $S1$. Here we report a similar plot for the model conditioned on R_{QUERY} data (see figure 22).

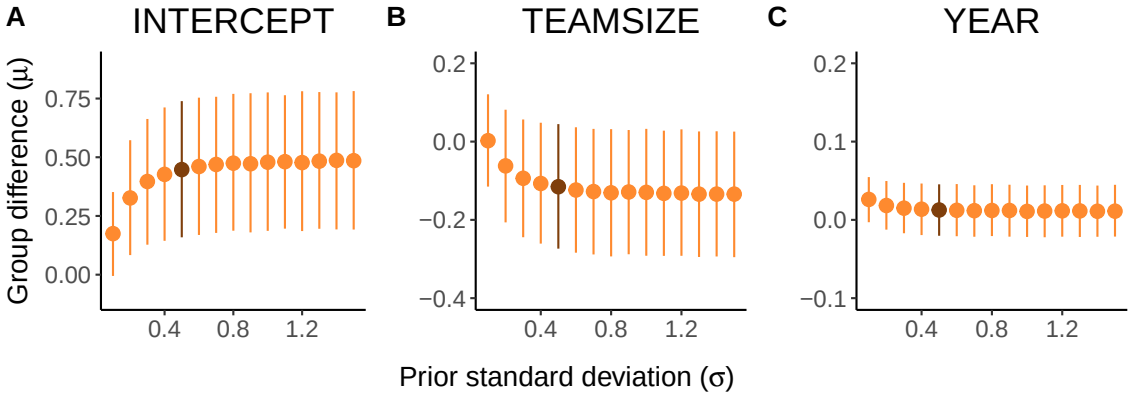

Figure 22: Prior sensitivity checks for the posterior difference between groups of our population-level effects. In all cases, the dot is the estimated difference between groups and lines are 95% credibility intervals. Dark color marks the standard deviation used in the model that we report and light color marks the sensitivity tests. **A)** $INTERCEPT$ difference between groups. **B)** $TEAMSIZE$ difference between groups (interaction effect). **C)** $YEAR$ difference between groups (interaction effect).

A.1.11 Prior and Posterior Predictive Checks

We conduct prior- and posterior predictive checks for models conditioned on the R_{FOS} data in the methods section for $S1$. Here we report the same plots for the model conditioned on the R_{QUERY} data set. Figure 23 shows ungrouped prior- and posterior predictive checks. The prior predictive check and all of the posterior predictive checks look reasonable, although the model does underestimate the frequency of $c_5 = 0$ somewhat. Figure 24 shows posterior predictive checks by condition (control, experiment). Again, the posterior checks look reasonable, although there is a clear regularization in the sense that the model produces a lower mean for the experiment group than what is actually observed, and the model produces a higher mean for the control group than what is actually observed.

