Comparative Analysis of Registered Reports and the Standard Research Literature

PhD Thesis

By

Aoife O'Mahony

In Partial Fulfillment

of the Requirements for the Degree of

Doctor of Philosophy

School of Psychology, Cardiff University

23rd October 2023

Abstract

The replication crisis revealed high levels of bias and questionable research practices (QRPs) in psychology research. Registered Reports (RRs) have been increasingly adopted as a possible solution, but there has been relatively little evidence of whether this novel publishing format appears to be working as intended to reduce bias and QRPs. This project sought to build a detailed database of RRs and closely matched standard reports (SRs), to investigate whether RRs perform better than SRs on indicators of quality, rigour, and transparency. 170 RRs were gathered, representing psychology, health, and related disciplines. Each RR was matched to 2 SRs and their characteristics were coded and compared between the two article types. Some brief descriptive analyses were also undertaken on a larger total sample of RRs (n = 359) that did not have a comparison sample, and a smaller sample of 12 RRs and 12 SRs was examined for signs of HARKing. Six key findings were observed. First, RRs exhibit lower rates of supported hypotheses and higher rates of unsupported hypotheses compared with SRs. Second, rates of open practices are higher among RRs than SRs. Third, RRs also appear to be more strongly associated with some methodological practices indicative of greater rigour and transparency. Fourth, author demographics and article citation rates revealed few differences between the article types. Fifth, higher citation rates were associated with more positive findings and fewer negative findings within both the SRs and RRs, but there was no statistically significant relationship between the journal impact factor and whether hypotheses were supported, for either article type. Finally, HARKing appeared to be non-existent in RRs, while some evidence of HARKing was observed in SRs, however, replication is needed. Overall, the evidence presented in this work demonstrates that, while there are still areas for improvement, RRs do appear to be working as intended in being associated with improved research practices, although further research is needed to determine causal impacts.

Table of Contents

Acknowledgements

Firstly, I would like to thank my supervisors Chris Chambers and Candice Morey for their consistent support and guidance throughout this process and without whom none of this would have been possible. Thanks also to Richard Morey for helpful conversations during annual reviews.

I am also grateful to the UK Reproducibility Network for the partial funding of this PhD studentship.

I would also like to acknowledge the hard work and dedication of the research assistants, interns and other students who were involved in these projects: Emma Chubb and Catrin McAdam's involvement in the matching and coding, particularly the early refinement of the protocols, were invaluable contributions. Thanks also to Akua Oye Ohemeng Owusu and Geraint Lewis for their involvement in the coding and matching processes. Furthermore, I am grateful for the efforts of the MSc students whose thesis projects informed the approach taken for the work later conducted for chapter 7. I am also particularly thankful to Molly White whose MSc thesis research was instrumental for the work reported in chapter 8.

Finally, I would like to thank my parents and other family members for all their support throughout this process.

Contributions to the Project

The work in this thesis was supported by several research assistants and MSc students. Their contributions are outlined below.

Matching and coding of articles

Emma Chubb completed a summer internship in June 2020 and subsequently continued working on the project as a research assistant. Emma initially provided informal feedback on the second draft of the coding protocol and first draft of the selection protocol before using these to independently match and code a small sample of papers. Following the double coding of this sub-sample, Emma and I compared our coding of this subset of articles and resolved discrepancies qualitatively through discussion. In total Emma was involved in the coding and checking of 21 Registered Reports (RRs) and 30 Standard Reports (SRs). She also matched 29 SRs and checked the matching of a further 9 SRs.

Catrin McAdam joined the team as an research assistant in August 2020. Catrin coded 29 articles (consisting of 14 RRs and 15 SRs) and matched 15 SRs. She was not involved in checking others' matching or coding.

Akua Oye Ohemeng Owusu and Geraint Lewis joined the team for a placement in the summer of 2021 as part of their MSc studies, and both contributed to coding and matching of articles. Oye coded 41 articles in total, consisting of one RR and 40 SRs, and was not involved in checking others' coding. She also matched 32 SRs and checked the matching of 8 SRs. Geraint matched and coded 2 of the SRs included in the database.

Informing citation study reported in Chapter 7

The work reported in chapter 7 was conducted myself, but the approach taken was informed by thesis projects done by a group of MSc students, in relation to citation rates of RRs and SRs. In particular, the choice to use three different sources to obtain the citations rates, and the decision to use the mean of these three sources as the metric to represent citations rates in the chapter 7 analysis, was informed by their having taken this approach in their own

projects. However, the data they collected/coded in their projects was not used in this thesis and instead, I collected the data for this chapter by myself at a later date.

Collaboration on data collection/coding in Chapter 8

The project reported in chapter 8, in relation to HARKing rates, was designed in collaboration with an MSc student (Molly White) as part of her thesis work. This included working together to formulate the research questions and design the study approach, as well as developing and testing the coding approach. While the student in question collected the new SR sample for this chapter and also performed the coding for this HARKing study, I checked each detail of this in depth and provided feedback on any judgements that needed to be refined. The resulting data was used for the results reported in chapter 8.

Chapter 1: Introduction

1.1. Replication Crisis and Questionable Research Practices

In recent years there has been growing concern regarding the rigour and credibility of research within the social and life sciences, with the inability to replicate many research findings leading to the concept of a 'reproducibility crisis' or 'replication crisis' (Maxwell et al., 2015; Pashler & Harris, 2012). This was initially brought to attention following some large-scale failures to replicate key findings across various disciplines (Cheung et al., 2016; Donnellan et al., 2015; Ebersole et al., 2016; Eerland et al., 2016; Harris et al., 2013; Klein et al., 2014; O'Donnell et al., 2018), with psychology receiving particular attention and concern. For example, the large-scale 'Reproducibility Project: Psychology' (Open Science Collaboration, 2015) attempted to independently replicate 100 effects from various areas of psychology but only managed to successfully replicate 39%. These findings have been supported by a more recent machine-learning study of psychology studies published in six major journals over 20 years which suggests that slightly more than half of these psychology papers would fail rather than pass replication tests (Youyou et al., 2023).

Disciplines such as social psychology have been the subject of particular concern, with a successful replication rate of only 25% reported by the Reproducibility Project (Open Science Collaboration, 2015). This broad finding is supported by Youyou et al.'s (2023) work which showed that replicability of psychology papers varied considerably by subfield, with social psychology showing an estimated replication score of 0.37 (37%), a finding that is slightly more encouraging than that previously reported, but still among the lowest of the subfields examined in this study. Other areas such as developmental, cognitive, and clinical psychology received estimated replication rates of 36%, 42%, and 44%, respectively, with areas such as organizational psychology and personality psychology showing only slightly more encouraging rates (50% and 55%, respectively). Overall, the evidence suggests that the widespread concern about the robustness and replicability of psychology's research findings is valid. While the problem is not restricted solely to psychology, the issues noted in this area have received considerable attention due to the apparent scale of the problem.

Contributing factors for this replication crisis are numerous but are thought to include biased research and reporting, overrepresentation of positive results, a lack of methodological rigour, low statistical power, lack of replication efforts, lack of transparency in reporting

(particularly selective reporting and undisclosed flexibility), unwillingness to share data and materials, and publication bias whereby null results are much less likely to be published than statistically significant findings (Ferguson & Heene, 2012; Chambers, 2017; Ware & Munafo, 2015). Intense pressure on researchers to publish prolifically, known as 'publish or perish' culture, has been recognised as being an important factor that may contribute to many of these other issues (Herndon, 2016; Rawat & Meena, 2014; Gopalakrishna et al., 2022; Grimes et al., 2018; Versteege, 2013). Due to this pressure to publish and the extreme emphasis on statistically significant findings in publication decisions, researchers may resort to concerning practices to increase the likelihood of their work being published (Bruton et al., 2020; John et al., 2012; Nosek et al., 2012; Herndon, 2016).

A number of these questionable research practices (QRPs) have become commonplace, including p-hacking and Hypothesizing After Results are Known (HARKing). P-hacking occurs when researchers consciously or unconsciously exploit 'degrees of freedom' in their choice of analysis approach and take advantage of this flexibility to report a desired result as if it were confirmatory (Simonsohn et al., 2014). HARKing, meanwhile, involves presenting a post-hoc hypothesis as if it was specified *a priori* (Kerr, 1998). This is problematic because the hypothesis presented has then been influenced by the data collected and is not representative of the researchers' initial expectations; as a result, HARKing produces unfalsifiable 'hypotheses' and can distort theory. These practices, and other forms of undisclosed flexibility in analysis are particularly concerning as these can inflate reported effect sizes and increase the risk of type 1 error, thereby increasing the likelihood of obtaining false positive findings (Świątkowski & Dompnier, 2017; Simmons et al., 2011; Wicherts et al., 2016).

As noted in chapters 3 and 8, p-hacking and HARKing are common in psychology. This is also true for a range of other questionable practices. For example, John et al. (2012) reports that almost 56% of their survey respondents admitted to collecting more data after checking and seeing that their result was not statistically significant, and almost 46% of respondents admitted to selectively reporting studies that "worked", i.e., that had statistically significant results. Furthermore, 38% reported deciding to exclude data after looking at how this impacted the results, and 27% admitted to reporting an unexpected finding as if it had been predicted *a priori*. Rates of each practice were even higher in the experimental group that were subjected to incentives for truth-telling. However, Fiedler and Schwarz (2016) criticise John et al.'s approach and interpretations. Using a more nuanced approach involving separate

measures of whether researchers had ever engaged in such practices, and the frequency of doing so, they showed lower rates of QRPs than those reported by John et al.

Looking across various other studies conducted on this topic, it is clear that estimates of the prevalence of QRPs vary widely. For example, QRPs were self-reported by 37.7% of researchers who have worked on ego-depletion research (Wolff et al., 2018), but in another study, self-reported rates of using at least one QRP were as high as 88% among Italian psychology researchers, with some arguing that the practices involved were not actually questionable or were defensible (Agnoli et al., 2017). Regardless of the exact prevalence rates, it is clear that use of these questionable practices is common and this undermines the confidence and trust that we can have in published research findings. Particular QRPs (e.g., p-hacking and HARKing) will be discussed in more detail in later chapters.

1.2. Impact of the Replication Crisis on Trust in Research

Understandably, awareness of psychology's low replicability rates and of the widespread use of questionable practices, has undermined confidence in the discipline's validity and credibility, affecting researchers' and students' trust in the discipline as a whole. Knowledge of the replication crisis and/or about the inability to replicate a particular study's findings, also appears to have some influence on the public's trust in research (Wingen et al., 2020; Hendriks et al., 2020). Furthermore, a survey of research participants showed that most considered QRPs to be unacceptable behaviour by researchers, and most viewed more open research practices positively (Bottesini et al., 2022). Other evidence has also shown that receiving information about scientific reforms increased trust in researchers and research, compared with only receiving information about the replication crisis and its causes (Methner et al., 2022). This is supported by Hendriks et al.'s (2020) work showing that study credibility and researcher trustworthiness were rated higher when laypeople learned of replication success, and lower when they learned of replication failure. However, Wingen et al. (2020) report conflicting findings, showing that although knowledge of low replicability did appear to reduce public trust in psychological research, providing explanations for this and information about increasing transparency and recovered replicability did not appear to repair the public's trust. This finding is further supported by Anvari and Lakens (2018) who showed that being informed of replication failures and QRPs did not result in significant differences in trust in future research, compared to the control group, although potentially did reduce trust in relation to past research. Meanwhile, the group in their study that also received information about reforms actually showed lower trust in future research compared to the

control group. Overall, then, the suggestion that awareness of the replication crisis and QRPs may reduce public trust in psychological research appears to be valid, but the evidence for how awareness of research reforms and improvements may influence public trust is mixed. Furthermore, a large-scale study found that members of the public did not consider the findings of Registered Reports (RRs) to be more credible than those of standard reports, regardless of how plausible or implausible these findings were (Costa et al., 2022).

The public is not the only population whose trust in research may be affected by their knowledge of the replication crisis and associated problems. Chopik et al. (2018) report decreased trust in psychology research among students following a single lecture on the replication crisis. Similar results were found by Sacco and Brown (2019) following a short training module about QRPs, although the effects appeared to be relatively short-term. Additionally, Pownall et al. (2023)'s recent quasi-experimental study demonstrated that even engaging in preregistration of their dissertation research did not seem to influence students' attitudes towards the acceptability of QRPs, at least according to the study's quantitative findings. Qualitative interviews conducted as part of the same study did however reveal some greater awareness of the benefits of preregistration for preventing QRPs, while both aspects of the study indicated more encouraging changes in students' knowledge of open science and their motivation to preregister research. In summary, therefore, increasing knowledge of the replication crisis and biased research does seem to impact on both the public and students' attitudes to science and trust in research, while knowledge of, or even engagement in, open practices, does not appear to consistently restore positive attitudes.

1.3. Positive Impacts of the Replication Crisis – the Credibility Revolution

Regardless of the public's views on current reforms, responses to the replication crisis have led to many positive changes in how research is done. This is due to intensive self-examination of particular disciplines and their working culture, the impassioned response of many researchers determined to address the issues and the underlying problems, and the subsequent uptake of more robust and transparent methodological practices. Such changes have led to the newer concept of the 'credibility revolution' (Vazire et al., 2018; Vazire et al., 2022; Korbmacher et al., 2023), involving the adoption of practices designed to increase the rigour and transparency of the work. Additionally, the growth of metascience research has been a positive outcome of the crisis, leading to greater scrutiny of psychology and science in general, and greater advocacy work regarding the use of more open and rigorous practices (Schooler 2014; Schooler 2019; Peterson & Panofsky, 2023).

In response to the issues identified as contributing to the crisis, there has been a movement towards improving research practices and adopting new standards in methodology and publishing. For example, there have been calls for more honest reporting of study procedures and analyses in scientific manuscripts, as well as greater documentation and accessibility of research materials such as data and code. Due to the high-profile replication attempts that initially drew attention to the crisis, the need for more replication attempts and particularly for independent replications, has also become more widely understood, in order to verify claims more thoroughly (Nosek et al., 2022). Encouraging or mandating open practices has also increased, through initiatives like open data policies, and incentives within journals such as badges for open practice (Wicherts et al., 2011; Steegan et al., 2016; Kidwell et al., 2016).

Awareness of study preregistration has also become more common in many disciplines, with greater uptake of this practice although this would still benefit from more widespread use (Nosek et al., 2018; Spitzer & Mueller, 2023). Preregistration will be discussed in more detail in subsequent chapters, particularly chapter 3. In short, however, this practice involves prespecifying research questions, hypotheses, study designs and analysis plans, prior to data collection (Wicherts et al., 2016), and ideally doing so publicly. This can therefore help to clarify which aspects of research have been planned *a priori* and which were introduced post hoc. However, preregistration alone seems to be insufficient to fully eliminate bias and QRPs, and greater accountability and incentives are needed (Mathieu et al., 2009; Ramagopalan et al., 2014). Furthermore, preregistration alone is insufficient to address the widespread issue of publication bias. Therefore, the Registered Reports publishing format aims to address many of these concerns in a more comprehensive way, as outlined in the following section.

1.4. Registered Reports

Registered Reports (RRs) are a major development in the effort to incentivise trustworthy and reliable research. RRs involve a two-stage publication process in which a stage 1 protocol consisting of the study's introduction, hypotheses, and methods section is peer reviewed and, if successful, gains in-principle acceptance (IPA) so that the subsequent stage 2 report will be published by the journal (or other reviewing platform) regardless of the findings, as long as the authors have adhered to their stage 1 protocol (Chambers, 2013; Chambers, 2019; Kiyonaga & Scimeca, 2019). Therefore, methods and analysis plans are essentially 'preregistered' with the journal, in a manner that holds the researchers accountable for adhering to their protocol. Furthermore, RRs aim to reduce publication bias because acceptance is not based on the existence of statistically significant results and instead is based

on the robustness of the methods and validity or importance of the question. This reduced pressure to produce 'interesting' and significant results should therefore remove the incentives that drive selective reporting and other QRPs (Chambers & Tzavella, 2022). Additionally, conducting peer review prior to data collection allows feedback to be incorporated into the study plan before data collection commences, thereby improving the overall quality of the study (Chambers, 2015; Chambers, 2019; Chambers & Tzavella, 2022).

Although this format was introduced in mainstream empirical science in 2012/2013 at the journals *Cortex*, *Perspectives on Psychological Science*, and *Social Psychology*, this type of approach is not a completely novel idea, having been previously suggested by various sources (Rosenthal, 1966 cited by Chambers & Tzavella, 2022; Walster & Cleary, 1970; Newcombe, 1987). These earlier proposals had not been carried forward and implemented, but a version of this format was introduced at the *European Journal of Parapsychology* in the 1970s and continued until the 1990s, although this appears to have been less detailed and rigorous than the current standards used for RRs (Wiseman et al. 2019; Johnson 1975, cited in Wiseman et al., 2019). Likewise, protocol review was used at the Lancet before being discontinued in 2015 (The Editors of the Lancet, 2015) but this did not offer in-principle acceptance regardless of the significance of the results (Chambers & Tzavella, 2022).

The modern introduction of the RR format in 2013 therefore filled an important gap in the publishing landscape and despite attracting some initial criticism, RRs have since been widely adopted (Chambers & Tzavella, 2022; DeHaven et al., 2019). They have now been offered by over 300 journals across multiple disciplines, either as a permanent format (n = 278) or as part of a special issue (n = 36). These represent disciplines far beyond psychology, including medicine, engineering, cancer research, ecology and environmental science, economics, law, politics, chemistry, and computer science (Center for Open Science, n.d.). Furthermore, hundreds of stage-2 RRs have been published to date, with many more stage 1 protocols having also been submitted.

Variant formats have also emerged, such as results-blind review where peer review is conducted after the study is completed but where the findings are known only to the authors, not to the editors or reviewers (e.g., Button et al., 2016). Though this approach is still vulnerable to QRPs by the authors, it may still help to protect against publication bias from the editors and reviewers. Furthermore, results-blind review has been used in conjunction with the Verification Reports format which focuses on re-analysing the data from previous

studies to verify their computational reproducibility[1] (Chambers, 2020). Another approach, post-publication peer reviewed RRs, has been adopted at some journals. This involves open public review of the stage 1 protocol rather than a more traditional closed peer review system (Chambers & Tzavella, 2022; Murray, 2017). Another major development in recent times has been the Peer Community In (PCI) Registered Reports initiative, launched in 2021. This non-profit platform provides independent peer review of stage 1 protocols and stage 2 preprints and if reviews are favourable, it provides a recommendation for the preprint to be published in one of their partnering journals without requiring any further peer review by that journal. This has the advantage of providing the author with a wider range of potential journals to choose from after just one submission process (Chambers & Tzavella, 2022; Peer Community In Registered Reports, n.d.).

Meanwhile, the original scope of the RR initiative has widened in recent years with the emergence of a small number of RR-funding partnerships, in which the stage 1 protocol is assessed by both a journal and funder simultaneously and so both funding and the in-principle acceptance to the journal are awarded concurrently (PLoS ONE Editors, 2017; Munafo, 2017). While the number of such initiatives that have been implemented is low, this type of joint review process appears promising and adaptable to the differing requirements of different funders and publishers. There are, however, important considerations acknowledged by various stakeholders, such as the need for good communication between the different stakeholders, and the potential workload impacts of the process (Drax et al., 2021; Clark et al., 2021).

1.5. Current Evidence regarding Registered Reports

While little empirical evidence existed regarding RRs when this PhD research began, a number of key studies have been published since then, shedding light on the format and the extent to which these appear to be working as intended. Allen and Mehler's (2019) study of RRs gave a promising first look at the proportions of hypotheses in RRs that were supported vs. unsupported. Since RRs aim to reduce the need for positive findings and provide space to share null results, it was anticipated that RRs should have a more realistic balance of supported and unsupported hypotheses than that seen in the standard research literature. Allen and Mehler's study provided the first evidence of this, showing rates of 60.5% of hypotheses

[1] Computational reproducibility refers to the "ability to recreate the same results as the original study (including tables, figures, and quantitative findings), using the same input data, computational methods, and conditions of analysis" (Framework for Open and Reproducible Research Training, 2021).

within the RRs being unsupported, compared with previous studies which report rates of only 5-20% in the standard research literature. Notwithstanding the fact that the authors did not select a control sample of standard reports, this study nevertheless provided promising initial evidence that RRs appear to be working as intended. These findings were then built on by Scheel et al. (2021) who also found that RRs had far fewer supported hypotheses than matched standard reports did (44% vs. 96%). As one of the proposed key benefits of the RR format is its potential to reduce false positive results, these findings confirming more realistic rates of hypothesis support in RRs were encouraging and have received considerable attention.

Other work has focused on the rates of open data and reproducibility of RRs (Obels et al., 2020), showing higher rates of data and/or code availability among RRs than among standard research articles within psychology, as well as RRs having greater computational reproducibility. Studies considering the format from a broader perspective have also been published by Soderberg et al. (2021) and Montoya et al. (2021). Soderberg et al. found that, using a peer-review-like process, reviewers perceived RRs to have much greater rigour and overall quality compared with standard articles. They also demonstrated that RRs had larger median sample sizes, greater sharing of materials and data, and more use of preregistration, than standard research articles, as well as performing better on a range of other criteria. However, the blinding of the reviewers regarding the type of article they were assessing may have been ineffective as most of the reviewers correctly identified whether the article that they reviewed was an RR or not. Therefore, this awareness may have influenced their responses and contributed to the considerable differences in how the different formats were rated.