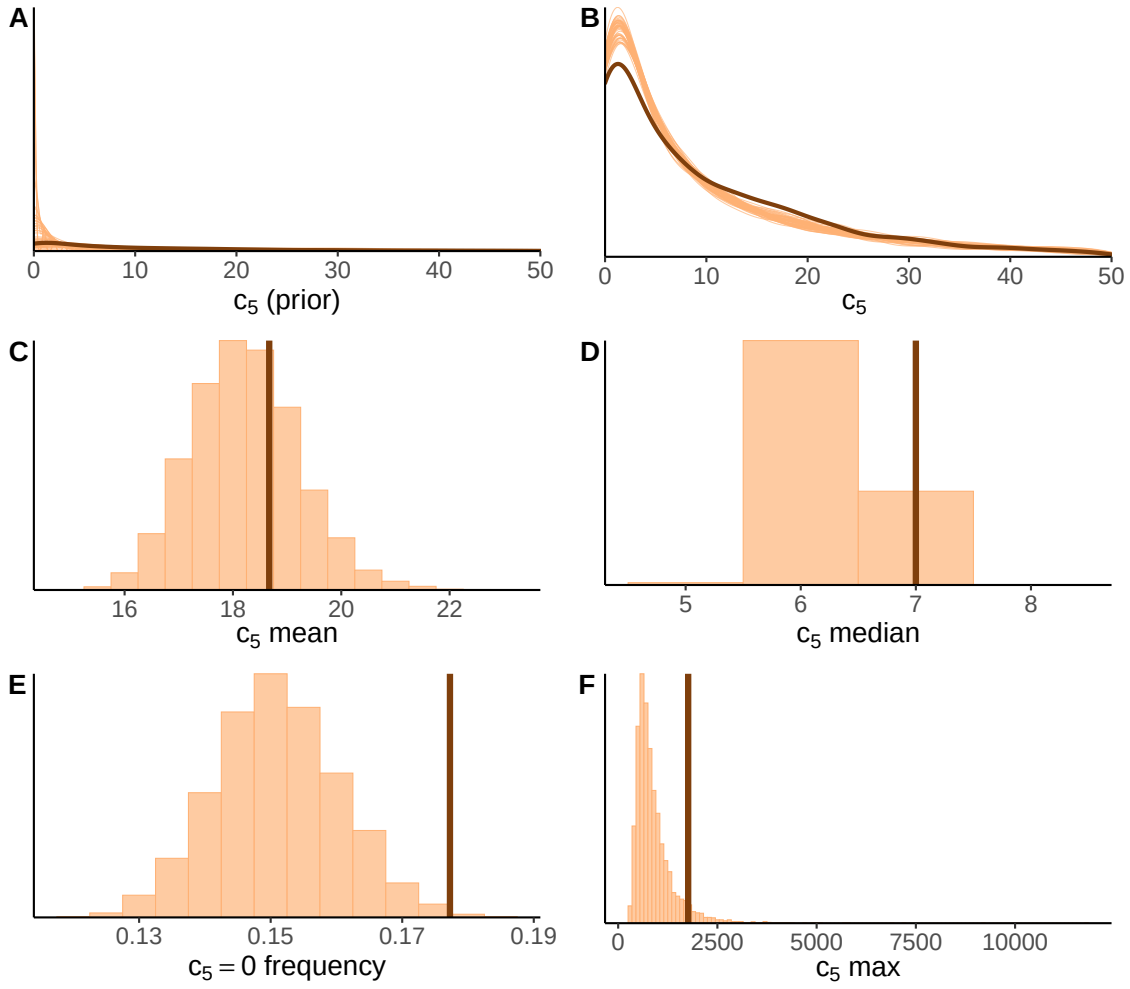

Figure 23: Prior and posterior predictive checks for the model conditioned on R_{QUERY} data. In all cases dark orange is the distribution (or value) of the actual data and light orange is samples from prior- or posterior distributions. **A)** Prior predictive distribution with x axis cutoff at $c_5 = 50$ with observed distribution overlaid. **B)** Posterior predictive distribution with x axis cutoff at $c_5 = 50$ with observed distribution overlaid. **C)** Posterior predictive distribution of the mean with observed value highlighted **D)** Posterior predictive distribution of the median with observed value highlighted. **E)** Posterior predictive distribution of the fraction of zero's with observed value overlaid. **F)** Posterior predictive distribution of the maximum value with observed maximum value overlaid.

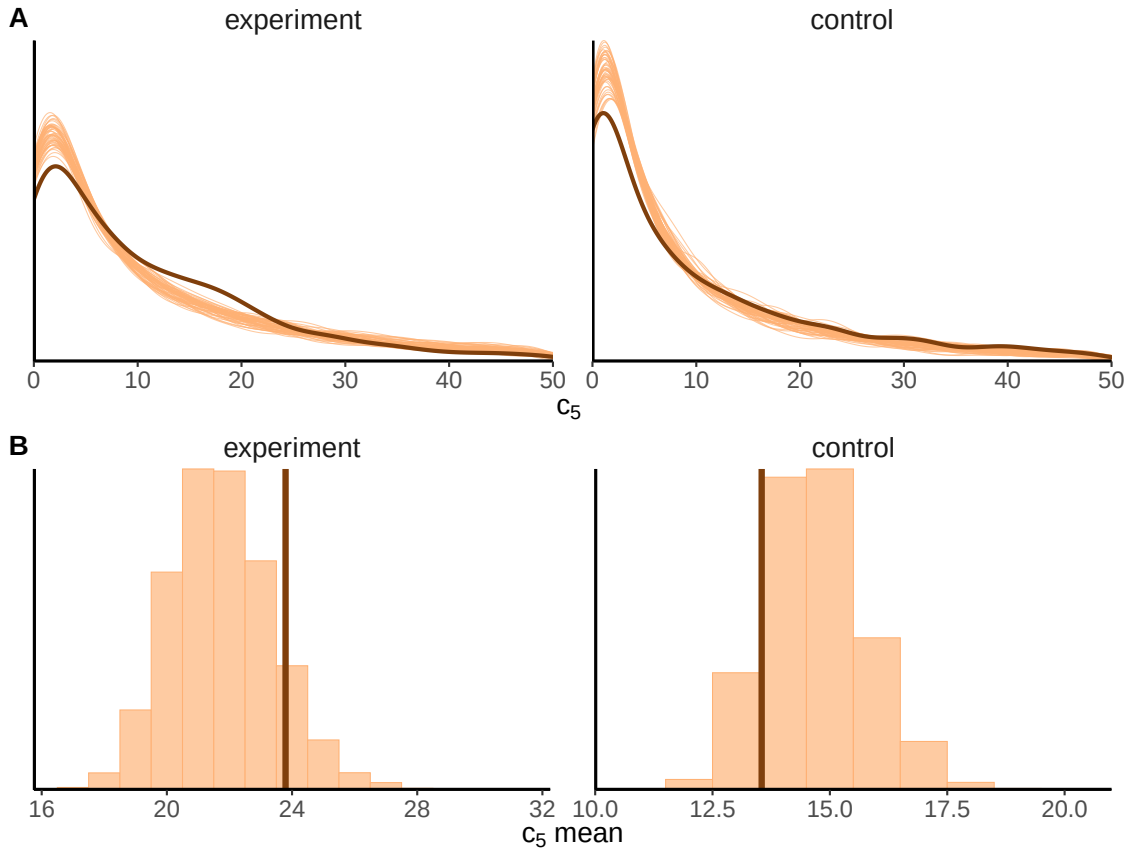

Figure 24: Grouped posterior predictive checks for the model condition on R_{QUERY} data. **A)** Density plot of the experiment group with 50 posterior predictive draws (light orange) and the actual posterior distribution of the experiment group overlaid (dark orange). **B)** Same for control (non-replication group). **C)** Histogram showing the mean of the experiment group for 50 posterior predictive draws (light orange) and line showing of mean for the actual experiment group (dark orange). **D)** Same for control (non-replication) group.

A.1.12 Updating Checks

We conduct prior-posterior updating checks for models conditioned on both the R_{FOS} and the R_{QUERY} data. The updating checks for the population-level effects of the model conditioned on R_{FOS} data is reported in methods section for $S1$. The updating checks for the population-level effects of the model conditioned on R_{QUERY} are shown in Figure 25.

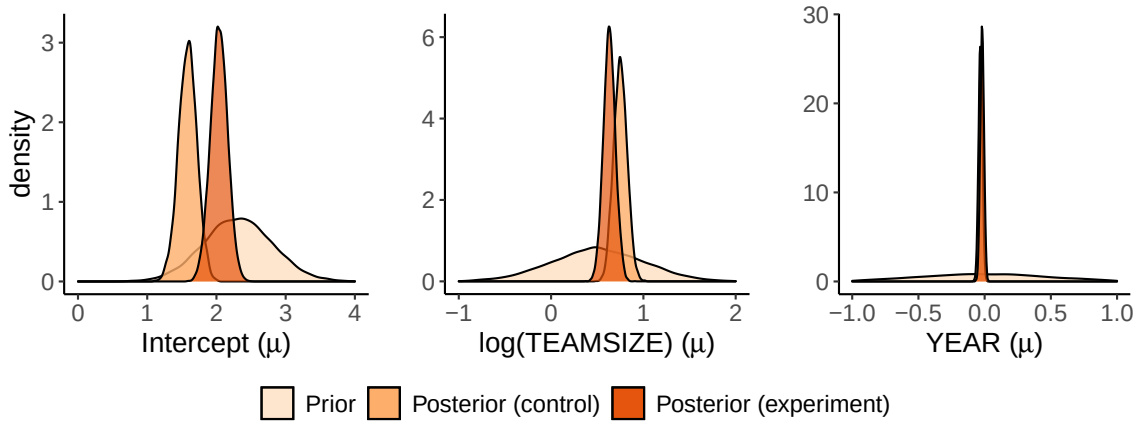

Figure 25: Prior-posterior updating checks for the population-level effects of the model conditioned on R_{QUERY} data. Since the prior for each group is the same, we plot this only once. **A)** Updating check for the $INTERCEPT$ parameter. **B)** Updating checks for $TEAMSIZE$ parameter. **C)** Updating check for the $YEAR$ parameter.

In figure 26 we show prior-posterior updating checks for group-level effects and the family-specific parameter of the model conditioned on R_{FOS} data (for R_{QUERY} see figure 27). We observe that the posterior has lower variance than the prior in all cases. For the updating plot of the standard deviation of random intercepts (figure 26A) we have a posterior for each group (control, experiment) and we might note that we observe slightly more variance in the experiment group. For the correlation between random effects (figure 26B) the posterior is only slightly narrower than the prior, indicating that there is only a weak (and uncertain) correlation between the random intercepts for each group. For the family-specific shape (ϕ) parameter (figure 26C) we observe that the prior is rather broad, although we placed a more informative prior on the shape parameter than the `brms` default. However, the model has updated following conditioning on data, and as we do not observe problems associated with sampling and computation this is fine. Because we have more data

in the R_{QUERY} data set than in the R_{FOS} data set, the variance is slightly lower for the updated parameters (posterior distributions) in this case.

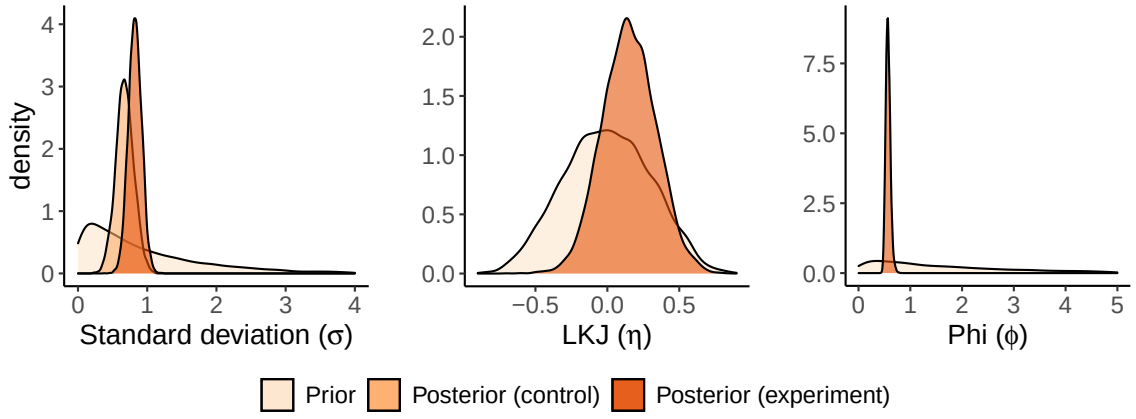

Figure 26: Prior-posterior updating checks for the group-level and family-specific effects of the model conditioned on R_{FOS} data. **A)** Standard deviation (σ) of our group-level effects. **B)** Correlation (R) of random effects. **C)** Family-specific shape (ϕ) parameter.

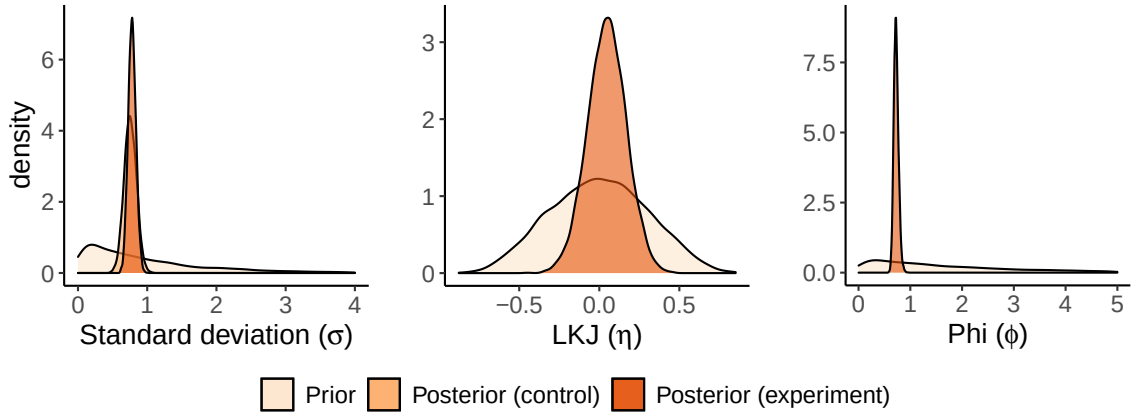

Figure 27: Prior-posterior updating checks for the group-level and family-specific effects of the model conditioned on R_{QUERY} data. **A)** Standard deviation (σ) of our group-level effects. **B)** Correlation (R) of random effects. **C)** Family-specific shape (ϕ) parameter.