Finally, Montoya et al.'s (2021) census of journal policies regarding RRs took an even broader view on the format, examining how RRs have been implemented across different journals. Their findings confirm that the majority of journals adopting RRs are in psychology, and that there is considerable variation in the amount of information available regarding the journals' RR policies. Furthermore, it appears to take an average of at least a year for journals to publish their first RR after they formally adopt the format, although this differs substantially between different journals. Their study also clarifies that exploratory analysis is not prohibited in RRs, although it may be slightly restricted e.g., needing to be presented in a separate section of the paper in order to be more clearly differentiated from the outcomes of confirmatory analysis. This is an important finding as it remains a common misconception

that exploratory analysis is not permitted in RRs. The authors also examined the extent to which power requirements were adopted by the journals, indicating that this was only the case in approximately 40% of the journals. They also found that most journals either required or encouraged open practices such as sharing of data and/or materials, while external preregistration of the stage 1 protocols (e.g., in a repository such as the Open Science Framework) was required in just over half of the included journals. This supports previous findings reported by Hardwicke and Ioannidis (2018) who found that many journals did not require accepted Stage 1 manuscripts to be made publicly available. The studies by Montoya et al. and Hardwicke and Ioannidis provide useful insights into how RRs have been adopted and implemented across journals and identify important areas for future improvements, but given their broader focus on journal policies, they lack a more granular view of the articles themselves and their specific characteristics.

1.6. Rationale for the Current Research

While a number of important metascientific studies of RRs have been published in the last couple of years, at the time the research for this thesis was started, there was very limited evidence of whether RRs actually appeared to be associated with their intended improvements and characteristics. Furthermore, most of the other research into RRs is hindered due to the difficulty of searching specifically for RRs in existing databases, particularly as many journals do not always correctly tag the publications as being an RR or occasionally incorrectly label standard research articles as being RRs, a problem that will be discussed in more detail in later chapters. The previous studies therefore mostly rely on the Centre for Open Science's Zotero library (Center for Open Science, n.d.) which contains a collection of stage 2 RRs. However, this resource is rarely updated, and a much more comprehensive database is needed. Therefore, this study aimed, initially, to assess the feasibility of creating a comprehensive database recording the characteristics of all published stage 2 RRs, and a yoked sample of regular research articles. It then aimed to use this to compare RRs and standard reports (SRs) on a wide range of different characteristics indicative of greater rigour and transparency, with the general expectation that RRs should be associated with these characteristics to a greater extent than SRs. The initial plan had been to investigate this in a much larger sample but there was not sufficient time available to match and code SRs for all 359 RRs that had been intended for inclusion. Therefore, the comparative sample that is used in most of these chapters consisted instead of 170 RRs and their 340 matched SRs. As the unmatched RRs had also been coded and checked, some

separate descriptive analyses were also run on this larger sample of RRs and these are reported in chapter 9. Despite this work relying on a smaller sample than had initially been intended, this still constitutes a larger sample of RRs than that used in the previous studies conducted on this topic.

The sample of SRs that constitute the control sample in the current study is more closely matched to the RRs than in the previous studies. Allen & Mehler's study does not provide its own control sample but instead compares their findings regarding the RRs to rates found in other studies of the standard research literature. Scheel et al. (2021) do provide a random sample of standard reports obtained by searching for articles that use the phrase "test* the hypothes*", as per the approach used by Fanelli (2010). Although this enabled direct comparisons of the RRs and SRs, the papers were not matched in terms of journal, topic or other characteristics. While their sample of comparison articles was limited to the same general timeframe as the RRs (published from 2013 to 2018), the RRs and SRs don't appear to have been individually matched on the specific timeframe of publication. Soderberg et al. (2021) did use a more detailed matching process but this was restricted to a small sample of only 29 RRs and 57 SRs. The current study therefore builds on previous efforts by providing a larger and closely matched control sample of SRs matched on journal of publication and timeframe, as well as matching as closely as possible on the topic, design, population, and if possible, sample size.

The studies that have emerged since this project began (e.g., Obels et al., 2020; Soderberg et al., 2021) have shed light on various characteristics of RRs and how these compare to the standard literature. However, many questions still remain to be answered about how RRs differ from standard research articles, and whether their theoretical benefits are realised. For example, there is currently insufficient evidence of whether RRs have more detailed methods sections, greater probability of including a sampling plan such as power calculations or Bayesian statistics, or greater use of manipulation checks, as compared with standard research articles. Further investigation into the rates of availability of data, code, and materials would also be beneficial as some of the current evidence is focused on the overall policy for RRs at the journal level rather than at article level. As many of these characteristics are indicators of greater methodological rigour, it is important to understand whether RRs actually differ on these characteristics compared with standard research articles. It could also be helpful to explore whether there are any differences in author demographics, or differences in the number of authors per paper, as more authors may indicate greater collaboration.

Author demographics for RRs have received little attention other than author seniority being briefly investigated by Chambers and Tzavella (2022), but more detailed insights into author characteristics could be beneficial.

1.7. Overview of Research Questions and Thesis Structure

Overall, this thesis seeks to determine whether RRs perform better than SRs on a wide range of criteria associated with greater quality, rigour and transparency. This chapter has sought to prove a brief introduction and overview regarding the RR format and the existing progress and evidence concerning its impacts and characteristics. The following chapter (Chapter 2: General Methods) will outline the overall process of creating the database and coding protocol. The subsequent chapters will then describe the specific research questions, hypotheses, coding processes and results for the different comparisons between RRs and SRs. Specifically, chapter 3 will outline the comparative analysis of the articles' hypotheses. While this chapter covers several different aspects of the hypotheses, it is particularly concerned with the comparative levels of supported and unsupported hypotheses, between the two article types. This chapter will show that RRs appear to be performing as expected, as demonstrated by much lower rates of supported hypotheses and higher rates of unsupported hypotheses, compared with the SRs. Chapter 4 will describe the comparative analysis of open practices, i.e., the sharing of data, analysis code, materials, and protocols, demonstrating that rates of these practices are higher among RRs than SRs. Chapter 5 will investigate a range of characteristics associated with the studies' methods and how these compare between the RRs and SRs. While the exact findings vary depending on the characteristics, this chapter also shows some encouraging evidence that RRs appear to be associated to a greater extent than SRs with practices indicating greater rigour and transparency. Chapter 6 will outline the comparative analysis of the author demographics, which reveal few differences between the article types and suggest that increasing the geographic diversity of RR authorship may be an important goal for future efforts. Chapter 7 will describe the comparative analysis of citation rates between the article types, as well as investigating whether there are associations between citation rates and levels of hypothesis support, or between journal impact factor and levels of hypothesis support. Chapter 8 will outline the comparison of signs of HARKing in RRs and SRs and will demonstrate that HARKing appears to be non-existent in RRs, while some evidence of this was found in SRs, suggesting that RRs may be performing as hoped in reducing this questionable research practice. Finally, chapter 9 will outline some descriptive analysis of a range of characteristics in a larger sample of RRs only, in order to yield a

broader overview of how common various practices are within RRs. The thesis will conclude with an overall discussion of the work in chapter 10, including how this relates to previous research, limitations of the work, and suggestions for future research and advocacy efforts. Overall, the evidence presented in this work demonstrates that, while there are still areas for improvement, RRs do appear to be working as intended in being associated with improved research practices, although further research is needed to determine causal impacts.

1.8. References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE, 12*(3), Article e0172792.

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

Anvari, F. & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology, 3(3),* 266-286.

Bruton, S. V., Medlin, M., Brown, M., & Sacco, D. F. (2020). Personal motivations and systemic incentives: scientists on questionable research practices. *Science and Engineering Ethics, 26*(3), 1531-1547.

Bottesini, J. G., Rhemtulla, M., & Vazire S. (2022). What do participants think of our research practices? An examination of behavioural psychology participants' preferences. *Royal Society Open Science 9*(4), Article 200048. https://doi.org/10.1098/rsos.200048

Button, K. S., Bal, L., Clark, A. & Shipley, T. (2016). Preventing the ends from justifying the means: withholding results to address publication bias in peer-review. *BMC Psychology, 4*(1), Article 59.

Center for Open Science (n.d.). *Registered Reports: Peer review before results are known to align scientific values and practices.* https://www.cos.io/initiatives/registered-reports

Chambers, C.D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex, 49*(3), 609-610.

Chambers, C. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 9-13.

Chambers, C. (2017). *The seven deadly sins of psychology*. Princeton University Press.

Chambers, C. (2019). The registered reports revolution: lessons in cultural reform. *Significance, 16*(4), 23-27.

Chambers, C. D. (2020). Verification Reports: A new article type at Cortex. *Cortex, 129*, A1-A3. doi: 10.1016/j.cortex.2020.04.020.

Chambers, C. D. & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*, 29-42.

Cheung, I., Campbell, L., LeBel, E., Ackerman, R. A., Aykutoglu, B., Bahník, Š., Bowen, J. D., Bredow, C. A., Bromberg, C., Caprariello, P. A., Carcedo, R. J., Carson, K. J., Cobb, R. J., Collins, N. L., Corretti, C. A., DiDonato, T. E., Ellithorpe, C., Fernández-Rouco, N., Fuglestad, P. T., . . . Yong, J. C. (2016). Registered Replication Report: Study 1 from Finkel, Rusbult, Kumashiro, & Hannon (2002). *Perspectives on Psychological Science, 11*, 750–764.

Clark, R., Drax, K., Chambers, C.D., Munafò, M., & Thompson, J. (2021). Evaluating Registered Reports Funding Partnerships: a feasibility study. *Wellcome Open Research, 6*, Article 231.

Costa, E., Inbar, Y., & Tannenbaum, D. (2022). Do Registered Reports make scientific findings more believable to the public? *Collabra: Psychology*, *8*(1), 32607. https://doi.org/10.1525/collabra.32607

Chopik, W. J., Bremner, R. H., Defever, A. M., & Keller, V. N. (2018). How (and whether) to teach undergraduates about the replication crisis in psychological science. *Teaching of Psychology*, *45*(2), 158–163.

DeHaven, A., Graf, C., Mellor, D., Morris, E., Moylan, E.C., Pedder, S., & Tan, S. (2019, September 17). *Registered reports: views from editors, reviewers and authors*. MetaArXiv. https://doi.org/10.31222/osf.io/ndvek

Donnellan, M. B., Lucas, R. E., & Cesario, J. (2015). On the association between loneliness and bathing habits: Nine replications of Bargh and Shalev (2012) Study 1. *Emotion*, *15*(1), 109–119.

Drax, K., Clark, R., Chambers, C. D., Munafò, M., & Thompson, J. (2021) A qualitative analysis of stakeholder experiences with Registered Reports Funding Partnerships. *Wellcome Open Research, 6*, Article 230.

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. V., 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.

Eerland, A., Sherrill, A. M., Magliano, J. P., Zwaan, R. A., Arnal, J. D., Aucoin, P., Berger, S. A., Birt, A. R., Capezza, N., Carlucci, M., Crocker, C., Ferretti, T. R., Kibbe, M. R., Knepp, M. M., Kurby, C. A., Melcher, J. M., Michael, S. W., Poirier, C., & Prenoveau, J. M. (2016). Registered Replication Report: Hart & Albarracín (2011). *Perspectives on Psychological Science, 11*(1), 158–171.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS One, 5*(4), e10068.

Ferguson, C.J., & Heene, M. (2012). The vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Fiedler, K. & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science, 7*(1), 45052.

Framework for Open and Reproducible Research Training (2021). *Computational reproducibility*. Framework for Open and Reproducible Research Training. https://forrt.org/glossary/computational-reproducibility/

Grand, J.A., Rogelberg, S.G., Banks, G.C., Landis, R.S., & Tonidandel, S. (2018). From outcome to process focus: fostering a more robust psychological science through registered reports and results-blind reviewing. *Perspectives on Psychological Science, 13*(4), 448-456.

Gopalakrishna, G., ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in The Netherlands. *PLoS ONE, 17*(2), Article e0263023.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. A. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science, 5*(1), Article 171511.

Hardwicke, T.E., & Ioannidis, J.P.A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2*, 793-796.

Hardwicke, T. E., Wallach, J. D., Kidwell, M. C., Bendixen, T., Cruwell, S., & Ioannidis, J. P. (2020). An empirical assessment of transparency and reproducibility-related research practices in the social sciences (2014-2017). *Royal Society Open Science, 7*(2), Article 190806. https://doi.org/10.1098/rsos.190806

Harris, C.R., Coburn, N., Rohrer, D., & Pashler, H. (2013). Two failures to replicate high-performance-goal priming effects. *PLoS ONE 8*(8), Article e72467.

Herndon, N. C. (2016). Research fraud and the publish or perish world of academia. *Journal of Marketing Channels, 23*(3), 91-96.

Hendriks, F., Kienhues, D., & Bromme, R. (2020). Replication crisis = trust crisis? The effect of successful vs failed replications on laypeople's trust in researchers and research. *Public Understanding of Science, 29*(3), 270-288.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Kerr, N.L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review, 2*(3), 196-217.

Kidwell, M.C., Lazarevic, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L.S., Kennett, C., Slowik, A., Sonnleitner, C., Hess-Holden, C., Errington, T. M., Fiedler, S., & Nosek, B.A. (2016). Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PLoS Biology, 14*(5), e1002456.

Kiyonaga, A., & Scimeca, J.M. (2019) Practical considerations for navigating registered reports. *Trends in Neuroscience, 42*(9), 568-572.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., 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

Korbmacher, M., Azevedo, F., Pennington, C., Hartmann, H., Pownall, M., Schmidt, K., Elsherif, M., Breznau, N., Robertson, O., Kalandadze, T., Yu, S., Baker, B., O'Mahony, A., Olsnes, J.,

Shaw, J., Gjoneska, B., Yamada, Y., Roer, J., Murphy, J., … Evans, T. (2023, May 29). *The replication crisis has led to positive structural, procedural, and community changes*. MetaArXiv. https://doi.org/10.31222/osf.io/r6cvx

Mathieu, S., Boutron, I., Moher, D., Altman, D.G., & Ravaud, P. (2009). Comparison of registered and published outcomes in randomized controlled trials. *JAMA, 302*(9), 977-984.

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), 487–498.

Methner, N., Dahme, B., & Menzel, C. (2022). The "replication crisis" and trust in psychological science: How reforms shape public trust in psychology. *Social Psychological Bulletin*. http://doi.org/10.23668/psycharchives.12192

Montoya, A. K., Krenzer, W. L. D., & Fossum, J. L. (2021). Opening the door to Registered Reports: Census of journals publishing Registered Reports (2013–2020). *Collabra: Psychology, 7*(1), 24404. https://doi.org/10.1525/collabra.24404

Munafò, M. R. (2017). Improving the efficiency of grant and journal peer review: Registered Reports funding. *Nicotine and Tobacco Research 19*, 773–773.

Murray, (2017, October 12th). Transparency meets transparency. *F1000 Blog Network*.

https://blog-f1000-com.abc.cardiff.ac.uk/2017/10/12/transparency-meets-transparency/

Newcombe, R. G. (1987). Towards a reduction in publication bias. *British Medical Journal*, 295, 656-659.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., Meller, D. T. (2018. The preregistration revolution. *PNAS, 115*(11), 2600-2606.

Nosek, B., Spies, J.R. & Matyl, M. (2012). Scientific utopia II: restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science, 7*(6), 615-631.

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Struhl, M. K., Nuijten, M. B., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., & Vazire, S. (2022). Replicability, Robustness, and Reproducibility in Psychological Science. *Annual Review of Psychology*, 73, 719-748.

Nuijten, M.B., Hartgerink, C.H.J., van Assen, M.A.L.M., Epskamp, S., & Wicherts, J.M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48*, 1205-1226.

O'Donnell, M., Nelson, L., Ackermann, E., & Aczel B. (2018). Registered replication report: Dijksterhuis & Van Knippenberg. *Perspectives on Psychological Science, 13*(2), 268-294.

Obels, P., Lakens, D., Coles, N., Gottfried, J. & Green, S. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 229-237.

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

Pashler, H. & Harris, C.R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science, 7*(6), 531-536.

Peer Community In Registered Reports (n.d.) https://rr.peercommunityin.org/about/about

Peterson, D. & Panofsky, A. (2023). Metascience as a scientific social movement. *Minerva, 61*, 147-174.

PLoS ONE Editors. (2017, September 26th). PLoS ONE partners with the Children's Tumor Foundation to trial Registered Reports. EveryONE: The PLoS ONE. https://blogs-plos-org.abc.cardiff.ac.uk/everyone/2017/09/26/registered-reports-with-ctf/

Pownall, M., Pennington, C. R., Norris, E., Juanchich, M., Smailes, D., Russell, S., Gooch, D., Evans, T. R., Persson, S., Mak, M. H. C., Tzavella, L, Monk, R., Gough, T.,Benwell, C. S. Y, Elsherif, M., Farran, E., Gallagher-Mitchell, T., Kendrick, L. T., … Clark, K. (2023, May 30). Evaluating the pedagogical effectiveness of study preregistration in the undergraduate dissertation. OSF. https://doi.org/10.17605/OSF.IO/5QSHG

Ramagopalan, S., Skingsley, A.P., Handunnetthi, L., Klingel, M., Magnus, D., Pakpoor, J., & Goldacre, B. (2014). Prevalence of primary outcome change in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Research, 3*, 77.

Rawat, S. & Meena, S. (2014). Publish or perish: Where are we heading? *Journal of Research in Medical Sciences, 19*(2), 87-89.

Sacco, D. F. & Brown, M. (2019). Assessing the efficacy of a training intervention to reduce acceptance of questionable research practices in psychology graduate students. *Journal of Empirical Research on Human Research Ethics, 14*(3), 209-218. doi:10.1177/1556264619840525

Scheel, A.M., Schijen, M.R.M.J., & Lakens, D. (2021). An excess of positive results: comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature, 515*, 9. https://doi.org/10.1038/515009a

Schooler, J. (2019). Metascience: The science of doing science. *Observer, 32*(9). https://www.psychologicalscience.org/observer/metascience-the-science-of-doing-science#:~:text=Metascience%2C%20also%20known%20as%20metaresearch%20or%20the%20science,of%20science%20and%20the%20study%20of%20scientific%20methods.

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.

Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014). P-curve: a key to the file drawer. *Journal of Educational Psychology: General, 143*(2), 534-547.

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J. G., Singleton Thorn, F., Vazire, S., Esterling, M. K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour, 5*, 990-997.

Sorokowski, P., Frackowiak, T., Kobylarek, A., Groyecka, D., & Blaszczynski, K. (2019). Registered reports as a method to increase the credibility of science: experimental study among psychological students. *Journal of Education, Culture and Society, 2,* 67-75.

Spitzer, L., & Mueller, S. (2023) Registered report: Survey on attitudes and experiences regarding preregistration in psychological research. *PLoS ONE 18*(3), Article e0281086.

Steegan, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712.

Świątkowski, W., & Dompnier, B. (2017). Replicability crisis in social psychology: Looking at the past to find new pathways for the future. *International Review of Social Psychology, 30*(1), 111-124.

The Editors of the Lancet (2015). Protocol review at The Lancet: 1997–2015. *The Lancet, 386*(10012), 2456-2457

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*(4), 411-417.

Vazire, S., Schiavone, S. R., & Bottesini, J. G. (2022). Credibility beyond replicability: Improving the four validities in psychological science. *Current Directions in Psychological Science, 31*(2), 162-168.

Versteege, N. (2013). The social work environment of researchers committing scientific misconduct. *Social Cosmos, 4*(1), 71-77.

Walster, G. W. & Cleary, T. A. (1970). A proposal for a new editorial policy in the social sciences. *The American Statistician, 24*(2), 16–19

Ware, J. J. & Munafo, M. R. (2015). Significance chasing in research practice: Causes, consequences and possible solutions. *Addiction, 110*(1), 4–8.

Wicherts, J.M., Bakker, M., & Molenarr, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of the reporting of statistical results. *PLoS ONE, 6*(11), e26828.

Wicherts, J.M., Veldkamp, C.L.S., Augusteijn, H.E.M., Bakker, M., van Aert, R.C.M. & van Assen, M.A.L. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Frontiers in Psychology, 7,* 1832.

Wingen, T., Berkessel, J. B., & English, B. (2020). No replication, no trust? How low

replicability influences trust in psychology. *Social Psychological and Personality Science, 11*(4), 454-463.

Wiseman, R., Watt, C., & Kornbrot, D. (2019). Registered reports: an early example and analysis. *PeerJ, 7,* e6232 https://doi.org/10.7717/peerj.6232

Wolff, W., Baumann, L., & Englert, C. (2018). Self-reports from behind the scenes: Questionable research practices and rates of replication in ego depletion research. *PLoS ONE, 13*(6): e0199554.

Youyou, W., Yang, Y., & Uzzi, B. (2023). A discipline-wide investigation of the replicability of Psychology papers over the past two decades. *Proceedings of the National Academy of Sciences, 120*(6) e2208863120

Chapter 2: General Methods

2.1. Overview of Chapter 2

This chapter will describe the overall approach used to develop the database structure and coding approach. It will first describe how the overall Registered Reports (RRs) sample for the database was gathered, followed by the selection of the Standard Reports (SRs) for the control sample. This comparative sample of RRs and SRs was used for the majority of the research in this thesis (chapters 3 to 7, and chapter 9). The subsequent sections of this chapter will then outline the process of developing the initial coding protocol, the processes of checking the coding, the creation of specific variables for the analysis, and a general overview of the analysis approach for chapters 3 to 6. Subjective ratings of the difficulty of the coding and any particular reasons contributing to this difficulty have been analysed and the results are outlined briefly. The relevance of the chosen SRs to their matched RRs was also coded during the initial coding process, and the results from the analysis of this data are presented.