A.1.13 Exploratory Data Analysis (EDA)

Below are plots of exploratory data analysis (EDA) for the R_{QUERY} data set. We present similar plots for R_{FOS} in the main text. In the first plot (figure 28) we investigate the outcome distribution (c_5) for both replication (experiment) and non-replication (control) condition. In the second plot (figure 29) we investigate the relationship between c_5 and our predictors, including the effect of log-transforming team size ($TEAMSIZE$).

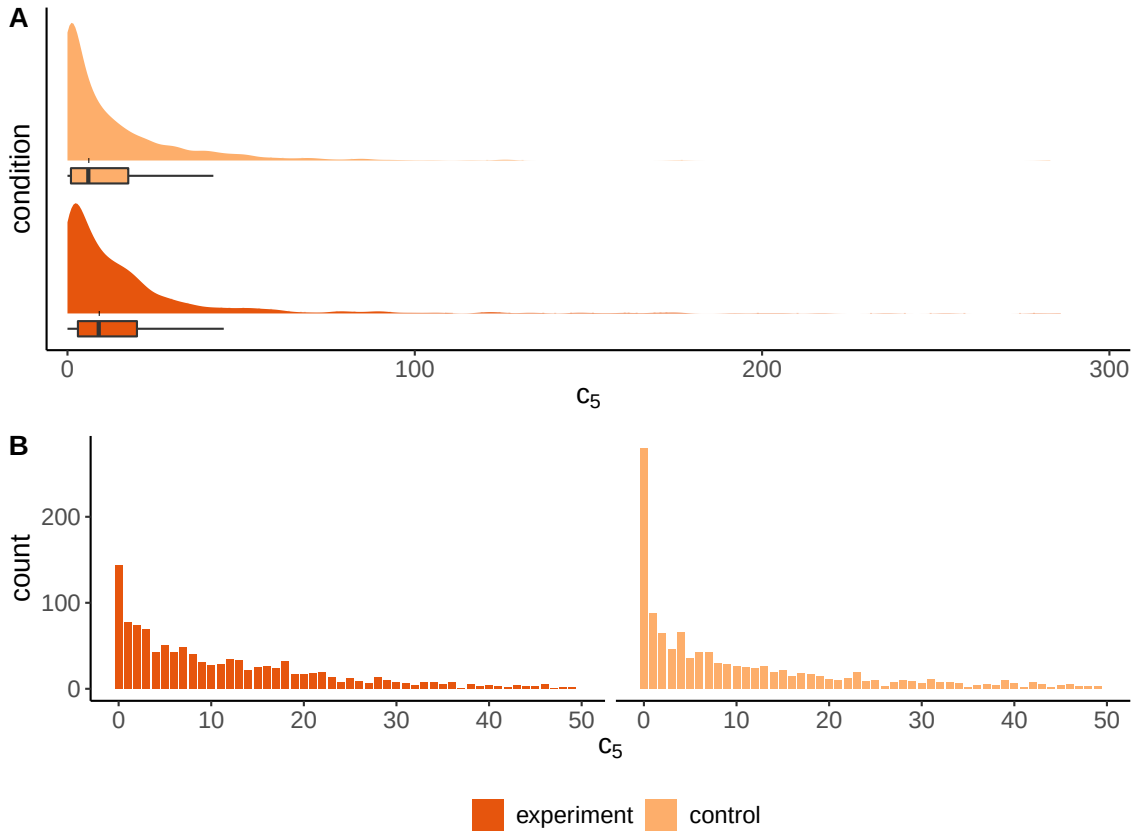

Figure 28: The figure visualizes the outcome distribution c_5 for both replication studies (experiment) and non-replication studies (control) and is based on R_{QUERY}. **A)** density plots and box-plots for each category. In order for the plot to be readable, we omit all studies with $c_5 > 300$. **B)** raw distributions for all studies with $c_5 <= 50$.

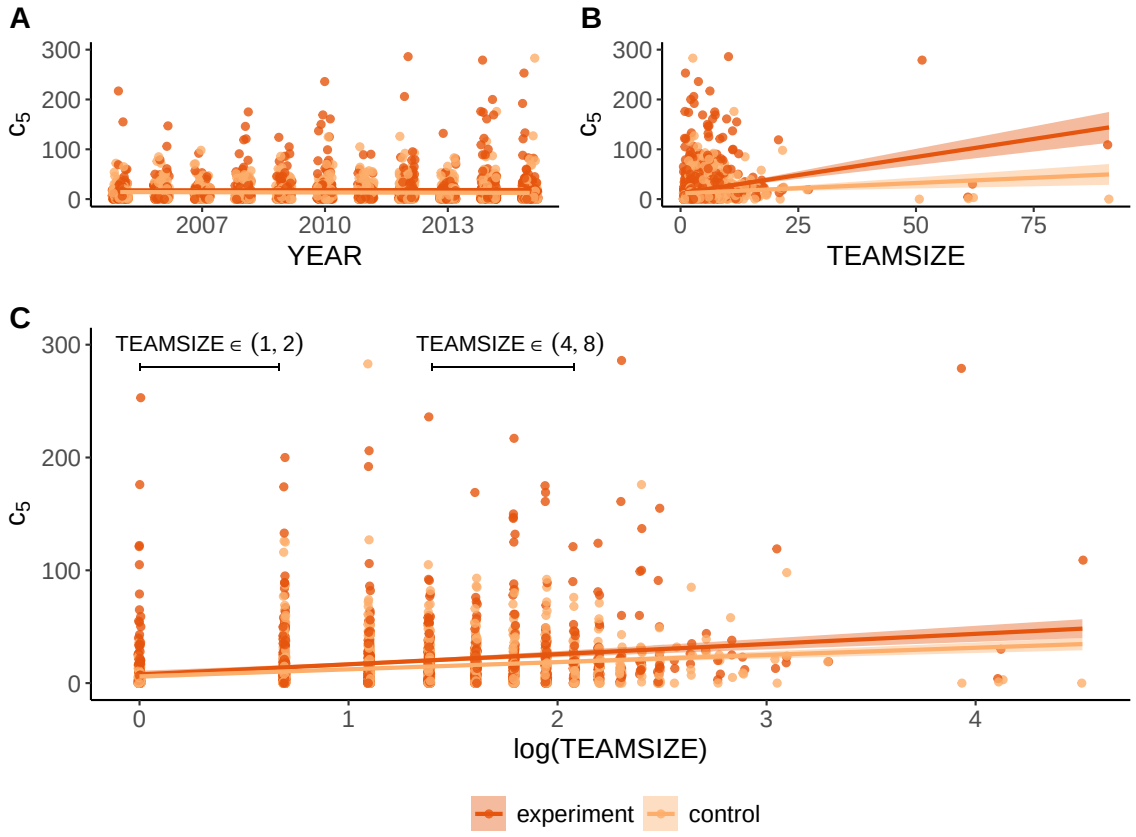

Figure 29: In all plots we display raw data-points and linear fit (including 95% confidence intervals) for the association between our predictors and outcome. All plots use R_{QUERY}. **A)** scatterplot and linear fit between year of publication ($YEAR$) and outcome (c_5). **B)** scatterplot and linear fit between team size ($TEAMSIZE$) and c_5. **C)** scatterplot and linear fit between the log-transformed team size ($log(TEAMSIZE)$) and c_5. Annotations emphasize the fact that for the log-transformed variable, the distance between a solo-authored paper and a duo-authored paper, and the distance between 4 and 8 authors is the same.

A.1.14 Selection Bias

The MAG has different document types, one of which is called ”Repository”. Repository papers are most often preprints. Importantly, if a paper is published first as a preprint (”Repository”) and subsequently as a peer-reviewed article (”Conference” or ”Journal”) then those two instances of the paper are linked by a ”FamilyId” in the MAG. This allows us to look at the fraction of papers that are published only as a preprint (”Repository”) and the fraction of papers that are first published as preprint (”Repository”) and subsequently as a peer-review article (”Conference” or ”Journal”). We start by selecting all preprints (”Repository”) published in psychology between 2005 and 2019 (including both full years). We restrict the sample to 2019 rather than 2020 because this allows for time for the preprint to be published as a journal article or at a conference subsequently (e.g. in 2020 and the first half of 2021). We then select all of the preprints that contain ”replicat” in their title. For both datasets (preprints that match ”replicat*” and all preprints) we find the fraction of the dataset that can be linked to another paper through a ’FamilyId’ and for which the other paper is (i) published as either a Journal article or at a Conference and (i) the other paper in the family is published subsequently. For the full data set we observe 14.770 out of 71.836 preprints that are subsequently published as either journal articles or at a conference (20.56%). For the preprints which contain ”replicat*” in their title we observe 36 out of 126 preprints that are subsequently published (28.57%). Although very few preprints in psychology (2005-2019) use the term ”replicat” in their title, this control check does not indicate that there is a bigger selection bias against publishing replication studies.

A.1.15 Computation

All Bayesian modeling and computation reported in this work relies on `brms` (Bürkner, 2017). `brms` is a package for the programming language *R* which implements Bayesian Multilevel Models using *Stan*. The `brms` package does not fit models itself, but relies on *Stan* as back-end (Bürkner, 2017). As such, `brms` has access to all samplers implemented in *Stan*, and we follow the default and use the state-of-the-art No-U-Turn Sampler (NUTS) (Gelman et al., 2020). Hamiltonian Monte Carlo (HMC), which NUTS is an extension of, is computationally more expensive than other algorithms, but it produces samples of much higher quality, which more than compensates for the computational price (Bürkner, 2017).

Using the NUTS algorithm we conduct full posterior estimation. NUTS is a Markov Chain

Monte Carlo (MCMC) algorithm which explores posterior space efficiently using gradient computation (Gelman et al., 2020; Bürkner, 2017). The chain simulation operates in multiple stages. The first important phase is warm-up, which serves two main purposes; (i) reduce the influence (bias) of starting values for sampling and (ii) provide information about the distribution which is used to tune parameters (Gelman et al., 2020). The sampling algorithm necessarily starts at some values in parameter space, but before accepting actual samples for inference, we want the model to be calibrated such that samples are taken from regions of parameter space closer to where the log posterior density is close to its expected value (Gelman et al., 2020). The second point (tuning) can be important, in particular to avoid divergences (https://www.rdocumentation.org/packages/brms/versions/2.16.3/topics/brm). We apply two tweaks to the NUTS sampler. We adjust the `brms` argument "adapt_delta" from 0.8 (the current default) to 0.99. This slows down the sampler (i.e. we pay more for each sample) but it also decreases the number of divergent transitions, which can threaten the validity of posterior inference (https://www.rdocumentation.org/packages/brms/versions/2.16.3/topics/brm). We also increase the "max_treedepth" argument from 10 (the current default) to 20. We allow the model 2000 steps/samples of warm-up which is double the default in `brms` and we run 4 MCMC chains.

A.2 S2: Supplementary Information

A.2.1 Prosocial Language

The list of prosocial terms is the same as the one employed by Murphy et al. (2020). Their list can be accessed here (https://github.com/everyxs/openScience/blob/master/code-data/input/Lancet%20Dictionaries.csv) The list includes the following 127 terms:

accepting, accommodat, affect, agreeable, aid, altruis, appreciat, approachable, assist, benefit, benevolen, biodivers, care, caring, charit, collective, commun, compassion, compliment, concern, confide, conscien, conservation, considerate, contribut, cooperat, cope, coping, courteous, courtesy, defend, dependab, dignity, donat, earth, ecolog, education, egalitar, empath, empower, encourag, environment, equal, ethic, everybod, everyone, facilitat, fair, forgiv, freed, genero, gentle, genuin, giv, goodhearted, greater good, guard, harmon, help, helpful, honest, honorable, honourable, hospit, human, impartial, inspiring, integrat, integrity, interact, invit, involv, justice, kids, kindness, listen, loyal, moral, NGO, nice, nonjudgmental, nonprofit, not-for-profit, nurtur, peace, philanthrop, prais, prejud, protect, reciproc, relia, relied, rely, respectful, responsib, responsiv, righteous, rights,

role model, selfless, sensitiv, serv, share, shari, shield, sincer, societ, solidarit, support, sustainab, sympath, taught, teach, team, tender, the people, therap, thoughtful, tolera, trust, tutor, underst, universal, unprejudiced, upright, virtuous, volunteer