Subsequently, the chapter will describe the coding processes used when gathering citation rates for the analysis outlined in chapter 7. Finally, an overview will be given of the sampling and coding approaches used in chapter 8 to investigate the comparative prevalence rates of HARKing within a small sample of RRs and SRs in cognitive psychology and neuroscience.

Although each of these aspects will be discussed in more specific detail within each subsequent chapter, this chapter aims to provide a more general overview of the processes and methods used before this is broken down into more detailed individual accounts.

2.2. Database Sample Creation

2.2.1. Identification of Registered Reports

The contents of the Centre for Open Science's (COS) Zotero library of Registered Reports were downloaded in October 2019 and these papers were checked to verify whether they were RRs. A small number of papers were rejected when they were found not to be RRs and this left a total of 177 RRs from this source. Subsequently, all journals included in the COS's Registered Reports Participating Journals list were searched to identify any other RRs available that had not been included in the Zotero library. Where necessary, several articles were verified via email, and articles within psychology were compared with an earlier record of those identified by Scheel et al. (2021) to confirm whether they were RRs.

Thereafter, email alerts were set up for Google Scholar, Elsevier, Science Direct, and Scopus, using the following terms: “registered report”, “registered reports”, “preregistered direct replication”, pre-registered direct replication”, “preregistered report”, “pre-registered report”, and “pre-results review”. These alerts were screened regularly to identify newly published RRs. An email alert was also set up for the journal Comprehensive Results in Social Psychology which only publishes RRs and so it was important to be aware of any papers published by this journal. Occasionally RRs were identified through personal recommendations and through Twitter. The sample to be included in the first version of the database was initially capped at 400 RRs which were gathered up to 14th October 2020 and represented 74 journals. No new RRs were added after this date.

However, a major issue when identifying and verifying RRs is that some papers do not use the term ‘Registered Report’, and that articles are sometimes falsely described as being an RR. Hardwicke and Ioannidis (2018) also report difficulties in clearly identifying RRs due to the lack of clear or consistent use of terminology for the format across journals. In line with this, a lack of clarity regarding the use of the term ‘Registered Report’ resulted in 40 coded articles being excluded from the final database sample in March 2022, resulting in a total RR sample of 359 papers. Although these excluded 40 articles from the Journal of Medical Internet Research (JMIR) all had International Registered Report ID numbers (IRRIDs), it was discovered that the journal used this ID number to refer to any paper they publish that has a published pre-registered protocol in a journal. It was initially very difficult to determine which of the articles in this journal with these ID numbers actually had a protocol published in JMIR Research Protocols (and so fit our definition of a RR because of the in-principle acceptance this granted), and those which had their protocol published elsewhere and did not in fact have in-principle acceptance at JMIR when they submitted their final report. Shortly before this issue came to light, JMIR changed their system for linking their stage 2 RRs with their stage 1 protocols in JMIR Research Protocols making it much easier to verify which articles were truly RRs. As the database sample was being finalised at the time, the decision was taken not to replace these articles and to instead cap the total RR sample at 360 papers. One more RR was later found to be a duplicate and was excluded but not replaced, resulting in a final total sample of 359 RRs. These 359 RRs represent a variety of different disciplines, including cancer biology, politics, and finance, though the majority of the RRs were from psychology and, to a lesser extent, neuroscience. However, the majority of the work in this thesis focuses on only 170 of these RRs, specifically those from psychology and closely

related areas (neuroscience, health, education). Although I had initially intended to conduct the comparative analysis on all the RRs that had been collected, the matching and coding processes were much more time-consuming that expected and so it was not feasible to adhere to this original plan. Instead, the 170 RRs in psychology and related areas that had been coded and checked and that had both of their SRs matched, coded and checked by March 2022, were chosen to use as the comparative sample, while the remaining RRs were only considered in the descriptive analysis reported in chapter 9.

As the RRs that were excluded in 2022 were not replaced, the end date for the RRs included in the database remained 14th October 2020.

Despite the challenges of accurately identifying the RRs, there is a clear difference between the number of RRs identified using this approach, and the 194 RRs currently in the COS's Zotero library which highlights the need for this new database.

2.2.2. Selection of Standard Reports

As mentioned in section 2.2.1. it had initially been intended that comparative standard reports (SRs) would be chosen for all 359 of the included RRs. However, due to time constraints, this was not feasible. Instead, 170 of the RRs which had SRs matched and coded and then had the coding checked, were included as the comparative sample. For each RR in the comparative sample, two SRs had been selected, giving a total sample of 170 RRs and 340 SRs for the comparative analyses (chapters 3 to 7). The rationale for including two SRs for each RR was in order to ensure there were adequate comparators for each RR, since the validity of the conclusions drawn from this study would be highly dependent on the control sample being well matched. In particular, the decision to choose two SRs per RR was anticipated to be important because this meant that two articles that were both relevant but in different ways could both be selected, ensuring a wider use of the matching criteria. For example, while one possible SR for an RR might match closely on the topic but not on the study design used, in other potential SRs the study design might be a much better match while the topic could be less relevant. Furthermore, the selection of two SRs per RR was in line with that approach taken by Scheel et al., (2021), which we were aware of when the current study was being designed and so this approach was also felt to be important in keeping the approach comparable to that taken by Scheel and colleagues.

SRs were chosen using a standardised protocol, which can be seen in Appendix 1. These were matched, wherever possible, on journal, timeframe, topic, design, population, and

sample size, in approximately that order. The first two characteristics were achieved by searching the same issue of the same journal that the RR was published in, or one issue either side of that, to identify likely matches. Within these limits, papers were matched as closely as possible on the topic and design, and if possible, on the population used. Sample size was also considered in this process but this was generally less of a priority than the other characteristics. Each of these six characteristics was then coded for their relevance to the RR it was matched to. Characteristics were each coded on a 4-point scale (Very irrelevant, Somewhat irrelevant, Somewhat relevant, and Very relevant). A cumulative score was then created from the sum of the ratings of each of these 6 characteristics to create an overall relevance score for the SR.

Although the standardised approach was possible to implement even across the wide range of different journals, there were some situations that required special consideration because the standardised approach was not completely feasible. For example, where RRs were published as part of a journal's special issue, SRs could not be identified from within the same issue and so were instead selected from the issues immediately before or after that special issue, in order to keep selection within the same general time frame of publication. This meant that the timeframe of these SRs could not be coded as 'Very relevant' because they had not come from the same issue of the journal as the RRs; the highest they could be rated for this characteristic was 'Somewhat relevant'. This was accepted as a necessary compromise in order to remain as close to the standardised protocol as possible.

A number of other challenges also arose in the selection of SRs. The use of a standardised protocol for identifying and selecting these articles ensured that they were matched to RRs in terms of the journal and relative time of publication. However, for many RRs there was simply not a clear match available within these issues in terms of the relevance of the topic or study designs. While this was a limitation, it seemed to be a necessary consequence of taking a consistent approach to this selection, and still represented a more stringent matching process than that used by previous studies. Furthermore, to confirm that the selection approach could be replicated appropriately, this matching protocol was used by four research assistants/interns to verify whether different researchers using this same approach could identify the same papers as suitable comparators. Checking reliability in the selection of SRs and in the level of agreement in the relevance coding, relied on a qualitative approach of checking each other's choices and discussing disagreements. While choices of SRs occasionally differed between researchers, these alternative selection choices were always

agreed to be reasonable choices, particularly within journal issues with many highly relevant papers available.

2.3. Development of Initial Coding Protocol for Database Sample

A structured database was created in Excel, initially with separate sheets for the RRs and the SRs. Development of the coding protocol was an iterative process. To determine the most feasible structure for the database and the range of response options necessary for each variable, a small selection of RRs was initially coded on a range of characteristics identified as being of interest. As the selected RRs were coded, a number of other characteristics of interest arose that were then added as variables to the database. For other variables, the difficulty of coding such diverse forms of each characteristic necessitated breaking some variables out into two or three separate variables, or into different levels; this maintained the stability of the database while also capturing these characteristics at the finest level of detail possible. For example, sample size was coded as the intended sample size, the pre-exclusion sample size, and the post-exclusion sample size, rather than just one single variable. In this way, characteristics were added or refined, and variables were recoded as necessary in order to determine the most suitable approach. Once the database structure and coding system was deemed relatively stable for RRs, a selection of SRs were coded to ensure that the structure was also stable for this article type.

Although the core coding approach has been stable, minor additions and clarifications have been necessary throughout the process in response to difficulties encountered, and alternative approaches that were deemed more informative or appropriate than those initially considered were adopted. Efforts were then made when double-checking older coding to ensure that the older coding complied with the more recent updates to the coding protocol. The final version of this initial coding protocol (version 6) is available in Appendix 2.

2.3.1. Database Structure

Initial coding and development of the protocol was an iterative process during which new variables were added, papers were re-coded repeatedly, and the structure of the database was altered to determine the most informative, stable, and user-friendly approach. This included splitting the structure into three separate levels of granularity: article level, study level and hypothesis level, so that details could be coded at as many of these levels as were applicable to the specific article. For example, while some papers may have only one study, with one hypothesis, others have multiple studies and may therefore state a general overall hypothesis

at an overarching article level. More specific hypotheses may be stated again or elaborated on within each study, and each of those studies may have multiple hypotheses or sub-hypotheses, necessitating the inclusion of the 'hypothesis level' within the database as well as a more general 'study level'. For example, Thai et al. (2019) conducted three studies in their SR. At an overarching (article-level) they predicted that "disparagement humor is better received if the source is from the group targeted by the humorous material, rather than an outgroup member", (Thai et al., 2019, p1). Within the first of their three studies, they sought to investigate this in relation to gay men: "It was hypothesized that, in the gay joke condition (but not the control joke condition), participants would perceive the joke and source more favorably if the source was gay than straight". This prediction was further broken down into four more specific predictions, each of which mapped onto specific outcome measures: "Specifically, the joke would be deemed more acceptable, less offensive, and more humorous, and the source's personality would be evaluated more favorably", (Thai et al., 2019, p2). Therefore, these four specific predictions were coded at hypothesis level, while the broader expectation for the study that encompassed these (i.e., that the joke and source would be received more favourably) was coded as a study level hypothesis. In the second of their studies, they focused on Asian people, again with a general prediction for the study and the four more specific predictions: "it was hypothesized that participants would evaluate the joke and source more favorably if the source was Asian, rather than Black or White. Specifically, the joke would be deemed more acceptable, less offensive and more humorous, and the source's personality would be rated more favorably", (Thai et al., 2019, p5). Likewise, their third study took a similar approach in relation to perceptions of the acceptability of jokes about both gay and Asian people. As a result, this article as a whole was coded as having one article-level hypothesis, three study-level hypotheses (one per study), and twelve hypothesis-level hypotheses (four per study). The coding of hypotheses at each of these levels was in contrast to the approach taken by Scheel et al. (2021) who coded only the first tested hypothesis for each paper regardless of what level they were stated at.

Many of the included characteristics were coded at each of these three levels as applicable. This allowed a clear and comprehensive approach to coding the article characteristics and it has proved to be stable and applicable to the diverse range of different RRs and SRs included. Further details are given about the specific coding processes used for each variable in the chapters that follow, but a brief overview of these characteristics is given in the following sections.

2.3.2. Characteristics Coded: Initial Coding Process

The characteristics coded during the initial coding process include: country; availability of protocol; number of authors; level of author seniority; statement of hypothesis at article, study and hypothesis level; articulation of hypothesis at each level; support for hypothesis at article, study and hypothesis level; articulation of support for hypothesis at each level; sample size (at planned, pre-exclusion, and post-exclusion, levels); sampling plan (e.g. use of frequentist power analysis, Bayesian analysis or other); inclusion of exploratory analysis; the nature of the exploratory research; clarity of any distinction made between confirmatory and exploratory research; inclusion of manipulation checks; whether manipulation checks have been passed; whether the study is a replication (i.e. original study or direct/indirect/internal direct/internal indirect replications); word count of methods section; number of studies reported in each article; and availability of data, code and materials.

2.3.3. Testing Clarity and Feasibility of the Protocols

The second draft of the coding protocol and first draft of the selection protocol were shared with an undergraduate intern in June 2020, who found the protocol clear and reasonably feasible to use. Initially the intern attempted to code papers using this protocol without additional assistance, and generally made good progress. Feedback was then provided on her coding and any deviations from the approach specified in the protocol were identified in order for these to be re-coded appropriately. Aspects of the protocol that could be clarified or additional details that would be beneficial were added based on her advice. Once the intern had become familiar enough with the coding process, a selection of articles was independently double coded and compared to identify discrepancies. These discrepancies were discussed at length until agreement was reached. In most cases, this was resolved when the intern agreed with my own approach and was due to my own greater level of familiarity with the coding protocol, as well as having more experience or knowledge of some of the research methods and topics used in the papers being coded.

Further minor clarifications and additions were made to the coding protocol when subsequent research assistants joined the team and learned to apply the protocols. While these changes were very minor and primarily involved additional examples and clarification rather than actual changes in the coding approach itself, these changes improved the quality of the coding protocol and have resulted in the final (6^{th} version) of the protocol. These additions were discussed among the coding team before being formally adopted and added into the coding protocol. During the double-checking process, efforts were made to check that the

coding of any refined characteristics was consistent with the most up-to-date approach, including updating this throughout the earlier coding.

The SR selection protocol was also tested by various research assistants in the same way but subjected to only minimal changes. The minor changes led to the creation of the second version of the SR selection protocol, which is the current version.

2.4. Double Coding & Double-Checking

In order to verify the accuracy and consistency of the coding, it was thoroughly checked either by the same myself at a later date, or by another team member. A small number of papers (approx. 15 RRs and 16 SRs) were independently double coded, as previously mentioned above. Of these, data regarding the number of discrepancies found between coders was available for 14 of these RRs and 13 of the SRs. The other few double-coded articles mentioned had been used as a training exercise when new coders joined the team at later dates; data on the discrepancies for those other papers was not formally recorded and disagreements were handled through discussion only. Based on the sample for which the data regarding the discrepancies was recorded (n = 14 RRs and 13 SRs), in total, SRs had more coding discrepancies than RRs did (i.e. 93 in total or a mean of 7.15, compared with a total of 71 and a mean of 5.07 discrepancies in RRs). Discrepancies in the coding were broadly categorised as either major or minor discrepancies. Major discrepancies tended to relate to the coding of hypothesis characteristics, or the level at which the information was being coded, as these could affect the structure of the rest of the coding for that article. Our coding of the SRs had a mean of 5.69 minor discrepancies compared with a mean of 3.79 in the RRs. The mean for the major discrepancies was only slightly higher for the SRs than for the RRs (1.46 vs. 1.29). Consensus on the correct approach to the coding of these discrepancies was then reached through discussion between the two coders. However, due to the time-consuming nature of the double-coding process, it quickly became clear that this approach was not feasible for the full database. Instead, coding was double-checked rather than double-coded, and was usually double-checked by myself rather than by a different coder, due to the time and resource constraints of the research assistants involved in the project.

To ensure that the earlier coding was updated and consistent with the later changes to the protocol, the earlier papers were double-checked, and any necessary changes were made to bring the approach in line with the newer updates. During this process, all other aspects of the coding of those papers were also double-checked. Having gained considerably more

experience in coding during the year between first coding these papers and double-checking them, some earlier uncertainties became easier to resolve, and in a small number of cases it was easier to identify partially stated hypotheses that were not as clear initially. Although the coding did not need to be changed in most instances, this checking of the earlier coding was necessary to ensure that the changes made to the protocol later in the process were also implemented for the earlier coding, to maintain consistency throughout the database.

Due to the difficulty of coding papers outside of our own areas of expertise, we had initially intended to bring in external experts to verify the accuracy of the coding of certain papers. This was considered particularly important for the coding of the eLife papers due to our lack of expertise in the areas of cancer biology and genetics, and also for some papers from the area of finance and accounting. However, due to time constraints this external review process was not feasible. Fortunately, as the comparative analysis has been restricted to articles from psychology and related disciplines, these more challenging papers were only included in the RR-only analysis reported in chapter 9.

Based on the insights gained while checking the earlier coding, a checklist template had been developed in Excel to guide external checkers in what to focus on specifically, and to provide a structured way to document whether they agree with the coding decisions made and any suggestions for alternative coding decisions. This had a specific section for the types of issues commonly experienced when coding the eLife papers in particular, such as lack of clarity about the sample sizes, and about the use of manipulation checks. The first draft of the checklist was shared with a research assistant for feedback, and no changes were considered necessary. It was intended that this checklist could be used to guide external reviewers when checking the accuracy of the coding in disciplines that the team were less familiar with. Additionally, as this checklist was developed before the decision was taken to check my own coding rather than relying on research assistants to do so, it was also thought that the checklist could be used whenever any team member checked another member's coding. However, as this approach was then not used, the checklist has not been used in practice, although it may still prove useful for any future coding of additional papers or other future work in this area.

2.5. Variable Creation for Database Chapters (Chapters 3 to 6, & 9)

Based on the data gathered from the initial coding process and the subsequent checking of this coding, a range of new variables were created for the analysis. In some cases, these

variables were also created at the sub-levels of the coding structure, but the majority were created as overall level variables in order to summarise these characteristics across the paper, particularly for some characteristics where the distinctions between the levels were less clear or less meaningful. In some cases, such as the hypothesis variables, both overall and sub-level variables were created. The specific process of how each individual variable was created is outlined in the relevant chapters that follow.

2.6. Analysis Approach for Database Chapters (Chapters 3 to 6, & 9)

The specific analyses are outlined in the following chapters. In general, however, descriptive statistics were examined. Where categorical variables were used, this relied on the mode and frequencies. Where proportion scores or other numerical variables were used the medians were typically of interest, particularly as the data were generally not normally distributed. Within the comparative sample, categorical variables were compared between article types using chi-square analysis, and proportion scores were compared using the Mann-Whitney U test. These analyses were conducted in JASP.

As described in chapter 9, descriptive statistics were also examined for the various characteristics within the full RR-only sample (n=359). As this sample of RRs contained those papers from the more challenging disciplines such as cancer biology and finance, this offered the opportunity to examine the characteristics within a larger and more diverse sample of RRs rather than only the smaller sub-sample of psychology and related disciplines used for the comparative analysis.

2.7. Examination of Coding Difficulty for Database Chapters (Chapters 3 to 6, & 9)

The overall perceived level of coding difficulty and reasons for this difficulty were coded during the initial coding process. These were analysed in order to understand any particular issues in the coding process and whether there were any differences in coding difficulty between the two article types. This coding difficulty was examined descriptively in the overall sample of all 359 of the included RRs, and in the comparative sample consisting of the 170 RRs and their 340 SRs.

2.7.1. Coding and Analysis of Coding Difficulty

During the initial coding process, the level of difficulty of the coding for that paper was considered and this was coded on a four-point scale from Very easy to Very difficult (Very easy, Somewhat easy, Somewhat difficult, and Very difficult). An overall judgement was made on difficulty and, in an additional column, the coder noted what aspects were

considered particularly difficult, if applicable. Although the reasons for any difficulty were initially provided as a free-text response, when the coding protocol was refined, standardised response options were created and these were used wherever possible. These include the following: Inferring hypotheses; Inferring support for hypotheses; Determining level of coding; Determining level of coding – hypotheses; Determining level of coding – methodological details; Identifying manipulation checks; Inferring support for manipulation checks; Lack of clarity regarding sample sizes; Determining number of studies in paper. However, this list of response options is not exhaustive and if another aspect of the paper caused particular difficulty then a free text response was used to document this in the coder's own words.

It is worth noting that the listing of these reasons for coding difficulty was restricted to characteristics that caused significant difficulties, and not necessarily just instances where information could not be inferred. For example, in many cases, certain characteristics would not be clearly stated in the paper and so could be easily coded as Unclear during the initial coding process (or Other, No, N/A, etc.) but these did not necessarily cause actual difficulty in the coding process. Instead, this was kept for characteristics that had caused significant challenges. For example, if the coder couldn't see any evidence of a sampling plan being used this could be coded as Unclear quite easily, and this characteristic did not need to be mentioned as a reason in this column, whereas if there was some evidence to suggest a different response was more appropriate for the coding of this characteristic but understanding the authors' explanations was very difficult and time consuming, so that it caused significant delays, difficulty or uncertainty of the eventual coding judgement, this could then be included as a reason for the coding difficulties.

Subsequently, an overall score was created to reflect how difficult the coding of the article had been overall. This was created on a scale of 1 representing 'Very easy', to 4 representing 'Very difficult'. Categorical variables were created for each of the characteristics that seemed to most commonly cause difficulty during the coding process. These included inferring the statement of the hypothesis, inferring the support for the hypothesis, determining the level at which to code information, inferring sample details (e.g. sample sizes or sampling plans), inferring the existence of manipulation checks, inferring whether manipulation checks were passed, inferring whether there were replication studies included in the paper, inferring whether exploratory analysis was included in the paper, difficulty due to a lack of familiarity with the topic or approach used in the paper, inferring the number of studies in the paper, and

finally, whether there was any other cause of coding difficulty. These categorical variables were each coded as either 1 or 0 to indicate whether there was or wasn't difficulty with each characteristic, respectively. Descriptive statistics were examined for each article type, particularly frequencies. Chi square analysis was also undertaken to examine whether there were any statistically significant differences between the two article types in the extent to which each characteristic caused difficulty during the initial coding process.