A.2.2 Stopwords

We use the nltk list of English stopwords. The list contains the following 179 terms:

i, me, my, myself, we, our, ours, ourselves, you, you're, you've, you'll, you'd, your, yours, yourself, yourselves, he, him, his, himself, she, she's, her, hers, herself, it, it's, its, itself, they, them, their, theirs, themselves, what, which, who, whom, this, that, that'll, these, those, am, is, are, was, were, be, been, being, have, has, had, having, do, does, did, doing, a, an, the, and, but, if, or, because, as, until, while, of, at, by, for, with, about, against, between, into, through, during, before, after, above, below, to, from, up, down, in, out, on, off, over, under, again, further, then, once, here, there, when, where, why, how, all, any, both, each, few, more, most, other, some, such, no, nor, not, only, own, same, so, than, too, very, s, t, can, will, just, don, don't, should, should've, now, d, ll, m, o, re, ve, y, ain, aren, aren't, couldn, couldn't, didn, didn't, doesn, doesn't, hadn, hadn't, hasn, hasn't, haven, haven't, isn, isn't, ma, mightn, mightn't, mustn, mustn't, needn, needn't, shan, shan't, shouldn, shouldn't, wasn, wasn't, weren, weren't, won, won't, wouldn, wouldn't,

A.2.3 Topics & Communities

We show all 100 topics from our topic model in chronological order in table 4 below. For the 50 topics which we select for further analysis, we provide the labeled community in the "Community" column. For the 50 topics that were dropped, the value in this column is "<NA>". We show the top 10 most strongly weighted words for each topic.

Topic	Community	10 most weighted words
0	Culture & Training	public, trust, health, result, engagement, right, perspective, build, author, increase
1	Reform Psychology	learn, preregistration, decade, student, power, low, method, lesson, course, teach

2	Publication	review, peer, politic, postpublication, oath, reviewer, peerreview, process, comment, sign
3	Culture & Training	initiative, reality, openness, reviewer, launch, pro, collaboration, student, interesting, international
4	Publication	paper, write, code, available, link, cite, submit, comment, author, really
5	Reform Psychology	social, medium, resolve, humanity, prize, responsible, badly, specify, psych, theory
6	Data & Policy	health, dataispower, digitalhealth, artofinsight, hippocraticoath, datasaveslive, precisionmedicine, bigdatame, responsibleai, patientsfirst
7	Culture & Training	scientist, trust, explain, badge, strike, error, young, citizen, mistake, rarely
8	<NA>	workshop, reproducible, day, th, slide, rsvp, march, national, june, openaire
9	<NA>	digest, weekly, openpharma, happen, peerreview, pharma, openforall, medical, clinicaltrial, datashare
10	<NA>	promote, survey, institution, finding, suggest, individual, evaluation, mean, india, repository
11	<NA>	post, petersuber, interesting, guest, ger, eng, write, washington, comment, late
12	<NA>	way, forward, catalyse, consortia, lead, guardian, communicate, address, assess, long
13	Reform Psychology	transparency, openness, increase, prize, encourage, incentive, integrity, guideline, commitment, improve

14	Data & Policy	europe, elsevi, horizon, corrupt, monitor, eu, sparc, analysis, let, dream
15	<NA>	week, openpharma, fan, digest, event, berlin, member, enjoy, welcome, celebrate
16	Culture & Training	practice, survey, principle, questionable, adopt, key, reward, versione, student, care
17	Publication	knowledge, decolonization, human, build, mind, step, enhance, data, user, repository
18	<NA>	challenge, opportunity, career, early, collaboration, tip, benefit, preregistration, perspective, address
19	<NA>	movement, value, improve, microsoft, commit, discuss, pace, slow, reform, culture
20	Culture & Training	change, culture, climate, incentive, reward, cultural, life, behaviour, norm, require
21	Data & Policy	share, code, data, presentation, experience, online, care, benefit, privacy, material
22	<NA>	good, thing, bad, point, idea, example, fail, luck, voice, generation
23	<NA>	university, library, utrecht, education, role, principle, eua, transition, finland, press
24	<NA>	join, come, discussion, event, webinar, discuss, pm, th, interested, panel
25	<NA>	global, unesco, recommendation, service, library, sustainability, development, consultation, coalition, scoss
26	<NA>	story, openedu, daily, twitter, scienceopen, personal, prize, star, hard, mediatalk

27	<NA>	tool, prize, digital, source, software, track, opensource, outbreak, roadmap, joint
28	OSF	framework, osf, webinar, introduction, connect, entire, improve, cycle, commons, scholarly
29	<NA>	european, cloud, eosc, commission, eu, launch, digital, monitor, expert, declaration
30	<NA>	know, let, thing, interested, love, ask, twitter, hub, ve, figshare
31	<NA>	thank, fan, late, author, add, slide, lot, follow, amazing, link
32	Publication	scientific, information, process, publication, result, understand, revolution, american, discovery, crack
33	<NA>	rickypo, action, eu, prize, amsterdam, data, manifesto, openaire, trend, infrastructure
34	Reform Psychology	reproducibility, openness, improve, critical, continue, replicability, increase, essential, solve, nosek
35	Reform Psychology	psychology, resolve, explain, forward, bad, theory, preregistration, specify, badly, personality
36	Publication	openaccess, oaweek, openresearch, openpharma, scholcomm, publication, librarian, publisher, repository, interview
37	Culture & Training	training, foster, handbook, course, event, online, management, fosteropenscience, library, method
38	OSF	free, software, source, online, feel, course, book, register, webinar, download

39	Culture & Training	make, reality, available, accessible, easy, sure, transparent, field, sense, reproducible
40	Data & Policy	innovation, technology, citizen, opportunity, commission, responsible, tech, aiforgood, digital, responsibleai
41	Culture & Training	community, build, member, twitter, utrecht, plo, resource, role, network, grow
42	Publication	center, service, launch, brand, announce, osf, release, cos, nosek, openness
43	Culture & Training	researcher, career, early, succeed, young, survey, discuss, incentive, earlycareer, provide
44	<NA>	job, current, opening, zpid, movement, state, ad, repository, manager, market
45	OSF	use, link, osf, term, create, method, file, analysis, student, presentation
46	Culture & Training	research, reproducible, management, assessment, medical, integrity, quality, accelerate, transparent, education
47	<NA>	today, day, daily, pm, slide, meeting, tomorrow, student, present, follow
48	Publication	publish, result, imagine, understand, scholarly, perish, choose, finding, trial, fail
49	<NA>	academic, librarian, publisher, library, scholcomm, build, structure, frustration, phd, student
50	Data & Policy	policy, platform, analysis, risk, alienate, highlevel, development, announce, shed, universal
51	<NA>	talk, slide, podcast, video, tomorrow, episode, ill, let, series, watch

52	<NA>	question, answer, ask, student, trust, interesting, chronicle, san, biological, twitter
53	Publication	impact, benefit, economic, factor, enhance, user, significant, evidence, career, increase
54	<NA>	say, thing, right, mean, eu, error, author, really, agree, head
55	<NA>	work, group, hard, student, team, personal, credit, light, real, secret
56	Data & Policy	report, register, registered, final, preregistration, result, state, stateofopendata, trial, progress
57	<NA>	check, resource, link, page, list, collection, interested, date, video, slide
58	<NA>	year, ago, student, past, ve, resolution, old, happy, million, progress
59	Data & Policy	datascience, bigdata, ai, iot, machinelearne, iiot, defstar, infographic, tech, mpgvip
60	Publication	article, author, badge, write, issue, aspect, statistical, collection, cite, ve
61	<NA>	think, term, school, ve, bad, really, lot, important, interesting, issue
62	<NA>	great, piece, idea, thread, resource, example, mind, lot, unlike, really
63	<NA>	world, understand, imagine, animal, mooc, little, develop, real, network, vision
64	Publication	publishing, platform, scholarly, launch, model, communication, african, opensource, forprofit, commission

65	Data & Policy	datum, management, sharing, analysis, available, key, versione, code, care, repository
66	<NA>	covid, pandemic, lesson, life, save, coronavirus, thread, reminder, corona, danger
67	<NA>	news, story, mcdawg, figshare, sharmanedit, oatp, creativecommon, plo, late, na
68	<NA>	like, really, feel, sound, thing, idea, yes, connect, try, student
69	<NA>	plan, national, action, education, eu, strategy, wikidata, academy, france, netherland
70	Reform Psychology	psychologys, away, wish, mean, field, real, debate, crisis, atlantic, lining
71	Culture & Training	support, library, infrastructure, eu, announce, cern, cos, member, digital, building
72	<NA>	blog, openaire, collaboration, scienceopen, eng, ger, plo, guest, conversation, star
73	<NA>	oa, link, infrastructure, action, foster, eu, english, green, source, guide
74	<NA>	rt, briandavidearp, infographic, routledgepsych, datascience, rtmachinelearnbot, defstar, rstat, learnrinaday, characteristic
75	Reform Psychology	crisis, replication, fix, statistical, solution, theory, measurement, error, failure, bad
76	<NA>	look, forward, sign, tomorrow, method, interesting, failure, group, solve, consider

77	<NA>	read, interesting, openpharma, interested, book, worth, digest, piece, openforall, late
78	Reform Psychology	case, replicationcrisis, method, miss, bigdata, bad, bias, badge, psych, late
79	<NA>	help, succeed, advance, solve, build, pandemic, zika, prepare, phdchat, citizenscience
80	<NA>	physics, bblogrt, scicomm, rtdnr, mattitgd, tgd, model, dark, arxiv, deleuze
81	<NA>	fair, principle, osfair, eosc, figshare, guide, september, infrastructure, digital, network
82	Publication	journal, editor, badge, editorial, club, author, guideline, predatory, submit, submission
83	Publication	preprint, service, server, brand, launch, asapbio, osf, author, trust, release
84	<NA>	new, announce, frontier, normal, launch, statistic, likely, true, idea, feature
85	<NA>	start, interview, revolution, point, matter, podcast, student, getting, interested, tenure
86	<NA>	psychedelic, snrtg, psychat, wellnesswednesday, amanda, scicomm, feilde, mind, countess, inside
87	<NA>	people, lot, thing, mean, maybe, twitter, tell, point, understand, really
88	<NA>	big, problem, fan, issue, solution, solve, step, address, reward, deal
89	Culture & Training	need, coalition, skill, incentive, infrastructure, funder, break, expert, issue, create

90	Publication	society, royal, launch, submission, human, festival, accept, citizen, introduce, raise
91	<NA>	want, fix, replicate, darpa, solve, wire, early, career, try, build
92	Data & Policy	opendata, openresearch, data, figshare, opensource, openacces, dataset, rdm, state, mt
93	Publication	access, publisher, source, publication, information, instrumentalist, pursuit, link, argument, book
94	Reform Psychology	study, replicate, fail, replication, result, preregistere, repeat, finding, effect, case
95	Reform Psychology	future, mean, shape, communication, scholarly, vox, lab, website, african, data
96	<NA>	time, space, consciousness, ipod, iphone, education, high, ve, real, lab
97	<NA>	edtech, sciencematter, itrtg, scicomm, flockbn, makesciencegreatagain, biophoton, biophotonic, climate, human
98	OSF	project, fund, lab, osf, student, cool, collaboration, collaborate, presentation, grant
99	Reform Psychology	conference, osc, presentation, th, international, berlin, speaker, online, poster, march

Table 4: Topics & Communities

A.2.4 Centrality

Top 10 accounts in the network of ”Reform Psychology” accounts ranked by eigenvector centrality is displayed in figure 30.