2.7.2. Results for Differences in Coding Difficulty

2.7.2.1. Comparative Results for Differences in Overall Coding Difficulty

As outlined above, an overall score from 1 to 4 was created as an overall measure of the coding difficulty for each paper, where 1 was Very easy and 4 was Very difficult. Descriptive statistics show that most articles had been coded as 2 (Somewhat easy) for both article types. Examination of the frequencies for each article type (as shown in Figure 1 below) also showed that there was no substantial difference in the difficulty of the coding between the two article types.

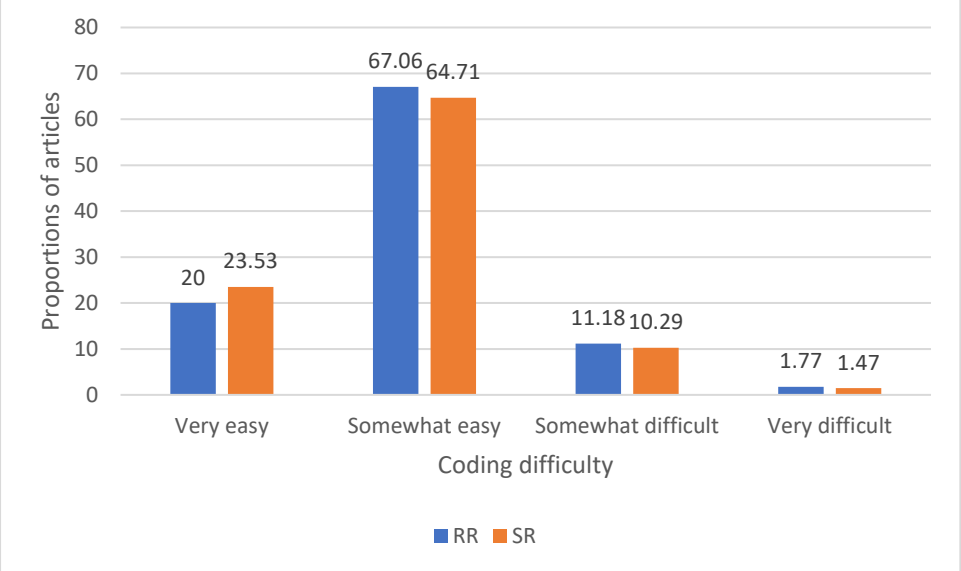

Figure 1: Bar graph depicting the overall coding difficulty scores in Registered Reports (RRs) and Standard Reports (SRs). The relative distributions were similar in both article types, and the majority of the SRs and RRs were rated as being 'Somewhat easy' to code.

As expected from these descriptive statistics, chi-square analysis showed that there was no significant difference between the two article types in the overall coding difficulty score: $x^2(3, N = 510) = 0.87$, $p = 0.83$.

2.7.2.2. Results for Coding Difficulty in Full RR-only Sample

Within the full RR-only sample (*n* = 359), descriptive statistics were also examined. This also showed that the mode for coding difficulty in the full RR-only sample was 2 (Somewhat easy). Frequencies also showed that most (61%) of the RRs had been considered 'Somewhat easy' to code (i.e., achieving a score of 2 on the scale), while 21.45% were considered 'Somewhat difficult' (achieving a score of 3). A further 15.04% were considered 'Very easy' to code (achieving a score of 1), while only 2.51% of the RRs were considered to be 'Very difficult' to code (i.e., achieving a score of 4).

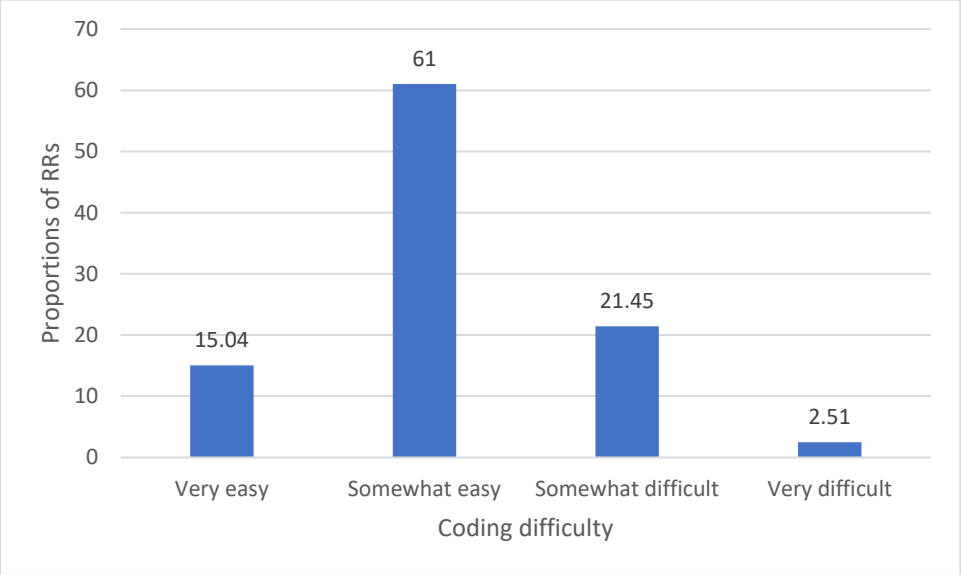

Figure 2: Bar graph showing the overall coding difficulty scores in the full Registered Report (RR) sample from chapter 9 (n=359). The majority of these RRs were rated as being 'Somewhat easy' to code.

2.7.3. Results for Comparative Analysis of Reasons for Coding Difficulty

Descriptive statistics and chi-square analysis was used to compare the reasons for coding difficulty between the two article types. In general, there were very few significant differences between the article types in the reasons for coding difficulties.

Within the full comparative sample, only two characteristics were found to have a statistically significant difference between the article types: determining the number of studies in the paper, and the use of manipulation checks and whether these were passed. Although for most

papers there was not much difficulty in determining the number of studies in the paper, where this issue did exist it was found to be more common in the RRs (4.12%) than in the SRs (0.88%). Chi-square analysis showed that there was a statistically significant difference between the article types in the extent of this issue: $x^2(1, N = 510) = 6.17, p = 0.01$.

There was significantly greater difficulty in determining whether manipulation checks were passed in the RRs than in the SRs: $x^2(1, N = 510) = 12.73, p < 0.001$. A significant difference was also found at first for difficulty in inferring whether manipulation checks were actually included, with this issue more likely in RRs (5.29%) than in SRs (0.88%). Chi-square analysis showed a significant difference between the article types in this type of coding difficulty: $x^2(1, N = 510) = 9.60, p = 0.002$. However, this could be because RRs were much more likely to include manipulation checks and so it is reasonable to think that this could lead to more instances of difficulty in coding these manipulation checks for RRs than for the fewer instances in SRs. The analysis was therefore run again after filtering out the papers that did not include manipulation checks. This showed that the difference was no longer statistically significant: $x^2(1, N = 119) = 1.37, p = 0.24$.

Where the comparisons of the other characteristics were also stronger or more informative when papers that did not include these characteristics were removed, the same filtering was undertaken and the analyses run on the restricted sample. Where this has occurred, it was referred to as the 'restricted comparative sample', to distinguish it from the 'full comparative sample'. In some instances, however, both the full and restricted samples were of interest. For example, filtering out the papers that had been coded as 'exploratory only' allowed analysis of the coding difficulty to be conducted on only the papers that included hypotheses which is helpful in understanding the difficulty of coding the hypotheses. However, the efforts to determine whether there were any hypotheses could also have caused considerable difficulty in coding the paper even if it was ultimately decided that there were no clear hypotheses. Therefore, where both the full and restricted samples had the potential to be informative, both were considered. A similar approach was taken to other characteristics if both a full and restricted sample was of interest but this did not change the outcomes of the comparations between the article types.

2.8. Examination of SR Relevance

2.8.1. Coding and Analysis of SR Relevance

The relevance of each SR to its matched RR was coded during the initial coding process, across six different characteristics, and an overall relevance score was derived. These six characteristics were the relevance of the journal and the timeframe, and the relative similarity of the research design, topic, population and/or sample size. These were each coded as either Very relevant, Somewhat relevant, Somewhat irrelevant, or Very irrelevant. An overall rating was derived from the sum of these six characteristics by taking Very irrelevant to be 0 and Very relevant to be 3. When these were summed for the six characteristics, it gave a total number out of a maximum possible total of 18. To determine an overall relevance ranking for the SR, SRs with a total score of 0 to 4 were considered to be Very irrelevant; SRs with a total score between 5 and 9 were considered to be Somewhat irrelevant; SRs with a total score between 10 and 14 were considered to be Somewhat relevant; and SRs with a total score between 15 and 18 were considered to be Very relevant.

From the data coded in the initial coding process, a single scale variable was created for the overall relevance of the SRs to their matched RRs. Using the original scale used in the initial coding process (Very irrelevant, Somewhat irrelevant, Somewhat relevant, and Very relevant), the responses were re-coded on a scale of 0 to 3, with 0 representing 'Very Irrelevant' and 3 indicating 'Very relevant'. In hindsight, however, using the sum score that was used to determine the overall relevance score from the sum of the six characteristics, would have been more informative to use as a variable for this as it could have shown more variation in the relevance scores than this 4-point scale.

As the SR relevance was only coded for the SRs, the sample only consisted of the 340 SRs within the comparative sample. Within this sample, descriptive statistics were examined for the overall SR relevance scale variable. Particular attention was paid to the frequencies in order to understand how frequent each level of the scale was.

2.8.2. Results for SR Relevance

Analysis of the numerical scale variable created for this characteristic showed that none of the SRs in this comparative sample had been coded as being Very irrelevant (0), and that most were considered to be Somewhat relevant (i.e., a modal score of 2). Examination of the frequencies also confirmed this, showing that 56.47% of the SRs had been considered 'Somewhat relevant' to their matched RRs (i.e., achieving a score of 2 on the scale), while a

further 36.47% were considered ‘Very relevant’ (achieving a score of 3 on the scale). Only the remaining 7.06% of SRs were considered to be ‘Somewhat irrelevant (achieving only a score of 1 on the scale).

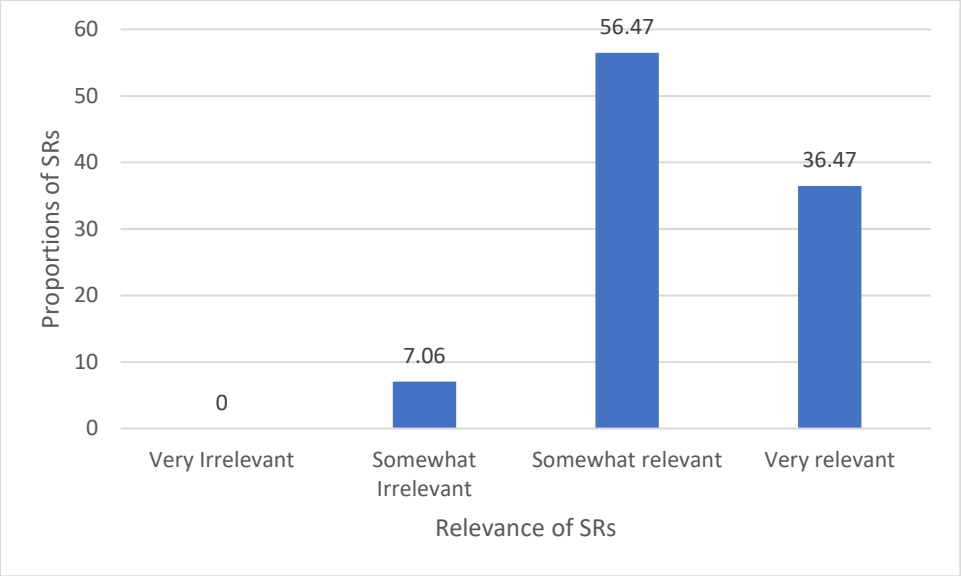

Figure 3: Bar chart showing the overall relevance scores for Standard Reports (SRs). The majority of the SRs were considered to be ‘Somewhat relevant’, with most of the remaining SRs being rated as ‘Very relevant’. Only a small proportion of the SRs were rated as only ‘Somewhat irrelevant’, while none were considered ‘Very Irrelevant’.

2.9. Interim Conclusion: Conclusion Regarding Main Database Creation

In conclusion, detailed coding and matching protocols were developed and used, and efforts were made to ensure that the coding was checked again before analysis. There were very few differences in the difficulty of coding the RRs and the SRs. Where differences did exist, these were in relation to the difficulty of coding the manipulation checks and support for the manipulation checks, as well as the difficulty of coding the number of studies in the paper, with greater difficulty in each of these characteristics in the RR sample than in the SR sample. While the aforementioned methods cover the majority of the research reported in this thesis, two additional projects were conducted with MSc students. In the following sections of this chapter, a very brief overview is given of the methods of these two additional studies, although these are reported in more detail in chapters 7 and 8.

2.10 Overview of Additional Projects (Chapters 7 and 8)

In addition to the overall database which forms the basis of most of this thesis, two projects were subsequently conducted based on work conducted with MSc students for their thesis projects. The project outlined in chapter 7 examines the rates of citations for each article type, and whether these citation rates are associated with the proportions of supported and unsupported hypotheses. Chapter 8 investigates whether there are any signs of HARKing evident in a very small sample of RRs and SRs. An overview of each of these projects is given below, and they are described in more detail in chapters 7 and 8 respectively.

2.10.1. Comparative Citation Rates Study Methods

As outlined in more detail in chapter 7, the full comparative sample was used to gather citation rates for each of the articles. The citation rates were gathered from three sources: Google Scholar, Scopus, and Web of Science, and the mean scores across these three sources were calculated in order to get a single score for each article. This had initially been conducted by MSc students on a much smaller sample from the database, and their approach was later replicated to gather more recent citation data for the full sample. A Mann Whitney U test was used to compare the citation rates between the two article types. Correlations were examined between the citation rates and the proportion of supported hypotheses, within each article type. Journal impact factors were also gathered from the most recent Clarivate report and when journals were not included in this source, these were sought online for the same year. Even so, a small number of journals did not have impact factors available and so these were excluded from the analysis when correlations were examined between the journal impact factor and the rate of supported hypotheses.

2.10.2. Overview of Comparative HARKing Rates Study Methods

2.10.2.1. Sample Creation

In order to investigate any signs of HARKing in the RRs and SRs, a much smaller sample of RRs (n = 12) was chosen from the main database, all in the areas of cognitive psychology and neuroscience. All of these articles needed to have a stage 1 protocol available in order to investigate signs of HARKing and so the availability of this influenced the choice of articles. As the existing SRs matched to these 12 RRs did not have preregistered protocols available, a different control sample was created. This new SR sample consisted of 12 SRs selected from the Open Science Framework (OSF) registry which were matched to the RRs on keywords or topic area, had a pre-registered protocol available, and whose results had already been published in a peer-reviewed journal.

2.10.2.2. Coding of HARKing

The process of coding the articles to investigate HARKing is described in more specific detail in chapter 8, but this involved extracting a range of details from the protocols and their corresponding final reports. These details included the exact text of the hypotheses stated in each document. Based on the information gathered during this initial coding process, a series of characteristics were then coded to represent whether there were any differences between the protocols and the final manuscripts. These are described in more detail in chapter 8. In short, however, responses were coded for whether there was a change of any kind in the wording of the hypothesis between the final manuscript and the protocol, whether any hypotheses were added to the final manuscript compared to the protocol, and whether any hypotheses were removed from the final manuscript compared to the protocol. Where changes had occurred in the wording of the hypothesis between the protocol and the final manuscript, the extent of such changes was documented, i.e., whether they were minor (consisting of wording changes only) or major (consisting of both wording changes and also changes to the meaning). Finally, the coding indicated whether there was considered to be evidence of HARKing, which was defined as at least one major change to at least one existing hypothesis, or the addition or removal of at least one hypothesis. Finally, a subjective rating of the coder's certainty in the accuracy of the coding was recorded. This new coding was conducted initially by a trained MSc student, before being checked in detail to ensure accuracy and consistency in the coding approach.

2.10.2.3. Analysis of HARKing Rates

Descriptive analysis was sufficient to detect the presence of any changes between the protocols and manuscripts. A number of categorical variables and proportion scores were created, as outlined in chapter 8, and a Wald test was used to test for any association between the article type and the rate of HARKing, to determine whether this rate was, as expected, higher in SRs than RRs. An independent samples *t*-test was used to compare the proportion of major changes (between the final manuscripts and the protocols) between RRs and SRs, i.e., the proportions of HARKing relative to the total number of changes. Another independent samples *t*-test was used to test for whether there was a difference between the RRs and the SRs in the proportion of minor changes between the protocols and manuscripts.

2.11. Conclusion

The aforementioned details provide an overview of the methods used to create the study's samples and the general development of the coding protocol. Further details about each of the

characteristics examined are outlined in more depth in each of the following chapters. Specifically, these sections examine the comparisons of the hypotheses (chapter 3); of the availability of data, code and materials (chapter 4); of the methodological variables or study characteristics (chapter 5); of the author demographics (chapter 6); of the citation rates (chapter 7); and of the rates of HARKing (chapter 8). Chapter 9 will then present an overview of descriptive statistics within the larger total sample of RRs only. Finally, chapter 10 will present an overall discussion and conclusion to the thesis.

2.12. References

Center for Open Science (n.d.). https://www.zotero.org/groups/479248/osf/collections/KEJP68G9 Accessed October 2019

Hardwicke, T. E. & Ioannidis, J.P.A. (2018). Mapping the universe of Registered Reports. *Nature Human Behaviour, 2,* 793-796.

Scheel, A.M., Schijen, M.R.M.J., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with Registered Reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Chapter 3: Comparative Analysis of Hypotheses in Registered Reports and Standard Reports

3.1. Introduction

This chapter will describe the phenomenon of confirmation bias and how this contributes to the over-representation of positive findings within scientific research. Issues such as p-hacking and publication bias will also be considered. The potential for RRs to mitigate such issues and promote the publication of null results will be briefly outlined. Subsequently, a comparative study of the hypotheses within RRs and SRs will be described and results will be discussed in light of the relevant literature.

3.1.1. Confirmation Bias

Confirmation bias is thought to be the most common cognitive bias and refers to a complex and heterogeneous phenomenon whereby people search for, interpret and remember information that supports their pre-existing beliefs (Nickerson, 1998). Therefore, they tend to give more weight to information that supports these beliefs while under-valuing information that contradicts them, requiring less evidence to accept a hypothesis than to reject it (Pyszczynski & Greenberg 1987; Beattie & Baron, 1988; Ditto & Lopez, 1992; Lord et al., 1979). This is generally considered to be an automatic and unintentional phenomenon rather than a deliberate pattern of thinking.

The ways in which confirmation bias can affect reasoning are numerous. Biased search for information can mean that people tend to search for information in a one-sided way which makes them more likely to find information that confirms their hypothesis or pre-existing beliefs, while ignoring or avoiding information that challenges this (Nickerson, 1998). Biased interpretation of that evidence also leads people to favour interpretations that support their preferred conclusion, while discounting or being critical of information that disconfirms this, i.e., 'disconfirmation bias' (Taber & Lodge, 2006; Ditto & Lopez, 1992; Talluri et al., 2018). Therefore, even when they are provided with the same evidence, people's interpretations could still be biased because of this phenomenon (Lord et al., 1979), particularly if the information provided is ambiguous, allowing more room for differing interpretations. Finally, biased recall of information can also affect people's thinking as selective memory prioritises confirming evidence and remembers this better than disconfirming evidence (Gilovich, 1983).

Confirmation bias impacts many key areas of daily life and society, particularly because it contributes to people having overconfidence in their own beliefs (Nickerson, 1998). This kind of bias may explain a number of other phenomena such as attitude polarization, whereby disagreements in attitudes or beliefs become more extreme even though both sides of the argument have access to the same evidence (Taber & Lodge, 2006; Lord et al., 1979); the idea of belief perseverance, whereby people are thought to continue to hold their beliefs even after they have been given evidence to prove that those beliefs are false (Ross et al., 1975; Anderson et al., 1980; Davies, 1997; Anderson 1983); the irrational primacy effect i.e. relying to a greater extent on information presented earlier while then undervaluing or ignoring information presented later on; and illusory correlation in which people incorrectly believe there to be a correlation between two different events or concepts (Chapman & Chapman, 1969). The strength of an individual's confirmation bias can be considerable, with its effect believed to be strongest for emotional issues, deeply entrenched beliefs, or other very strongly desired outcomes (Taber & Lodge, 2006). The potential strength of confirmation bias can be particularly seen in the related "backfire effect", whereby encountering evidence that challenges their beliefs can actually strengthen the person's belief or commitment to their own view (Silverman, 2011).

The practical impacts of confirmation bias can be seen in people's approaches to believing and seeking information presented on social media and news sites (Ling, 2020; Modgil et al., 2021; Workman, 2018), seeking and sharing of health information (Zhao et al., 2020; Meppelink et al., 2019), recruitment of job candidates (Agarwal, 2018; Consul et al., 2021), political beliefs (Taber & Lodge, 2006), confidence in political candidates (Westen et al., 2006; van Erkel & Thijssen, 2016), juries' decision-making (Pennington & Hastie, 1993; Hudachek & Quigley-McBride, 2022), forensics (Kassin et al., 2013; Cooper & Meterko 2019), community policing (Schlosser et al., 2021), medical diagnosis (Howard, 2018; Pines, 2006; Mendel et al., 2011; Prakash et al., 2017; Tschan et al., 2009), and stereotyping judgements or perceptions of other people (Darley & Gross, 1983). Furthermore, this form of bias clearly has the potential to impact on reasoning within scientific research, with strong evidence of confirmation bias and failure to test alternative hypotheses noted by Mynatt et al. (1977) in an experimental study that used a simulated research environment. However, when participants in their study did obtain explicit falsifying information, they did use this information to reject incorrect hypotheses.