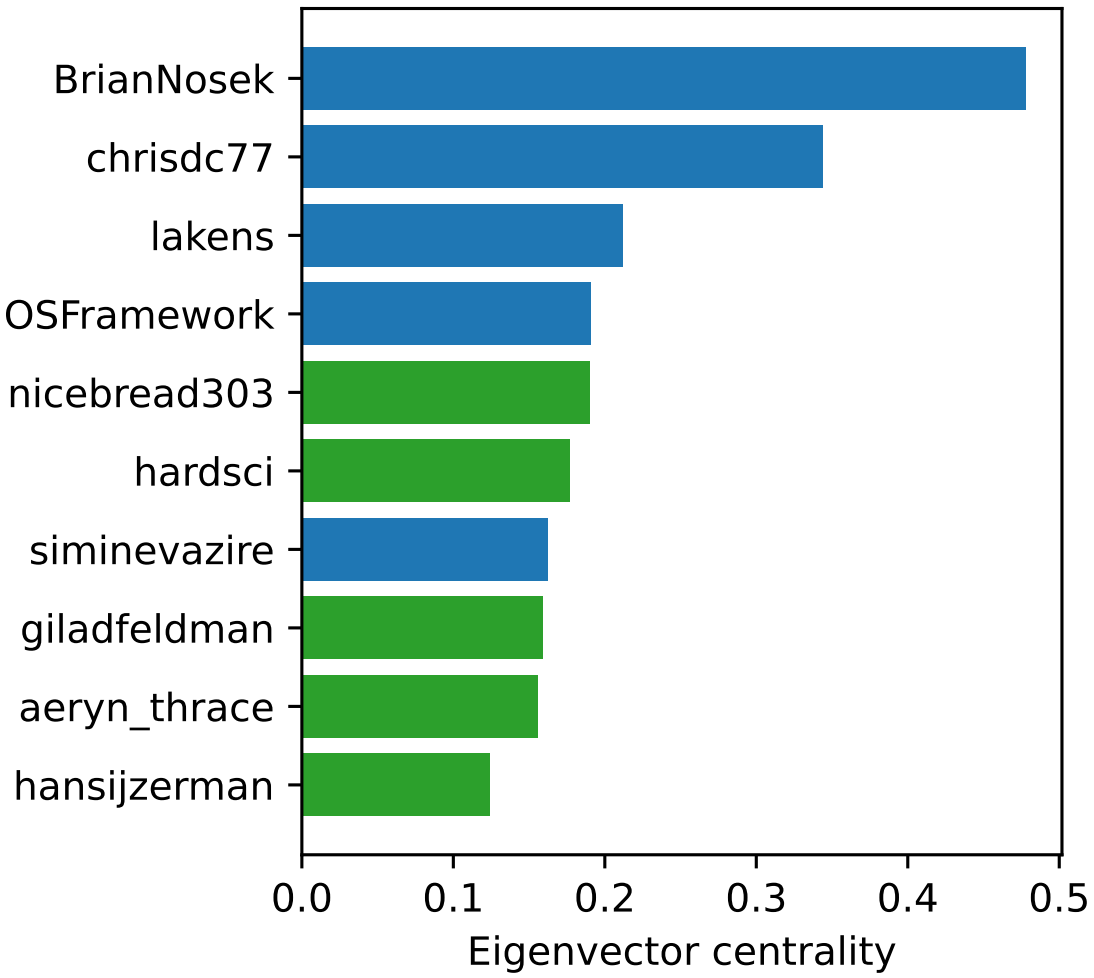

Figure 30: Top 10 most central accounts ranked by eigenvector centrality. Brian Nosek (@BrianNosek) the most central account, consistent with the centrality ranking by weighted degree. Olivia Guest (@o_guest) drops out of the top 10 when we rank by eigenvector centrality.

A.2.5 Data Curation

There are caveats at multiple levels of our analysis. Firstly, there is the caveat that our data is contingent upon the queries that we chose (Open Science and Replication Crisis) as well as choices in data curation. Recall that for our main corpus we restricted ourselves to accounts with at least two posts in both ”Open Science” and ”Replication Crisis”. This means that we are biased to find accounts which are involved in the ”Open Science” community. It is possible that some ”Reform Psychology” voices do not engage with discussions of ”Open Science”. However, we do not believe that this should be the case for the majority of reformers, since the ”Open Science” groups is very central to the reform agenda. This is supported by the fact that we match both proponents (e.g. Brian Nosek) and critics (e.g. Olivia Guest). We attempted other queries, but generally encountered issues of matching too much (e.g. ”replication” matches a lot of COVID-19 related material) or too little (e.g. ”replicability” returns few matches). As such, the final queries (Open Science and Replication Crisis) represent our best attempt at isolating ”Reform Psychology” while obtaining a meaningfully large corpus.

A.2.6 Semantic Analysis Limitations

Naturally, there are also limitations to the ”objectivity” of the semantic analysis (including topic modeling and semantic linkage networks). There are both choices of exclusion (i.e. we exclude 50 out of 100 topics) and then there is labelling of topics and communities (e.g. ”Reform Psychology” could also reasonably have been called ”Methods & Statistics in Psychology”). I view the analysis as exploratory, and as substantiating historical and sociological claims, such as those in Flis (2022), with an analysis that is supported by data. However, this does not mean that the analysis is exempt from subjective choices, and it is possible that different choices at various levels could have somewhat altered the final representation that we suggest here.