Others have suggested that confirmation bias occurs as part of a broader 'positive test strategy', in which people tend to test cases that they expect or know have the property of interest, rather than cases that they think lack that property (Klayman & Ha, 1987). Within scientific research, using a positive test strategy has been criticised as being theoretically improper as it is considered incompatible with falsification. However, Klayman and Ha (1987) have suggested that this is not necessarily the case and that it can be a useful heuristic when testing a hypothesis, particularly when cognitive demands are high or when there is a lack of definite, task-specific information. However, they do concede that this approach can also increase the likelihood of errors or inefficiencies in the testing process.

Confirmation bias can also be seen in the evaluation of scientific outputs. For example, Mahoney (1977) showed that reviewers for journals were strongly biased against manuscripts containing results that did not support their own theoretical perspective. Similarly, Koehler (1993) found that scientists rated reports more highly when they were consistent their own beliefs than when they were inconsistent with them. Furthermore, this effect was mediated by the strength of the scientists' prior beliefs, with stronger prior beliefs associated with greater agreement. More recent evidence shows similar findings, with researchers rating abstracts more highly when they confirm the researchers' own beliefs about the topic (Hergovich et al., 2010).

Confirmation bias is an unwitting occurrence rather than a deliberate strategy and so cannot be completely avoided (Nickerson, 1998). However, the consideration of alternative or competing hypotheses has been suggested as a way of reducing bias, because restricting one's attention to a single favoured hypothesis may lead to researchers not adequately considering alternative views, and not considering the relevance of information to these other views (Nickerson, 1988). Considering these alternative hypotheses may therefore allow a more balanced approach to gathering and interpreting information. In particular, Nickerson (1998) highlights the importance of competing hypotheses in enabling researchers to consider the diagnosticity of their observations, i.e., the extent to which an observation is consistent with a particular hypothesis as opposed to a different hypothesis. If only a single hypothesis is considered, this hinders the researcher's ability to determine the diagnosticity of that observation (Doherty et al., 1979; Fischhoff & Beyth-Marom, 1983; Klayman & Ha, 1987; Wolf et al., 1985; Wolf et al., 1988). In line with this, the value of testing competing hypotheses has been demonstrated among professionals during a complex intelligence analysis task (Lehner et al., 2008). Substantial confirmation bias was demonstrated among

less experienced participants, which manifested as participants weighing evidence that supported their preferred hypothesis more heavily than evidence that disconfirmed it. Among these less-experienced participants, confirmation bias was substantially reduced after using analysis of competing hypotheses as an approach.

In summary, confirmation bias is evidently a widespread phenomenon with considerable impact on individual cognition and on wider society. Although this is only one form of bias that manifests within scientific research, other forms of bias such as hindsight bias will be considered in more detail in chapter 8 which focuses on HARKing. Thus far the current chapter has only briefly considered the ways in which bias manifests within research. The following section of the introduction will briefly outline some related issues that intersect with confirmation bias, primarily the over-representation of positive results in the literature, questionable research practices such as p-hacking, and publication bias and the suppression of null results. Subsequently, the potential for RRs to help address these concerns and the existing evidence base for this will be considered.

3.1.2. P-hacking, Publication Bias, and Null Results

There has been widespread concern in recent years that positive results (i.e., supported hypotheses) are disproportionately common in the published literature in many disciplines, including psychology (Simmons et al., 2011; Heene & Ferguson, 2017; Lilienfeld et al., 2011; Jureidini et al., 2020; Schwarcz, 2019; Haeffel, 2022; Singal, 2021; Bakker et al., 2012). Specifically, Fanelli et al. (2012) reports that 95% of results they examined within psychology were positive. Similar findings are reported by Vasilev (2013) who found that 95.4% of articles they examined in European psychology journals had at least one supported hypothesis, while 73% of the papers found support for all of their tested hypotheses. This is concerning because a disproportionately high number of supported hypotheses is inevitably unrealistic, and raises questions about the validity of those conclusions, particularly in light of recent efforts to replicate key findings. For example, the Open Science Collaboration (2015) found that less than half of their replication studies could be considered to have successfully replicated the original finding. Other sources show similar outcomes, with many key findings not being successfully replicated or showing much less convincing results in the form of lower effect sizes or more conflicting or mixed evidence (Klein et al., 2018; Wagge et al., 2019).

Confirmation bias may contribute to this lack of robust findings, by leading researchers to believe their own biased conclusions and to have undeserved faith in the validity of their potentially biased research methods. Furthermore, publication bias, which can be considered a manifestation of confirmation bias, has also been suggested as an important contributor to this overabundance of supported predictions. Journals' almost universal preference for statistically significant findings leads to an over-representation in the literature of studies that have positive and tidy or 'interesting' findings, as opposed to null or inconclusive findings (Ferguson & Heene, 2012; Mervis, 2014; Franco et al., 2014; Kuhberger et al., 2014). The lack of null results published therefore means that the available body of published literature on most topics is biased and is not a fully accurate representation of the sum total of work actually done on that topic. This is also known as the file drawer problem (Rosenthal, 1979), whereby null, inconclusive or 'uninteresting' findings are not shared, thereby giving the impression that the evidence for any particular effect or finding is overwhelmingly positive. Although recent efforts to incentivise and provide outlets for the publication of null results have been encouraging, including special issues and even full journals specifically for this purpose (Journal of Trial and Error, n.d.; Devine et al., 2020; Journal of Articles in Support of the Null Hypothesis, n.d.), the overall status quo remains firmly in favour of positive findings. This may be justified in many cases, given how challenging null findings can be to interpret and the additional work that is often required for a null pattern of results to be considered convincing (for example, collecting more data, and having tried multiple ways of analysis).

However, this norm for positive results within publishing has widespread implications, particularly because of how it intersects with academic incentive structures and its 'publish or perish' culture. 'Publish or perish' refers to the expectation for researchers to publish large numbers of articles in order to demonstrate their productivity and remain competitive in hiring and promotion decisions, as the number of publications and their metrics are thought to be increasingly relied on by administrators as metrics of the researcher's success and productivity. Although some evidence suggests that such metrics are less influential in hiring decisions than researchers believe them to be (Abbott et al., 2010), the perception of their importance nonetheless creates immense pressure for researchers to publish prolifically and leads them to worry that those who don't do so may find themselves at a disadvantage compared to their peers (Rawat & Meena, 2014).

This combination of pressure to publish, combined with widespread publication bias in favour of significant findings, can incentivise some concerning practices by researchers in

order to demonstrate statistically significant results and so, increase the likelihood of their work being published (Bruton et al., 2020; John et al., 2012; Nosek et al., 2012). For example, it can lead to the prioritisation of smaller, rapidly-conducted projects rather than work that may be slower but more rigorous, as well as the practice of salami-slicing (splitting up the results of a project into multiple different papers to maximise the number of outputs from it; Menon & Muraleedhara, 2016). More concerningly, however, a large proportion of published research is thought to be affected by bias and selective reporting, in the form of p-hacking (Munafo et al., 2017).

P-hacking occurs due to 'degrees of freedom' existing for researchers in their choice of analysis approach, allowing for multiple testing of hypotheses (De Groot 1956/2014). These degrees of freedom can increase the chance of false positive findings (Wicherts et al., 2016; Ioannidis, 2005; Simmons et al., 2011) and inflate estimates of effect sizes (Wicherts et al., 2016; Ioannidis, 2008). Furthermore, this flexibility can be exploited to obtain a desired result, by exploring the data and 'fishing' for significant results, then reporting these in a manner that gives the illusion of a straightforward and planned approach (Simonsohn et al., 2014; Rubin, 2017; Andrade, 2021). This can therefore lead to hidden flexibility in the analysis and reporting of research studies which in turn can contribute to difficulties in replicating findings due to the lack of transparency about what was actually done, as well as contributing to the overrepresentation of positive findings (Simmons et al., 2011; Wicherts et al., 2016; Stefan & Schonbrodt, 2023; Bruns & Ioannidis, 2016; Head et al., 2015; Reis & Friese, 2022). Furthermore, this opportunity for researchers to explore different analysis approaches and cherry-pick results leaves them open to confirmation bias, allowing them to make biased conclusions that support their own beliefs but that may not be the most accurate reflection of the evidence.

However, having multiple possible options for analysing data is not necessarily a problem in itself. As Silberzahn et al. (2018) conclude, it may be difficult to avoid having some degree of variation in the data analysis approach due to the many reasonable choices that need to be made during this process, even when expert researchers approach this without biased or questionable motives. Therefore, transparent reporting of these analytical choices and where possible, pre-specified analysis plans, are vital. In particular, clearly distinguishing between pre-planned confirmatory analysis and exploratory analysis, is important in clearly communicating these choices to the reader (Reis & Friese, 2022). Many other potential solutions have also been proposed to reduce and detect p-hacking, including reporting Bayes

factors alongside p-values to allow readers to interpret the findings more critically, reducing false positive rates by reducing the threshold for statistical significance (Stefan & Schonbrodt, 2023), making data (including raw data) and analysis code available to allow for independent verification of the findings (Sijtsma, 2016), and crowdsourcing data analysis to allow greater transparency regarding how degrees of freedom in the analysis affect the results (Silberzahn et al., 2018). Preregistration has also been proposed as a potentially promising solution to the issue of analytical flexibility and selective reporting (Wagenmakers et al., 2012). Preregistration involves pre-specifying detailed research questions, hypotheses, design and analysis plans, prior to data collection (Wicherts et al., 2016). Such plans must be specific, precise and comprehensive in order to account for the many different contingencies that need to be considered at each step of the process (Wicherts et al., 2016). Otherwise, many degrees of freedom can still exist and be exploited by the researchers. Furthermore, researchers typically are not held accountable for adhering to their preregistered protocol, and deviations appear to occur frequently (Mathieu et al., 2009; Ramagopalan et al., 2014).

3.1.3. Lack of Clarity or Identifiability of Stated Hypotheses

Alongside risk of bias in hypothesis-testing, a major concern is the lack of clarity and specificity of hypotheses stated in the literature, and even whether these can be identified at all. This variation with which hypotheses were introduced was also noted by Scheel et al. (2021); although most of the hypotheses they identified within RRs used specific terms such as "hypothes*", "replicat*", or "test" to indicate this, they found 64 unique phrases for this in total across the 97 hypotheses they included in their study. This inconsistency in how hypotheses are described may in turn lead to a lack of clarity about whether a specific prediction has actually been stated. Such poorly-defined articulations of hypotheses even within RRs has also been highlighted by Scheel (2022) who found that many hypotheses they examined were so unclear that it was impossible to determine how they were operationalised or tested. Furthermore, the articulations of these frequently differed throughout the article, leading to confusing and incongruent statements. This led to extreme difficulty in determining whether these hypotheses had actually been supported by the data. Although Scheel states that the authors claimed to have provided evidence in relation to their hypothesis, it was typically not possible to verify this because of how vague the hypothesis articulations had been. This level of ambiguity has also been reported by other researchers (Farrar et al., 2020; Edelsbrunner and Thurn, 2020). This is a concern as such vaguely-defined predictions, even if pre-specified, leave room for researchers to interpret their

evidence in a biased manner or to find ways that support these predictions and serve their goal of achieving statistically significant results. Furthermore, although preregistration has been proposed as a potential solution for poorly-defined hypotheses, multiple studies have shown that this does not appear to have been effective in resolving this issue (van den Akker, 2021; Bakker et al., 2020; Claesen et al., 2021). For example, these were so unclear that Bakker et al. (2020) reported inter-rater reliability rates of as low as 14% when coding the number of hypotheses stated in preregistrations.

3.1.4. Registered Reports as a Potential Solution

Preregistration has been recommended as a potential solution to a number of questionable research practices, including p-hacking and selective reporting, as well as for ambiguously stated hypotheses. However, as mentioned in sections 3.1.2. and 3.1.3, preregistration does not appear to have been as successful as hoped in addressing such issues. Furthermore, pre-registration alone is insufficient to address the widespread issues of "publish or perish" culture, or publication bias since this does not necessarily increase the likelihood of journals accepting a paper.

Registered Reports, however, aim to address this issue more comprehensively. As the hypotheses are specified in advance and the in-principle acceptance (IPA) decision is made prior to data collection, this means that the subsequent stage 2 report will be published by the journal regardless of the findings, as long as the authors have adhered to their stage 1 protocol (Chambers, 2013; Chambers, 2019; Kiyonaga & Scimeca, 2019). Therefore, methods and analysis plans are essentially 'preregistered' with the journal or review platform, in a manner that holds the researchers accountable for adhering to their protocol. Furthermore, RRs aim to reduce publication bias because acceptance is not based on the existence of statistically significant results and instead is based on the importance or validity of the research question and the robustness of the methods. This reduced pressure to produce 'interesting' and significant results should therefore remove the incentives that drive selective reporting and other QRPs that lead to falsely supported hypotheses (Chambers & Tzavella, 2020).

There is encouraging initial evidence of RRs having significantly higher rates of null findings, which could indicate that RRs reduce the rate of false positive results, in line with expectations. For example, a recent study of the psychology literature by Scheel et al. (2021) showed a rate of 96% supported hypotheses in standard research articles compared with only

44% in RRs. Similarly, Allen and Mehler (2019) also showed high rates of null results among 113 RRs in psychology and biomedical research, with RRs showing 60.5% unsupported hypotheses, compared with previous studies which report rates of only 5-20% null results within standard research papers. While many factors may be involved in such differences, part of this reduction could be related to a reduction in biased reporting within RRs. Further research into these differences is warranted, particularly involving the full range of disciplines that have adopted or trialled RRs, and with a more closely matched SR sample.

The current study therefore builds on the existing literature in a number of ways, including through a partial replication (and extension) of Scheel et al.'s (2021) study. Their approach tested only the first hypothesis stated within each paper, and without regard for its level of granularity within the paper. In contrast, the current study includes all hypotheses identified within each paper and distinguishes them by the level of granularity at which they occur in the paper, reflecting the overall structure and complexity of the articles and the interdependent nature of hypotheses throughout these different levels.

Furthermore, the sample of standard reports (SRs) in the current study is more closely matched to the RRs than in the previous studies. Allen & Mehler's study does not provide its own control sample but instead compares their findings regarding the RRs to rates found in other studies of the standard research literature. Scheel et al. (2021) do provide a random sample of SRs obtained by searching for articles that use the phrase "test* the hypothes*", as per the approach used by Fanelli (2010). Although this enabled direct comparisons of the RRs and SRs, the papers were not matched in terms of journal, topic or other characteristics. While their sample of SRs was limited to the same general timeframe as their RRs (published from 2013 to 2018), the articles don't appear to have been individually matched on the specific timeframe of publication. The current study therefore builds on previous efforts by providing a comparable sample of standard reports matched on journal of publication and timeframe, as well as matching as closely as possible on the topic, design, population, and if possible, sample size., as outlined in Chapter 2 (General Methods). In addition, this study will compare several other characteristics regarding the hypotheses presented in these different article types, including the identifiability of the hypothesis statements, and the inclusion of competing hypotheses and whether these actually received mutually exclusive support.

3.1.5. Research Questions

This chapter will investigate five main research questions comparing the hypotheses in RRs and SRs. These research questions and, where applicable, their corresponding hypotheses, are outlined in Table 1 below.

Table 1

Overview of research questions and hypotheses

Research Question	Hypotheses
RQ 1: Are there differences between RRs and SRs in how identifiable the statements of the hypotheses are?	H1: RRs will be more likely than SRs to contain clearly identifiable hypotheses.
RQ 2: Are there differences in the rates of supported hypotheses between RRs and SRs?	H2: The average proportion of supported hypotheses will be lower in the RRs than in the SRs
RQ 3: How does the granularity of the hypotheses differ between RRs and SRs (i.e., the structure of how hypotheses are presented within the paper, across article, study and hypothesis level)	H3: RRs will be more likely than SRs to contain hypotheses at the more granular levels (e.g., hypothesis-level).
RQ 4: Do the article types differ in their use of competing hypotheses, e.g., are RRs more likely to include hypotheses that the authors believe will be mutually exclusive?	No specific hypotheses were specified for RQ 4.
RQ 5: To what extent do the competing hypotheses actually have mutually exclusive support (i.e., are two competing hypotheses actually then confirmed and not confirmed, respectively).	No specific hypotheses were specified for RQ 5

The following methods section will first outline the overall process of coding the hypotheses, before breaking out into specific sections that will describe the specific coding processes used to create the variables for each of the different research questions. Specifically, this chapter will cover the comparisons between RRs and SRs in terms of how identifiable the statements of the hypotheses are, whether hypotheses were supported, the granularity of the hypotheses, the inclusion of competing hypotheses, and whether the competing hypotheses actually had mutually exclusive support.

3.2. Methods

3.2.1. Overview of Initial Coding Process for Hypothesis Variables

As described more broadly in the General Methods chapter, content analysis was used to gather the initial data, using a detailed coding protocol that was developed through an iterative process. Four characteristics about the hypotheses were coded at each level during the initial coding process: whether the statement of the hypothesis was identifiable, the articulation of the statement of hypothesis, whether the hypothesis was supported, and the articulation of support for the hypothesis.

As outlined in the Chapter 2 (General Methods), hypotheses (and other details) could be stated at more than one level of the paper, depending on how the information was presented and the components of the paper, and so information could be coded at up to three levels of granularity (article, study, and hypothesis level). Variables were coded at each level that they occurred at, for maximum clarity.

Article level was considered to only exist in the coding of a paper when the article contained multiple studies. In such cases, there may be important information at the overall article level that is not captured within the individual studies, such as an overarching general hypothesis that the authors hope to answer using a series of studies, or some overall analysis (e.g. internal meta-analysis conducted by pooling the data from all of the studies after they have been conducted). If an article contained only one study, then there was unlikely to be any additional useful information at the article level and so characteristics were coded as N/A for the article level in such cases.

Whether a study only had one hypothesis, or if there were multiple sub-hypotheses stated within the study, these would be at hypothesis level. Wherever possible, we defaulted to the lowest level of granularity possible (hypothesis level) when coding these characteristics.

Variables were coded at study level if they occurred at a general study level above any specific breakdown by hypothesis-level hypotheses. It was decided to default to the lowest level of granularity possible when coding information presented within a study, e.g., if a study contained only one hypothesis, it was technically stated at both study and hypothesis level, but it was more important to code it at hypothesis level, as this was the lower level of granularity. As these details were coded at hypothesis level, these characteristics were then coded as No (for hypothesis being stated) and N/A (for other hypothesis columns) at study

level, assuming there were no relevant details at study level once the details coded at hypothesis level are accounted for.

This chapter will focus on the main methods and results in relation to the comparative analyses of hypothesis-related variables. Specific details regarding the coding of each characteristic are given in the relevant sections of the chapter and are further outlined in the coding protocol in Appendix 2.

An interim analysis was conducted by a research assistant in August 2020 using a small sample of the data regarding the support for the hypotheses. This chi-square analysis confirmed our expectations that the SRs had higher rates of supported hypotheses than the RRs did. The process of independent double-coding most of the sample used for this interim analysis also helped to validate the coding process early on, ensuring that the protocol could be used by another researcher with minimal guidance. Where assistance was needed or discrepancies were identified, these were typically due to the research assistant being an undergraduate student with less experience and knowledge, as well as lack of compliance with the protocol approach. Discrepancies were discussed and resolved prior to the interim analysis and the protocol was refined in places to improve the clarity or specificity of the guidance.

Furthermore, the perceived difficulty of coding the hypothesis variables was examined when the coding of the full sample was complete, and the specific methods and analysis used for this are outlined in the online Supplementary Appendix 1[2]. In short, there was no significant difference between RRs and SRs in the overall perceived difficulty of coding the articles, and specifically, no significant difference in the difficulty of inferring hypotheses, inferring support for hypotheses, or determining the level of coding.

3.2.2. Coding & Analysis of Identifiability of Hypothesis Statements

This aspect of the study investigated whether there were differences between RRs and SRs in how easily identifiable the statements of their hypotheses were. It was predicted that hypotheses would be clearly identifiable more frequently in RRs than in SRs, due to these having been subjected to peer review. This characteristic was often referred to as the 'clarity of the hypothesis statements' during the coding process and is still referred to this as such in the coding protocol (to preserve the accuracy of that document, as it was used during the

[2] The online Supplementary Appendices are available on this project's OSF page: https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

coding process). In this thesis, however, this characteristic has now been rephrased as the hypotheses' "identifiability", in order to more clearly indicate that this was a reflection of whether hypotheses seemed to be present in the article rather than a judgement on how specific or precise those statements were.

Data was gathered on whether each hypothesis was clearly or partially identifiable. For the purposes of this study, a hypothesis was considered to be any prediction of the outcome of the study. This was coded separately for each level of the paper, as applicable. Responses were initially coded as Yes, No, or Partially, or, in cases of competing hypotheses, Yes – competing and Partially – competing.

In many cases a statement of the hypothesis was not clearly identifiable but could still be inferred from clear research questions or attempts to replicate previous studies. This resulted in the inclusion of 'partially' stated as a response option when the protocol was developed. If unsure about whether the potential 'partially stated' hypothesis was substantial enough to be included, I aimed to be as inclusive as possible; if I could reasonably infer what the authors hoped to find, it could be coded as being partially stated/identifiable. However, if it was extremely difficult to judge whether something could be considered a hypothesis it was considered best to just code it as No if needed, i.e., That this was not a hypothesis. In order to justify the judgements of whether the hypothesis had been clearly or partially identifiable, quotes were also gathered from the paper in order to show the articulations of the hypothesis that the coding judgement was based on.

In general, clearly identifiable hypotheses included specific phrases such as 'we hypothesised that…', 'we predicted…', 'we expect that' or 'we tested the hypothesis that X affects Y'. Partially identifiable hypotheses were much less clear. These were often determined from research questions or aims that were specified by the authors, and sometimes in statements such as 'we aimed to explore whether……' or 'we investigated the effect of X on Y'; such statements were often very exploratory and so they may be coded as No but if there were sufficient details alongside this to make it possible to determine the author's predictions even without this being explicitly stated, this could then be included as a 'Partially' identifiable hypothesis. If this was not possible and the aims or research questions were clearly exploratory, these would be coded as No instead.

3.2.2.1. Coding of Article-Level Hypothesis Statements

As previously stated, hypothesis information was coded at each of the three sub-levels of the papers. For an explanation of these three levels please see the description in the General Methods chapter, particularly the example of the paper by Thai et al., (2019).

Here we focus on whether hypotheses were identifiable (or could be inferred) at the article level (i.e., at an overall/overarching level in a multi-study paper). A hypothesis might be stated at article level if multiple studies reported within the paper investigate the same overarching hypothesis that has been stated at that broader level, particularly where a result for that hypothesis has also been stated at an overarching level of the paper (e.g., in the general discussion section of the multi-study paper) rather than in a more granular way. Such a paper may also contain further hypotheses stated at study or hypothesis level but those would then be coded at those levels as well if applicable. Ideally, an article level hypothesis would be stated within the introduction section but sometimes it may not be stated explicitly until the discussion section. If an article contained only one study, then there was unlikely to be any additional useful information at the article level above and beyond what was contained at study or hypothesis level, and so the article-level characteristics were coded as N/A in those cases.

3.2.2.2. Coding of Hypothesis-Level Hypothesis Statements

Hypotheses coded at hypothesis level were typically the most granular and specific predictions within a paper, ideally mapping on to particular outcomes or analyses within the study. However, wherever possible, we sought to default to the lowest level of granularity possible when coding these characteristics and therefore the hypothesis level of the study was coded before the study level. As a result of this, even when the predictions were not as outcome-specific as mentioned above, they might still be coded at hypothesis level by default. Therefore, if the study contained only one hypothesis, this information would be coded at the hypothesis level rather than study level because of the need to default to the more granular level where possible. Usually if there were multiple hypotheses within a paper these would be coded as being present at hypothesis level since they would be most likely to be the most specific predictions within the study. As reported in section 3.3.3, the vast majority of hypotheses within each paper were coded at hypothesis-level. If a paper had no identifiable hypothesis at any level (i.e., no clearly identifiable or partially identifiable hypothesis), there was no hypothesis level; therefore, this was coded as ‘No’ for whether a

hypothesis-level hypothesis was stated, and the remaining hypothesis-related columns at hypothesis level were then coded as N/A.

3.2.2.3. Coding of Study-Level Hypothesis Statements

Once the hypothesis level of the coding sheet had been populated for a particular paper, the descriptions of the studies were checked for any remaining hypothesis statements within that study. As the more specific hypotheses had already been coded, any remaining hypotheses that did not seem to fit into hypothesis level generally constituted a much broader statement of that study's overall prediction, which encompassed or summarised the more specific hypotheses that had been coded at hypothesis level collectively, or that otherwise constituted a general collective articulation of these more granular hypotheses. Often, once the hypothesis-level hypotheses had been accounted for, there was not any study-level hypothesis or articulation left over, so this was then coded as 'No' for whether this was stated and the subsequent three columns of information in the coding sheet (articulation of hypothesis, support for hypothesis, and articulation of support for hypothesis) were coded as N/A. If, however, there was some overarching, more general prediction for the study above what had been coded at hypothesis level, this would be coded as a study-level hypothesis.

3.2.2.4. Separating out Articulations

Sometimes, articulations presented as a single sentence actually contained multiple predictions and so they needed to be broken down into separate predictions. If breaking a sentence down into separate hypotheses, from an articulation that was originally presented as one sentence, this was coded this as 'Yes' or 'Partially' as appropriate but used an asterisk and explained the situation in the notes section for greater clarity when checking back over the coding. These instances were typically coded as 'Partially' due to the need to break them down further to create clear definitive statements.

3.2.2.5. Moving Hypotheses (Article to Hypothesis Level)

If hypotheses were only stated at article level and none of the more granular levels, it was considered whether it was appropriate to move these to a lower level of granularity. For example, if all the hypotheses stated at article level were clearly tested in each of the studies within a paper, and there was clear evidence of their findings in relation to these hypotheses within each study (rather than an aggregate conclusion across the studies only), these hypotheses that had only been explicitly stated at article level could be moved to be coded at hypothesis level for each of these studies instead. In this case, hypotheses were coded as Yes or Partially at hypothesis level as appropriate, but an asterisk was added to explain in the

notes column that it had been moved from article level. Using asterisks in this manner to indicate hypotheses being moved was only necessary when moving from article level to hypothesis level as there was more of a distinction between these levels than there would be between study level and hypothesis level.

Often the decision of whether to take this approach was influenced by how the authors had presented the results or the support for the hypotheses: if this was presented more at an overall article level, continuing to code these hypotheses at article level was often considered to be best, whereas if they gave clear information within each study about the support for each of the hypotheses which were originally stated at article level, it could be more appropriate to move the coding of the hypotheses and their support to hypothesis level to reflect this. If support for these hypotheses existed both within the studies and at the overall article level, it could be appropriate to use both of these approaches simultaneously. While this approach was part of the standardised protocol, decisions about whether this action was appropriate in a given paper were made on a case-by-case basis, based on the individual structure of the paper and how it presented the information.

3.2.2.6. Variable Creation & Analysis for Hypothesis Statements

Based on the information gathered during the initial coding process, variables were created at each of the three levels of the paper and at a cumulative overall level, for what proportion of hypotheses were clearly identifiable, and what proportion were partially identifiable. This gave a total of eight variables for the 'identifiability of hypothesis statements' characteristic (i.e., two per level). Values for each variable were given as proportion scores. While the counts had also been coded, the proportion scores that were created from these were considered more informative for use in the analysis. For the three sub-level variables, each represented the proportion of hypotheses that were identifiable either clearly or partially, respectively, at that specific coding level, whereas the overall variable represented the total proportion of all the hypotheses in the paper that were clearly identifiable, or partially identifiable, respectively.

In order to analyse whether there were differences between the article types in the identifiability of the hypothesis statements, the analysis was restricted to those articles that contained hypotheses. A total of 129 articles (30 RRs and 99 SRs) had to be removed from this analysis as they did not contain any hypotheses, as far as could be determined. Therefore, far fewer RRs omitted hypotheses than SRs did (approximately 18% vs. 29%, respectively).

Following the exclusion of these unclear responses, the comparative sample for this research question consisted of 381 articles in total, which included 140 RRs and 241 SRs. Descriptive statistics were examined, with particular attention paid to the median proportions. Comparative analysis was conducted using independent-samples Mann-Whitney U tests to examine the differences in the proportions of clearly and partially identifiable hypotheses between the two article types. Data from the overall level and the three sublevels were analysed, but the results described in this chapter focus primarily on the overall-level results. Some additional details are available in the online Supplementary Appendix 2.[3]

3.2.3. Coding & Analysis of Support for Hypotheses

During the initial coding process, data were gathered on whether the hypotheses were supported. This was coded separately for the hypotheses that had been stated at each sub-level of the paper (article, study, and hypothesis level), as applicable. Response options were Yes, No, Partially, N/A or Unclear. When coding the support for competing hypotheses, the options Yes – competing, Partially – competing, No – competing, and Unclear – competing were used**.**

Where support was not stated explicitly, the research findings were interpreted to determine whether they supported the hypothesis. The response option 'Unclear' was only used in instances where it was impossible to tell. N/A was used if no hypothesis had been stated at that particular coding level and hypothesis support at this level was therefore, non-applicable. If some aspects of a hypothesis were supported but others were not, this was be coded as 'Partially', while fully supported hypotheses were coded as 'Yes' and fully unsupported hypotheses were coded as 'No'. In order to justify the coder's judgements of whether the hypothesis was supported, quotes were also gathered from the paper in order to show the articulations of the support or lack of support for the hypothesis that the coding judgement was based on.

Based on the information gathered during the initial coding process, five variables were created at each of the three levels of the paper and at an overall level, giving a total of 20 variables. Values for each variable were given as proportion scores. At each level, these variables represented the proportion of hypotheses at that level that were fully supported, that were partially supported, that were either fully or partially supported (combined), that were

[3] The online Supplementary Appendices are available on this project's OSF page: https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

not supported, or that were unclear. For the sub-level variables, each represented the proportion at that specific level, whereas the overall-level variables represented the total proportion across all the hypotheses in the whole paper. In order to calculate the proportions, the number of hypotheses involved in the calculation at the overall level was also coded during this second round of coding but the raw counts were not used further for the analysis itself.

In order to analyse whether there were differences between the article types in the proportions of support for hypotheses, this analysis was also restricted to those 381 articles (140 RRs and 241 SRs) that contained hypotheses, as per the analysis of the hypothesis statements. Descriptive statistics were examined, particularly the median proportions. Comparative analysis was conducted using independent-samples Mann-Whitney U tests to examine the differences in the proportions of support for hypotheses between the two article types. Data from the overall level and the three sublevels were analysed. Exploratory analysis was also conducted comparing the rates of unclear support for hypotheses between the two article types.

3.2.4. Granularity of Hypotheses

The granularity of the hypotheses was not explicitly coded for during the initial coding process. However, as previously outlined, the hypotheses were coded at three different levels (article, study, and hypothesis level), and so the structure of the hypothesis coding was used to create specific variables to reflect the granularity of the hypothesis structure within the paper. Based on the structure of the coding, three variables were created to reflect this structure: the proportion of hypotheses in the paper that were stated at article level, the proportion of hypotheses in the paper that were stated at study level, and the proportion of hypotheses in the paper that were stated at hypothesis level. The median proportion of hypotheses at each level was compared between RRs and SRs using the Mann-Whiney U test.

3.2.5. Competing Hypotheses

As previously outlined in the descriptions of how the statement of hypotheses and support for hypotheses were coded (sections 3.2.2. and 3.2.3.), specific response options were included for coding competing hypotheses. When coding whether competing hypotheses had been clearly or partially identified, the same approach was taken as when coding non-competing hypotheses, except that the response options Yes – competing and Partially – competing were

used. Competing hypotheses were coded at each of the three sub-levels at which they occurred, as applicable. When coding whether competing hypotheses had been supported, the same approach was taken as when coding the support for non-competing hypotheses, except that the following response options were used: Yes – competing / Partially – competing / No – competing / Unclear – competing. Support for the competing hypotheses were coded at each of the three sub-levels at which these hypotheses occurred, as applicable.

Based on the data gained from the first round of coding hypothesis statements, two variables were created at each sub-level, and at an overall level, to reflect whether competing hypotheses had been included in the paper: a categorical variable for whether competing hypotheses were included, with a response of 0 or 1 to indicate no or yes respectively, and a proportion score for the proportion of the hypotheses at that level (or overall level) that were competing.

Based on the data gained from the first round of coding hypothesis support, two variables were also created at each sub-level, and at an overall level, to reflect whether the competing hypotheses actually had mutually exclusive support: a categorical variable for whether any mutually exclusive support was found for a set of competing hypotheses, with a response of 0 or 1 to indicate no or yes respectively, and a proportion score for what proportion of the competing hypotheses at that level (or overall level) had mutually exclusive support.

Descriptive statistics were examined within the comparative samples of RRs and SRs. Chi-square analysis was used to compare the categorical variables (inclusion of competing hypotheses, and whether there was mutually exclusive support) between the two article types, while independent-samples Mann-Whitney U tests were used to compare the proportion scores (proportion of hypotheses that were competing, and proportion of competing hypotheses that had mutually exclusive support). The results for the competing hypotheses and their support are restricted to the overall level analysis only, due to the low numbers of competing hypotheses included, and thus the extremely small numbers of valid cases at the sub-levels (particularly article and study level). This is particularly the case for the ‘mutually exclusive support’ variables as there were only 30 papers in total with valid data for this.

3.3. Results

3.3.1. Results for Differences in Identifiability of Hypothesis Statements

3.3.1.1. Overall Proportions of Clearly and Partially Identifiable Hypotheses

Within the 381 articles that had identifiable hypotheses, there was little difference between the article types in the proportion of clearly or partially identifiable hypotheses included in the articles overall. Figure 4 below displays the counts of clearly and partially identifiable hypotheses per article type.

As outlined in the methods section, the counts were used to create proportion scores for the proportion of the hypotheses in each article that were clearly identifiable and partially identifiable, respectively. When the proportion scores for the clearly identifiable hypotheses were investigated, both RRs and SRs had a median proportion of 1, and the mean proportion was very similar in RRs (M = 0.69, SD = 0.43) and in SRs (M = 0.68, SD = 0.43). Furthermore, an independent-samples Mann Whitney U test showed no statistically significant difference between the two article types in the proportion of clearly identifiable hypotheses at the overall level (U = 17119, p = 0.79).

Likewise, there was little difference between the article types in the proportion of partially identifiable hypotheses included in the articles overall, with both having a median proportion of 0, and the mean proportion similar in RRs (M = 0.32, SD = 0.43) and SRs (M = 0.33, SD = 0.43). Results of a Mann-Whitney U test showed no statistically significant difference between the two article types in the proportion of partially identifiable hypotheses at the overall level (U = 16621, p = 0.79).

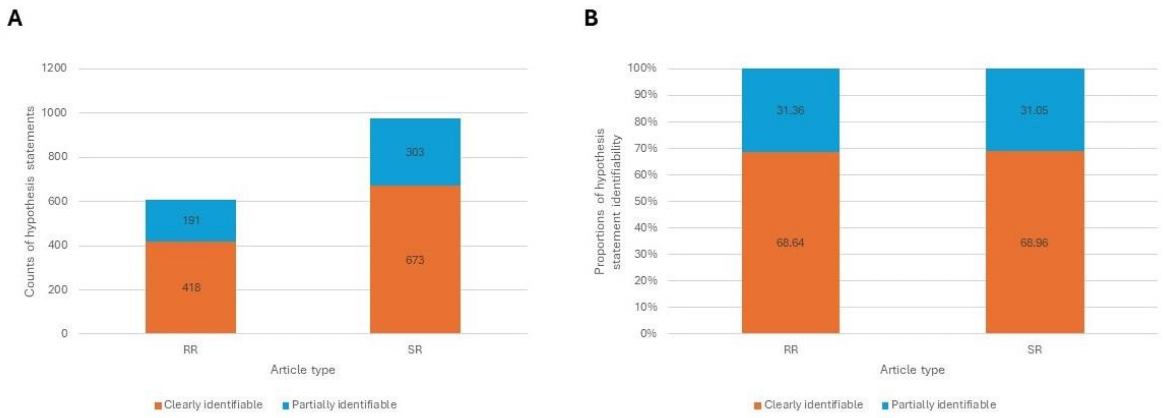

Figure 4: ***A.*** *Stacked bar chart showing the total counts of clearly and partially identifiable hypotheses per article type.* ***B.*** *Stacked bar chart showing the proportions of clearly and partially identifiable hypotheses per article type.*

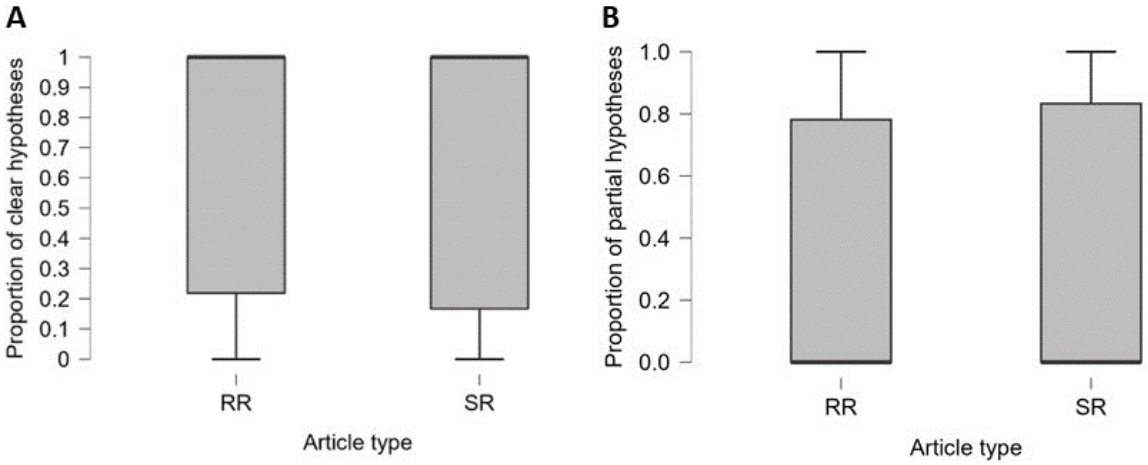

Figure 5: Boxplots showing the median proportions and distributions for the identifiability of hypothesis statements. ***A****. Boxplot A shows the median proportion and distribution of clearly identifiable hypotheses in RRs and SRs. The median proportion was 1 for both article types, and there was very little difference in the distribution of the values between these two formats.* ***B.*** *Boxplot B shows the median proportion and distribution of partially identifiable hypotheses for both article types, and also shows very little difference in the distribution of these values between the two article types.*

3.3.1.2. Sub-Level Results: Proportions of Clearly and Partially Identifiable Hypotheses

The same analysis approach was taken to examine this pattern at each of the three sub-levels (article, study and hypothesis level), and the results are outlined in more detail in the online Supplementary Appendix 2. Like the overall level analysis, the results of independent-samples Mann-Whitney U tests conducted at each of these levels all showed no statistically significant difference between the article types. For the proportions of clearly identifiable hypotheses, the pattern of medians was the same for all of the levels (i.e., median was 1 for both the RRs and SRs), while the means showed little difference between the article types at each of the levels.

Similarly, when the three sub-levels were examined, no significant differences were found between the article types in the proportion of partially identifiable hypotheses included,

although this did approach significance at article level. More specific details about the sub-level analysis for partially identifiable hypotheses in the comparative sample can be found in the online Supplementary Appendix 2.

3.3.2. Results for Differences in Support for Hypotheses

Overall, the results showed higher rates of supported hypotheses and lower rates of unsupported hypotheses in SRs than in RRs. The breakdown of the counts per level of hypothesis support is shown in Figure 6 below. In figure 7, the counts of supported and unsupported hypotheses are displayed, when clearly and partially supported hypotheses are combined as one and unclear support is excluded.

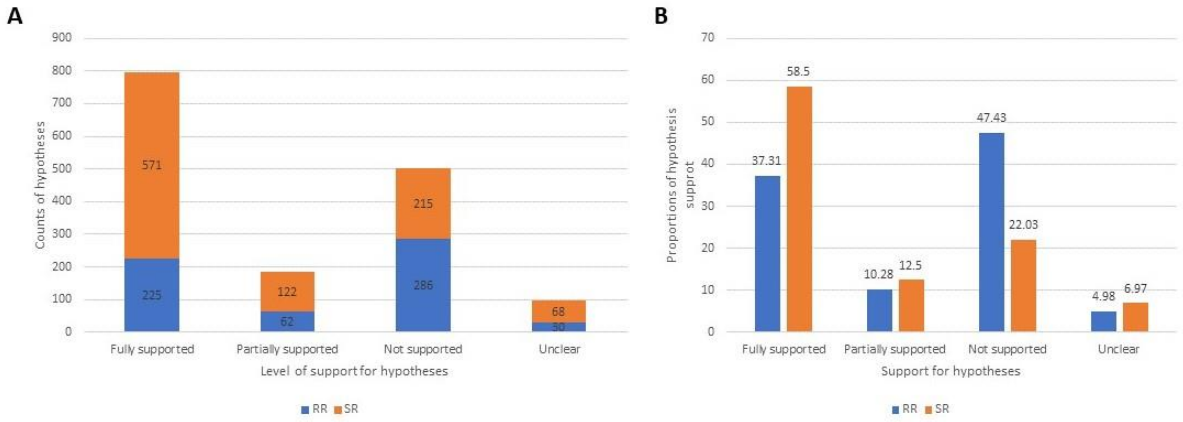

Figure 6: ***A.*** *Bar chart showing the total counts of the different levels of support for the hypotheses.* ***B.*** *Bar chart showing the proportions of the different levels of support for the hypotheses in both article types.*

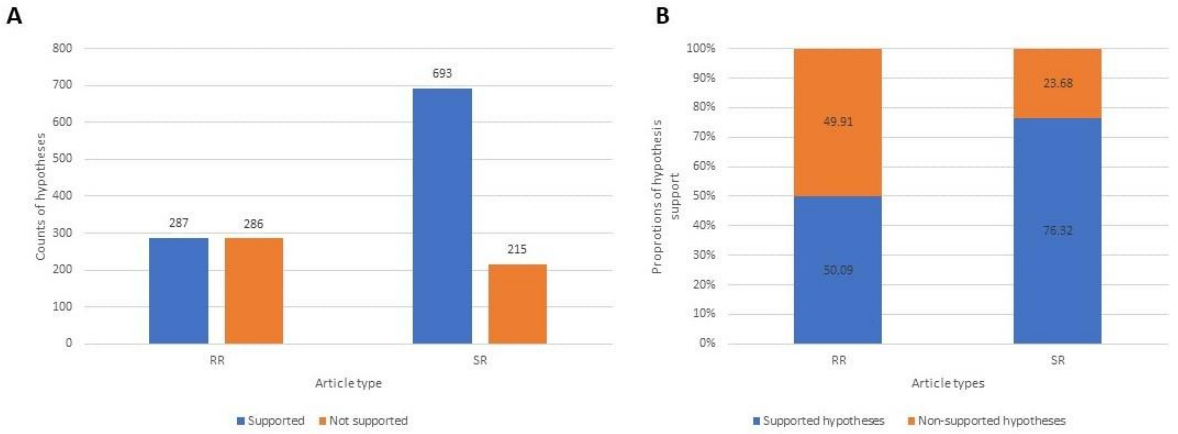

Figure 7: ***A.*** *Bar chart showing the counts of (clearly and partially) supported and non-supported hypotheses in RRs and in SRs. This shows almost equal proportions of supported and non-supported hypotheses in RRs, while far more hypotheses were supported than unsupported in the SRs.* ***B.*** *Bar chart showing the proportion of combined supported vs. non-supported hypotheses in RRs and SRs*[4].

While the raw data displayed in the figures above are based on the number of counts for each level of support, the results of the analyses reported in the following sections (3.3.2.1. to 3.3.2.5.) used the proportion scores that were obtained from this data.

3.3.2.1. Proportions of Fully Supported Hypotheses

The median proportion of fully supported hypotheses was higher among the SRs (0.67) than among the RRs (0.25). The same general pattern was seen in the mean proportions, with fully supported hypotheses more common among the SRs (M = 0.62, SD = 0.38) than the RRs (M = 0.35, SD = 0.35). An independent-samples Mann-Whitney U test showed that there was a statistically significant difference between the article types in the proportion of fully supported hypotheses (U = 10386.00, $p < 0.001$).

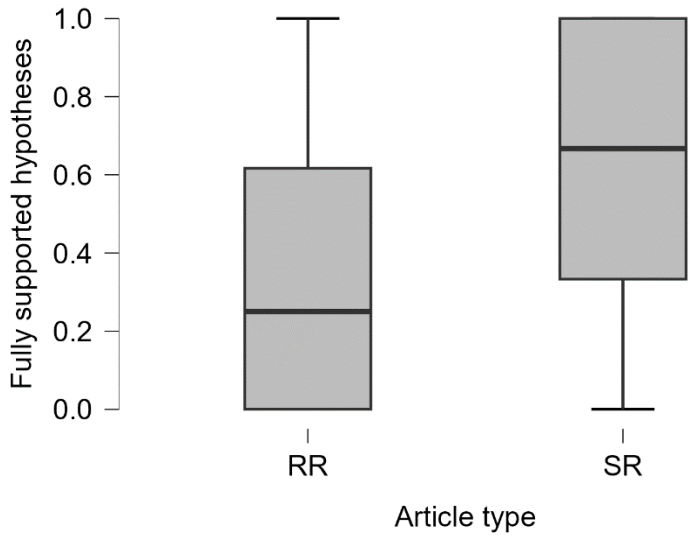

Figure 8: Boxplot showing median proportions and distributions of fully supported hypotheses in RRs and SRs.

[4] For the purposes of Figure 7, the hypotheses whose support was unclear has been omitted, and thus the figure's percentages for the rate of supported hypotheses may differ slightly from the percentages obtained by combining the percentages of fully and partially supported hypotheses shown in figure 6.

The same general pattern was seen for fully supported hypotheses at each of the three sub-levels, with SRs having much higher average proportions of fully supported hypotheses than RRs, and all independent-samples Mann-Whitney U test comparisons showing a statistically significant difference. Specifically, analysis at the article level showed a median proportion of 1 for the SRs and 0 for the RRs, while SRs also had a higher mean proportion (M = 0.64, SD = 0.45) than RRs (M = 0.18, SD = 0.36), and a statistically significant difference between the article types (U = 351.00, $p < 0.001$). Results of the analysis at study level also showed a median proportion of 1 for SRs and 0 for RRs, a higher mean proportion among SRs (M = 0.66, SD = 0.47) than RRs (M = 0.26, SD = 0.44), and a statistically significant difference (U = 161.50, p = 0.005). Hypothesis-level results also showed a median proportion of 0.67 in SRs and 0.33 in RRs, a higher mean proportion in SRs (M = 0.62, SD = 0.38) than in RRs (M = 0.38, SD = 0.36), and a statistically significant difference (U = 9689.50, $p < 0.001$).

3.3.2.2. Proportions of Supported Hypotheses (Combined Fully & Partially Supported)

The results for the partially supported variable on its own are available in Supplementary Appendix 2[5], while this section reports the findings of the analysis when the clearly and partially supported hypotheses are combined to form a single, more inclusive 'supported' variable.

The median proportion of supported hypotheses was higher among SRs (1.00) than among RRs (0.5). This was also the case with the mean proportions, with SRs having a higher mean proportion (M = 0.78, SD = 0.31) than RRs (M = 0.46, SD = 0.38). The independent-samples Mann-Whitney U test results show a statistically significant difference between the two article types in the proportion of supported hypotheses (U = 8977.00, $p < 0.001$).

[5] Supplementary Appendices are available on the project's OSF page: https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

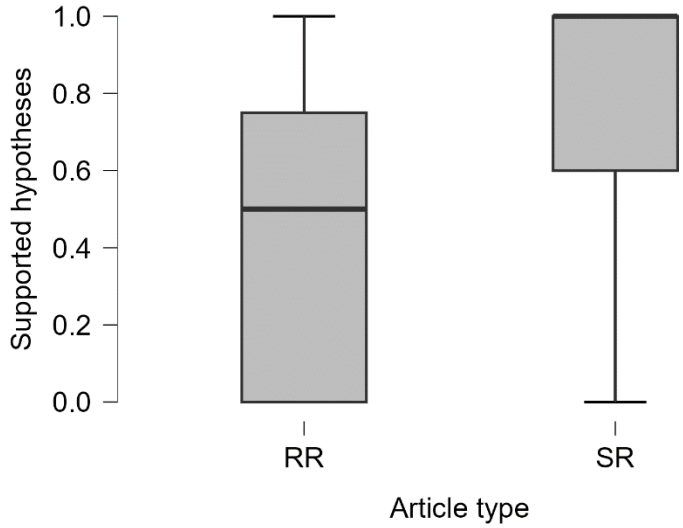

Figure 9: Boxplot showing median proportions and distributions of supported (full and partially) hypotheses in RRs and SRs.

The same general pattern was seen at each of the three sub-levels, with SRs having higher average proportions of supported hypotheses than RRs did. Specifically, analysis at the article level showed the median proportion of supported hypotheses in SRs was 1, compared with 0.31 in RRs. Likewise, the mean proportion was higher in SRs (M = 0.82, SD = 0.35) than in RRs (M = 0.45, SD = 0.47). The independent-samples Mann-Whitney U test showed a statistically significant difference between the two article types in the proportion of supported hypotheses at article level (U = 441.00, $p < 0.001$).

Results of the analysis at study level also showed a median proportion of 1 for SRs and 0.5 for RRs, and a higher mean proportion in SRs (M = 0.83, SD = 0.37) than in RRs (M = 0.5, SD = 0.5). The independent-samples Mann-Whitney U test results showed a statistically significant difference between the two article types in the proportion of supported hypotheses at study level (U = 185.00, p = 0.01). Finally, the analysis of supported hypotheses at hypothesis level also showed a median proportion of 1 for SRs, and 0.5 for RRs), as well as a higher mean proportion in SRs (M = 0.78, SD = 0.31) than in RRs (M = 0.48, SD = 0.37). Results of the independent-samples Mann Whitney U test showed a statistically significant difference between the article types in the proportions of supported hypotheses at hypothesis level (U = 8077.00, $p < 0.001$).

3.3.2.3. Proportions of Non-Supported Hypotheses

The median proportion of non-supported hypotheses was higher among the RRs (0.5) than among the SRs (0.0). Likewise, the mean proportions of non-supported hypotheses were

higher among the RRs (*M* = 0.51, *SD* = 0.38) than in SRs (*M* = 0.17, *SD* = 0.28). An independent-samples Mann-Whitney U test showed that there was a statistically significant difference between the two article types in the proportion of non-supported hypotheses (*U* = 25357.00, *p* < 0.001).

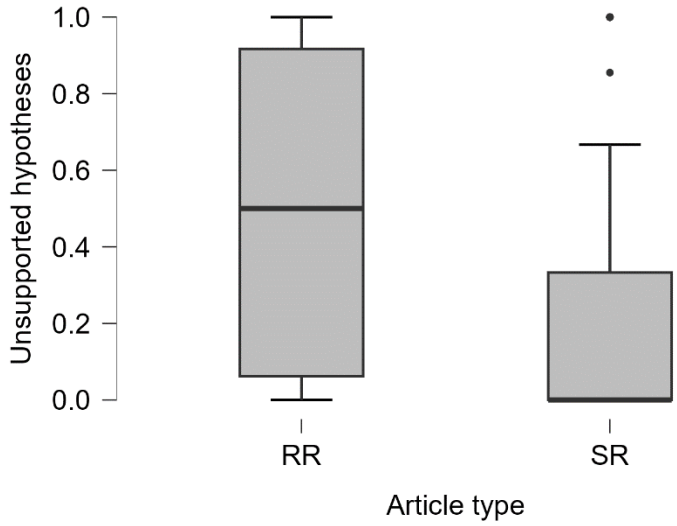

Figure 10: Boxplot showing the median proportions and distributions of non-supported hypotheses in RRs and SRs.

A similar pattern was seen from the analyses at the three sub-levels, with average proportions of non-supported hypotheses higher in RRs than in SRs and the independent-samples Mann-Whitney U tests showing statistically significant differences between the article types at each sub-level. Specifically, at article level, RRs had a higher average proportion of non-supported hypotheses (median 0.5, *M* = 0.50, *SD* = 0.46) than SRs did (median 0, *M* = 0.14, *SD* = 0.31). The results of the Mann-Whitney U test showed that there was a statistically significant difference between the two article types in the median proportion of non-supported hypotheses at article level (*U* =1061.00, *p* < 0.001).

At study level, both article types had a median proportion of 0, but RRs had a higher mean proportion (*M* = 0.45, *SD* = 0.50) than SRs (*M* = 0.11, *SD* = 0.32). The Mann Whitney U test showed a statistically significant difference between the article types in the proportion of non-supported hypotheses at study level (*U* = 385.00, *p* = 0.01). At hypothesis level, RRs also had a higher average proportion of non-supported hypotheses (median 0.5, *M* = 0.49, *SD* = 0.38) than SRs did (median 0, *M* = 0.16, *SD* = 0.27). The Mann Whitney U test showed a

statistically significant difference between the two article types in the proportion of non-supported hypotheses at hypothesis level (U = 22395.00, p < 0.001).

3.3.2.4. Proportions of Hypotheses with Unclear Support

In interpreting whether hypotheses were supported during the first round of coding, it could not always be clearly determined whether these were supported. In some cases this was because the information did not seem to be given in the paper or it could not be clearly identified. In other cases, although the evidence or articulation of support could be found in the paper, it could not be interpreted by the coder, either due to a lack of clarity in the information given, or because of the coder's lack of understanding or familiarity with the methods or topic used. In such instances the support for the hypothesis was coded as Unclear. This was checked again during the second round of coding (variable creation) in order to try to clarify the remaining uncertainties. However, this often could still not be achieved and so these remained coded as Unclear. In order to understand to what extent such issues affected the current dataset and whether this differed between the article types, descriptive statistics and independent-samples Mann-Whitney U test comparisons were examined.

The median for the proportion of hypotheses with unclear support was 0 for both article types, and the mean proportions were similar between RRs (M = 0.03, SD = 0.13) and SRs (M = 0.06, SD = 0.18). Mann Whitney test results showed that there was no statistically significant difference between the two article types in the proportion of hypotheses with unclear support (U = 16077.50, p = 0.18).

Similarly, at each of the three sub-levels, there was no statistically significant difference between the article types in the proportion of unclear support for the hypotheses. At article level, the median proportion of unclear support was 0 for both article types, and similar mean proportions were found for RRs (M = 0.06, SD = 0.22) and SRs (M = 0.04, SD = 0.18). Mann Whitney U test results showed that there was no statistically significant difference between the two article types in the proportion of hypotheses at article level where the support was considered unclear or impossible to determine by the coder (U = 761.00, p = 0.89).

At study level, the median proportion was 0 for both article types and the mean proportions were similar between RRs (M = 0.05, SD = 0.22) and SRs (M = 0.06, SD = 0.22), while the Mann Whitney U test showed no statistically significant difference between the article types (U = 267.50, p = 0.75). At hypothesis level, the median proportion for both article types was also 0. The mean proportion of unclear support was 0.03 (SD = 0.13) for RRs and 0.06 (SD =

0.19) for SRs. The Mann Whitney U test showed no statistically significant difference in the proportion of hypothesis support that was considered unclear by the coder, at hypothesis level (U = 14465.00, p = 0.33).

3.3.3. Results for Differences in Granularity of Hypotheses

Overall, the average proportions of hypotheses at each of the three levels of granularity did not differ between article types, and no statistically significant difference was found from the independent-samples Mann-Whitney U tests. The counts for the numbers of hypotheses at each of the three levels are displayed in Figure 11.

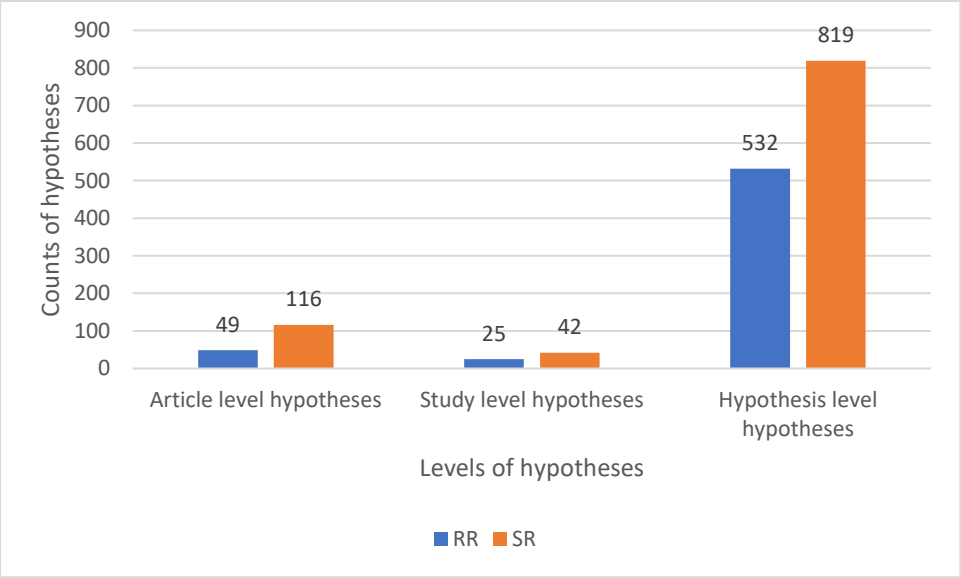

Figure 11: Counts of hypotheses at each of the three levels within the articles. The figure shows that both RRs and SRs had most of their hypotheses at hypothesis level, although smaller numbers were also stated at an overarching article level, or within studies but at a broader level than the more specific hypothesis level (i.e., study-level). As this sample included a larger number of SR articles than RRs, there are greater numbers of hypotheses represented here for SRs than for RRs.

Specifically, the median proportion of article-level hypotheses was 0 for both SRs and RRs. Likewise, the mean proportion was similar for RRs (M = 0.08, SD = 0.23) and SRs (M = 0.12, SD = 0.26). The results of the Mann Whitney U test showed no significant difference between the article types in the proportion of article-level hypotheses (U = 15877.50, p = 0.19). The median proportion of study-level hypotheses was similar for both RRs (median 0, M = 0.03 SD = 0.08) and SRs (median 0, M = 0.03, SD = 0.09). Mann Whitney U tests

showed no significant difference between article types in the proportion of study-level hypotheses (U = 17458.00, p = 0.325). The median proportion of hypothesis-level hypotheses was 1 for both article types. The mean proportion of hypotheses at the hypothesis level of granularity was similar for RRs (M = 0.89, SD = 0.24) and SRs (M = 0.86, SD = 0.27). The Mann Whitney U test results showed no significant difference between the article types in the proportion of hypothesis-level hypotheses (U = 17299.00, p = 0.62). Overall, then the article types did not show significant differences in the proportions of hypotheses at each level.

3.3.4. Results for Differences in Competing Hypotheses

The categorical measure of whether competing hypotheses were included (at the overall level) showed that these were only represented in a small proportion of the overall sample and that the rates of inclusion were very similar between RRs (7.86%) and SRs (7.88%). Chi square analysis did not show a significant difference between the article types in the inclusion of competing hypotheses: $x^2(1, N = 381) = 8.686x10^{-5}$, p = 0.99.

The median proportion of competing hypotheses at the overall level was 0 for both article types, and the means were also similar for RRs (M = 0.60, SD = 0.21) and SRs (M = 0.60, SD = 0.22). The Mann-Whitney U test comparing the proportion of competing hypotheses between the two article types, was non-significant (U = 17031.50, p = 0.74).

Within the 30 articles that had competing hypotheses (18 SRs and 12 RRs), frequencies for the categorical support variable showed that SRs had a higher rate of mutually exclusive support for their competing hypotheses than RRs did (72.22% vs. 58.33%). However, a chi-square analysis showed no significant difference in this between the article types, which may be due to the very small sample available for this analysis: $x^2(1, N = 30) = 0.63$, p = 0.43.

The median proportion of mutually exclusive support for competing hypotheses was higher in the SRs than in the RRs (1 vs. 0.572). Likewise, the mean proportion was higher in SRs than in RRs (M = 0.67, SD = 0.44, vs. M = 0.51, SD = 0.51). However, the independent-sample Mann-Whitney U test showed no significant difference between the article types in the proportion of mutually exclusive support for competing hypotheses (U = 95.5, p = 0.57), which is likely due to the very small sample size.

3.4. Discussion

The current chapter compared the characteristics of hypotheses between RRs and SRs. Key outcome measures were the identifiability of the hypothesis statements, the support for the hypotheses, the granularity of the hypotheses within the articles, whether competing

hypotheses were included, and whether mutually exclusive support was found for those competing hypotheses.

3.4.1 Recap of Main Findings

Preliminary analysis revealed significant differences between RRs and SRs in the proportions of supported hypotheses. Overall, RRs showed much lower proportions of supported hypotheses than SRs did, and higher rates of unsupported hypotheses. There was no significant difference found between the article types in the perceived difficulty of coding the hypothesis variables, or in the proportion of hypotheses for which the support couldn't be determined (i.e., proportions of unclear support for the hypotheses). There were also no significant differences noted in how easily the stated hypotheses were identified, the level of granularity of the hypotheses, the inclusion of competing hypotheses, or in the existence of mutually exclusive support for the competing hypotheses.

Table 2

Overview of support for chapter 3 hypotheses

Hypotheses	Conclusions
H1: RRs will be more likely than SRs to contain clearly identifiable hypotheses.	H1 was not supported. There was no significant difference between RRs and SRs in the identifiability of hypothesis statements
H2: The average proportion of supported hypotheses will be lower in the RRs than in the SRs	H2 was supported. The proportion of supported hypotheses was significantly lower in RRs than in SRs.
H3: RRs will be more likely than SRs to contain hypotheses at the more granular levels (e.g., hypothesis-level).	H3 was not supported. There was no significant difference between RRs and SRs granularity of the hypotheses

3.4.2. Discussion of Results & Comparison to Previous Research

The identifiability of hypotheses in RRs and SRs does not appear to have been formally compared between RRs and SRs (although Scheel et al. do discuss this issue within RRs more generally) and so the inclusion of this comparison in this study presents a novel finding. This finding that there was no difference in the identifiability of the hypotheses is somewhat surprising since the emphasis on pre-specifying and peer-reviewing detailed hypotheses is a core component of the RR model and so it was expected that these would be clearer in the RRs than in the SRs. As the comparison articles came from the same journals within the same timeframe, this may be a reflection of the journals' approaches and requirements in terms of

how hypotheses are presented. Alternatively, this could be driven by HARKing in SRs giving the appearance of clearer and more identifiable hypotheses by generating hypotheses from the results and retrofitting these onto the study's rationale, allowing these hypothesis statements to be defined more explicitly and precisely than they might be if they had been predicted *a priori.*

In any case, the same ambiguities affecting hypotheses stated in standard reports also appear to be an issue in the RRs. This supports Scheel's (2022) reports of an extreme lack of clarity and specificity in relation to the hypotheses/claims included in RRs. Similar findings have been reported by various authors in relation to hypotheses of preregistered studies more generally (van den Akker 2021; Bakker et al., 2020; Claesen et al., 2021). While it was hoped that RRs may improve on this due to the intensive peer review of the study protocols, the current study seems to suggest that this has not been the case, at least within the sample analysed here. This is concerning since, as previously stated, this ambiguity may give researchers the flexibility that allows bias to creep into their interpretations of the support for those poorly-defined hypotheses. However, it may be unrealistic to expect RRs to entirely resolve this issue given how widespread it appears to be, affecting null-hypothesis significance testing on a broad scale across many disciplines. Therefore, RRs alone may not be sufficient to effectively improve the identifiability, clarity and specificity of hypothesis statements and instead a much wider-scale shift in research practices may be required. For example, Scheel et al. (2020) point to the need for hypotheses to be created based on a sound derivation chain, and for there to be much greater emphasis on exploratory research before turning to confirmatory research to test hypotheses, since the underlying theoretical basis is generally not sufficiently established to allow well-defined and testable hypotheses to be created.

As previously mentioned, however, an alternative possibility is that the comparison of the identifiability of the hypotheses could be confounded by HARKing in the SRs. If so, then the similarity in how clearly hypotheses can be identified in the two article types could suggest that the RR process does improve the identifiability of the hypotheses, at least to the same extent that HARKing pretends to improve it in SRs. While investigating this is beyond the scope of the current study, this possibility offers one possible explanation for the results found.

The granularity of the hypothesis statements does not appear to have been previously investigated. It was also somewhat surprising that this didn't differ significantly between the article types since the RRs generally had a much more clearly structured layout and typically seemed to have hypotheses more clearly indicated, based on personal experience when coding the articles. However, the statistical analysis did not support this general sense of the granularity. It is also possible that the more complex granularity or structure may contribute to the difficulties in inferring the statement of the hypotheses and the support for the hypotheses, where there are multiple levels involved, for example, matching the support at a particular level to its statement, and disentangling the multi-level structure of the hypotheses stated, and so this may have influenced the data gathered. It certainly did influence the difficulties experienced when coding these details, with determining the coding level being reported as an issue for both RRs and SRs. The coding of the causes of difficulty did not consider the intersecting nature of some of these causes but anecdotally this issue was primarily in relation to coding the levels of the hypotheses and their support, and these three aspects were the main major challenge of the coding process. With that in mind it is important to be cautious regarding the results of these analyses as the coding process was inevitably subjective, as is common in such study designs.

The significantly higher rates of confirmed hypotheses within the SRs than the RRs confirmed H2 (see section 3.1.5 above), and also supports previous findings reported by Scheel et al. (2021) and by Allen and Mehler (2019). For example, depending on the statistic used, proportions of supported hypotheses were found to be almost twice as common in SRs as in RRs in the current study. This is a similar pattern to that found in Scheel et al.'s (2021) study which found that approximately 95 to 96% of SRs had supported hypotheses, compared with between 43 and 50% of the RRs, depending on whether replications were included. Furthermore, the median proportions of supported hypotheses (fully and partially supported combined) from the current study are very similar to the proportions reported in Scheel's work.

These similar findings are encouraging, particularly as there were some key differences in the methods used between these studies. Specifically, the current analysis was conducted on all hypotheses stated within these papers. As acknowledged by Scheel et al., their focus on only the first stated hypothesis may have influenced their findings, as the first hypotheses may not be representative of all hypotheses within a particular paper: they suggest that authors of SRs may be more likely to present their supported hypotheses first, whereas authors of RRs may

be more likely to present their hypotheses in chronological order instead. Therefore, they speculated whether supported hypotheses may be slightly overrepresented among the SR data they used, compared with their RR data. However, the current study's analysis across all of the hypotheses revealed very similar results which supports their common findings. The current study also differs from Scheel's approach in a number of ways, including a more closely matched control sample, and a more granular approach to the coding, which could also contribute to any differences from the findings reported by the previous studies. Additionally, as only three of the journals in Scheel et al.'s dataset provided both RRs and SRs, one potential confound considered in their discussion was the possibility that journals which publish RRs may have other editorial policies that promote rigour and transparency of reporting, such as the TOP guidelines, or open data policies, and so the quality of *all* articles within the journals they sourced RRs from may have been higher anyway. Therefore, the differences between the RRs from these journals and the SRs from different journals may be more striking than if the SRs had all been sourced from within the same journals as the RRs. The current study addresses this concern by selecting SRs from within the same journal as the RR (and within the same timeframe) as these would be subjected to any other relevant editorial policies. Despite these differences, results for hypothesis support were closely aligned overall, pointing to robustness of the overall conclusion that hypotheses in RRs are less likely to be confirmed.

3.4.3. Limitations

It is important to be cautious in interpreting findings and to acknowledge that the differences noted between the two article types in these preliminary analyses could be due to a range of factors. For example, the type of hypotheses researchers choose to investigate using RRs could differ from those they are likely to investigate using a standard approach. There has been some speculation that RRs may be used by researchers to test hypotheses which they are more sceptical of and which they expect may be less likely to be supported, because the in-principle acceptance decision protects authors against publication bias when testing risky predictions (Chambers & Tzavella, 2022). Alternatively, the RR format may appeal to researchers who are more conscientious about the quality/rigour of their approach, which could contribute to higher standards in the RRs than in the SRs (Soderberg et al., 2021).

Additionally, while any differences found between the article types may indicate a reduced level of bias in the conduct and reporting of the studies, we must acknowledge the limitations of the observational data involved and the fact that causality cannot be inferred from this

dataset. Nevertheless, it is hoped that this study will give an indication of the differences between these article types.

Coders' awareness of the article types being coded also raises the possibility of bias in our interpretations of the article characteristics. However, it was not considered feasible within the current study for the coders to be blinded to the article type, particularly as, in order to be truly blinded to this, certain information would need to be redacted from the articles, especially from the RRs. This was also highlighted by Scheel et al. (2021) as a limitation of their study. While experimental or quasi-experimental approaches to the coding of articles would be beneficial in the future, the current study represents a first look at the differences between these article types and so hopes to inform future causal studies.

Furthermore, as the process of developing the protocol was necessarily iterative and so had to be allowed to be flexible, this study was not deemed suitable for preregistration. This introduces the potential for bias in the coding approach and the analysis. In order to mitigate the effects of any potential bias, the coding was checked as previously described; while those checking the coding could also be biased by their own knowledge of the article type, we believed that we were likely to be able to at least verify the accuracy of the coding and identify any clear errors in interpretation or judgements. Furthermore, by making the protocol and database open to the public at the end of the study, I would encourage the wider community to verify this approach and to help in identifying any systematic bias that may have influenced our coding judgements.

Finally, Scheel et al. (2021) noted that their reliance on the use of the term 'test* the hypothes*' to gather their SR sample may have been a limitation as it may not have been representative of the full range of literature that focuses on testing hypotheses. Our approach avoids this issue by not using this term as the matching criterion. However, we instead have some SRs which do not appear to have tested any hypotheses and instead are purely exploratory (or at least must be assumed to be purely exploratory). As RRs are primarily intended to be for confirmatory research, comparing them against any articles which are not clearly confirmatory is a potential limitation of the current study's matching approach, although the fact that there are two SRs available for each RR should help to mitigate this in many cases.

3.4.4. Implications of Study and Future Directions

The comparisons conducted here demonstrate how RRs differ from standard research papers, and so could be helpful in informing journal policies and decision-making regarding use of the RR format, as well as other open science initiatives. In particular, the findings illustrate the areas in which RRs do not appear to have been as successful as would be hoped, such as in the identifiability of hypothesis statements, and the inclusion of competing hypotheses. Therefore, these represent areas for further improvement of the RR model and how it is implemented within journals.

Due to the difficulty of inferring hypotheses from unclear articulations, it is recommended that the identifiability, clarity and specificity of hypotheses may be an important area for future metascience studies to investigate in order to understand the true extent of the problem, as well as a priority for journals to address by raising standards for what is considered an acceptable statement of the study's hypothesis. This is particularly necessary to address as the lack of clarity regarding what the hypotheses actually are leaves these open for misinterpretation by researchers who are seeking opportunities to tweak their hypotheses post-hoc in order to line up with the results they have obtained, i.e., HARKing (which is discussed in more detail in chapter 8).

The low rate of competing hypotheses included in the studies was not particularly surprising given the low application of this approach in published research generally/within psychology. Likewise, the lack of a significant difference in this between the article types was somewhat expected after the relatively small sample size for this analysis became clear. Although I might initially have expected the RRs to have higher rates of competing hypotheses in an effort to more rigorously test their predictions, the small sample size for this analysis meant that only a very large effect would be possible to detect, if any effect were present. In future, larger samples should help to clarify this in a more robust way. In the meantime, efforts should be made to promote the use of competing hypotheses in order to test predictions more rigorously and journal editors could be instrumental in encouraging this, perhaps through journal policies or themed special issues. However, in many cases, theories may not be sufficiently developed to be tested using confirmatory hypothesis-testing approaches and so greater use of exploratory and descriptive research that is clearly declared as such, would also be important.

3.5. Conclusion

This chapter has outlined some specific details of the process of coding the hypothesis variables and has described the results obtained for the comparative analysis in relation to identifiability of hypothesis statements, support for hypotheses, granularity of hypotheses within the papers, inclusion of competing hypotheses, and the extent to which those competing hypotheses actually have mutually exclusive support. In summary, the evidence found that RRs had substantially lower rates of supported hypotheses than SRs, which is in line with our hypotheses and with previous research. No reliable differences were observed between the article types in how identifiable the statements of the hypotheses were, the granularity of the hypotheses within the papers or the inclusion of competing hypotheses and rates of mutually exclusive support for these. Findings regarding the rates of support for hypotheses support the conclusions of previous metascientific studies.

3.6. References

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Metrics: do metrics matter? *Nature, 465*, 860-862.

Agarwal, P., (2018, October 19). Here is how bias can affect recruitment in your organisation. https://www.forbes.com/sites/pragyaagarwaleurope/2018/10/19/how-can-bias-during-interviewsaffect-recruitment-in-your-organisation

Allen, C., & Mehler, D.M.A. (2019). Open science challenges, benefits and topics in early career and beyond. *PloS Biology, 17*(5), e3000246.

Anderson, C.A. (1983). Abstract and concrete data in the perseverance of social theories: When weak data lead to unshakeable beliefs. *Journal of Experimental Social Psychology, 19*(2), 93-108

Anderson, C. A., Lepper, M. R., & Ross, L. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39*(6), 1037–1049.

Andrade, C. (2021). HARKing, cherry-picking, p-hacking, fishing expeditions, and data dredging and mining as questionable research practices. *Journal of Clinical Psychiatry, 82*(1), Article 20f13804. Doi: 10.4088/JCP.20f13804

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554.

Bakker, M., Veldkamp, C. L. S., van Assen, M. A. L. M., Crompvoets, E. A. V., Ong, H. H., Nosek, B. A., Soderberg, C. K., Mellor, D., & Wicherts, J. M. (2020) Ensuring the quality and specificity of preregistrations. *PloS Biology, 18*(12), e3000937.

Beattie, J., & Baron, J. (1988). Confirmation and matching biases in hypothesis testing. *The Quarterly Journal of Experimental Psychology Section A, 40*(2), 269–297.

Bruns, S. B. & Ioannidis, J. P. A. (2016). P-curve and p-hacking in observational research. *PLOS ONE, 11*(2), e0149144

Bruton, S.V., Medlin, M., Brown, M., & Sacco, D.F. (2020). Personal motivations and systemic incentives: scientists on questionable research practices. *Science and Engineering Ethics*. Retrieved from https://link.springer.com/article/10.1007/s11948-020-00182-9

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex, 49*(3), 609-610.

Chambers, C. (2019). The registered reports revolution: lessons in cultural reform. *Significance, 16*(4), 23-27.

Chambers, C.D. & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*, 29–42.

Chapman, L. J., & Chapman, J. P. (1969). Illusory correlation as an obstacle to the use of valid psychodiagnostic signs. *Journal of Abnormal Psychology, 74*(3), 271–280.

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: An assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science, 8*, 211037.

Consul, N., Strx, R., DeBenedectis, C.M., & Kagetsu, N.J. (2021). Mitigating unconscious bias in recruitment and hiring. *Journal of the American College of Radiology, 18*(6), 769-773

Cooper G. S., & Meterko V. (2019). Cognitive bias research in forensic science: A systematic review. *Forensic Science International, 297*, 35–46.

Darley, J.M. & Gross, P.H. (1983). A hypothesis-confirming bias in labeling effects. *Journal of Personality and Social Psychology*, *44*(1), 20-33.

Davies, M. F. (1997). Belief persistence after evidential discrediting: The impact of generated versus provided explanations on the likelihood of discredited outcomes. *Journal of Experimental Social Psychology, 33*(6), 561–578.

De Groot, A. D. (1956/2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica, 148*, 188-194

Devine, S., Bautista Perpinyà, M., Delrue, V., Gaillard, S., Jorna, T. F. K., Meer, M. van der, Millett, L., Pozzebon, C., & Visser, J. (2020). Science fails. Let's publish. *Journal of Trial & Error*, *1*(1), 1-5. https://doi.org/10.36850/ed1

Ditto, P. H. & Lopez, D. F. (1992). Motivated skepticism. *Journal of Personality and Social Psychology, 63*(4), 568-584. Doi: 10.1037/0022-3514.63.4.568.

Doherty, M. E., Mynatt, C. R., Tweney, R. D., & Schiavo, M. D. (1979). Pseudodiagnosticity. *Acta Psychologica, 43*(2), 111–121.

Edelsbrunner, P. A., & Thurn, C. (2020, April 22). *Improving the utility of non-significant results for educational research*. PsyArXiv. https://doi.org/10.31234/osf.io/j93a2

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PloS One, 5*(4), e10068.

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics, 90*, 891-904.

Faria, R. (2018). *Research misconduct as white-collar crime.* Palgrave Macmillan.

Farrar, B. G., Altschul, D. M., Fischer, J., van der Mescht, J., Placì, S., Troisi, C. A., Vernouillet, A., Clayton, N. S., & Ostojić, L. (2020). Trialling meta-research in comparative cognition: Claims and statistical inference in animal physical cognition. *Animal Behavior and Cognition*, *7*(3), 419–444.

Ferguson, C.J., & Heene, M. (2012). The vast graveyard of undead theories: publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science, 7*(6), 555-561.

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), p1502-1505.

Fischhoff, B., & Beyth-Marom, R. (1983). Hypothesis evaluation from a Bayesian perspective. *Psychological Review, 90*, 239-260

Gilovich, T. (1983). Biased evaluation and persistence in gambling. *Journal of Personality and Social Psychology, 44*(6), 1110–1126.

Haeffel, G.J. (2022). Psychology needs to get tired of winning. *Royal Society Open Science, 9*(6), 1-8.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLOS Biology, 13*(3), e1002106.

Heene, M. & Ferguson, C.J. (2017). Psychological science's aversion to the null, and why many of the things you think are true, aren't. In S.O. Lilienfeld & I.D. Waldman (Eds.) *Psychological science under scrutiny: Recent challenges and proposed solutions*. Wiley-Blackwell, pp 34–52.

Hergovich, A., Schott, R. & Burger, C. (2010). Biased evaluation of abstracts depending on topic and conclusion: Further evidence of a confirmation bias within scientific psychology. *Current Psychology, 29*, 188–209.

Howard, J. (2018). *Cognitive errors and diagnostic mistakes: A case-based guide to critical thinking in medicine*. Springer.

Hudachek, L. & Quigley-McBride, A. (2022). Juror perceptions of opposing expert forensic psychologists. *Psychology, Public Policy, and Law, 28*(2), 213-225. Doi: 10.1037/law0000334.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PloS Medicine, 2*, e124. Doi: 10.1371/journal.pmed.0020124

Ioannidis, J. P. (2008). Why most discovered true associations are inflated. *Epidemiology, 19*(5), 640-8. Doi: 10.1097/EDE.0b013e31818131e7.

John, L.K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science, 23*(5), 524-532.

Journal of Trial and Error (n.d.). https://journal.trialanderror.org/

Journal of Articles in Support of the Null Hypothesis (n.d.) *About JASNH.* https://www.jasnh.com/about.html

Jureidini, J. & McHenry, L.B. (2020). *The illusion of evidence based medicine: Exposing the crisis of credibility in clinical research.* Wakefield Press.

Kassin, S. M. , Dror, I. E. & Kukucka, J. (2013). The forensic confirmation bias. *Journal of Applied Research in Memory and Cognition, 2*(1), 42-52. Doi: 10.1016/j.jarmac.2013.01.001.

Kiyonaga, A., & Scimeca, J.M. (2019) Practical considerations for navigating registered reports. *Trends in Neuroscience, 42*(9), 568-572.

Klayman, J. & Ha, Y.W. (1987). Confirmation, disconfirmation, and information in hypothesis testing. *Psychological Review, 94*(2), 211-228.

Klein, R.A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Alper, S., Aveyard, M., Axt, J. R., Babalola, M. T., Bahník, S., Batra, R., Berkics, M., Bernstein, M. J., Berry, D. R., Bialobrzeska, O., Binan, E. D., Bocian, K., Brandt, M. J., … 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. Doi:10.1177/2515245918810225

Koehler, J. J. (1993). The influence of prior beliefs on scientific judgments of evidence quality. *Organizational Behavior and Human Decision Processes, 56*(1), 28-55.

Kuhberger, A., Fritz, A, & Scherndl, T., (2014). Publication bias in psychology: A diagnosis based on the correlation between effect size and sample size. *PLOS ONE, 9*(9), e105825.

Lehner, P. E., Adelman, L., Cheikes, B. A., & Brown, M. J. (2008). Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans, 38*(3), 584-592. Doi: 10.1109/TSMCA.2008.918634.

Lilienfeld, S.O., Lynn, S.J., Ruscio, J., & Beyerstein, B.L. (2011). *50 great myths of popular psychology: Shattering widespread misconceptions about human behavior*. Wiley-Blackwell.

Ling, R. (2020) Confirmation bias in the era of mobile news consumption: The social and psychological dimensions. *Digital Journalism*, 8(5), 596-604, DOI: 10.1080/21670811.2020.1766987

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*(11), 2098–2109. https://doi-org.abc.cardiff.ac.uk/10.1037/0022-3514.37.11.2098

Mahoney, M.J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*, 161–175.

Mathieu, S., Boutron, I., Moher, D., Altman, D.G., & Ravaud, P. (2009). Comparison of registered and published outcomes in randomized controlled trials. *JAMA, 302*(9), 977-984.

Meppelink, C. S., Smit, E. G., Fransen, M. L. & Diviani, N. (2019) "I was right about vaccination": Confirmation bias and health literacy in online health information seeking, *Journal of Health Communication, 24*(2), 129-140. DOI: 10.1080/10810730.2019.1583701

Mervis, J. (2014). Why null results rarely see the light of day: "File drawer" study proposes registry for unpublished social science data. *Science, 345*(6200), 992.

Modgil, S., Singh, R.K., Gupta, S., & Dennehy, D. (2021). A confirmation bias view on social media induced polarisation during Covid-19. *Information Systems Frontiers*.

Mendel, R., Traut-Mattausch, E., Jonas, E., Leucht, S., Kane, J., Maino, K., Kissling, W., & Hamann, J. (2011). Confirmation bias: Why psychiatrists stick to wrong preliminary diagnoses. *Psychological Medicine, 41*(12), 2651-2659. Doi:10.1017/S0033291711000808

Menon, V. & Muraleedharan, A. (2016). Salami slicing of data sets: What the young researcher needs to know. *Indian Journal of Psychological Medicine, 38*(6), 577-578. Doi: 10.4103/0253-7176.194906.

Munafò, M., Nosek, B., Bishop, D., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021.

Mynatt, C. R., Doherty, M. E., & Tweney, R. D. (1977). Confirmation bias in a simulated research environment: An experimental study of scientific inference. *Quarterly Journal of Experimental Psychology, 29*(1), 85–95.

Nickerson, R.S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175-220.

Nosek, B., Spies, J.R. & Matyl, M. (2012). Scientific utopia II: restructuring incentives and practices to promote truth over publishability. *Perspectives in Psychological Science, 7*(6), 615-631.

Obels, P., Lakens, D., Coles, N. A., Gottfried, J., Green, S. A. (2020). Analysis of open data and computational reproducibility in Registered Reports in psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 229-237. Doi:10.1177/2515245920918872

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

Pennington, N., & Hastie, R. (1993). The story model for juror decision making. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 192–221). Cambridge University Press.

Pines, J.M. (2006). Profiles in patient safety: Confirmation bias in emergency medicine. *Academic Emergency Medicine, 13*(1), 90-94.

Prakash S., Bihari S., Need P., Sprick C., & Schuwirth L. (2017). Immersive high fidelity simulation of critically ill patients to study cognitive errors: a pilot study. *BMC Medical Education, 17*, 36.

Pyszczynski, T. and Greenberg, J. (1987) Toward an integration of cognitive and motivational perspectives on social inference: A biased hypothesis testing model. *Advances in Experimental Social Psychology, 20,* 297-340. https://doi.org/10.1016/S0065-2601(08)60417-7

Ramagopalan, S., Skingsley, A.P., Handunnetthi, L., Klingel, M., Magnus, D., Pakpoor, J., & Goldacre, B. (2014). Prevalence of primary outcome change in clinical trials registered on ClinicalTrials.gov: a cross-sectional study. *F1000Research, 3*, 77.

Rawat, S. & Meena, S. (2014). Publish or perish: where are heading? *Journal of Research in Medical Sciences, 19*(2), 87-89.

Reis, D. & Friese, M. (2022). The myriad forms of p-hacking. In W. O'Donohue, A. Masuda, & S. Lilienfeld (Eds.) *Avoiding questionable research practices in applied psychology* (pp 101-121). Springer.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638–641. https://doi-org.abc.cardiff.ac.uk/10.1037/0033-2909.86.3.638

Ross, L., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*(5), 880–892.

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology, 21*, 308-320. Doi: 10.1037/gpr0000128

Scheel, A. M., Tiokhin, L., Isager, P. M., & Lakens, D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4), 744–755. https://doi-org.abc.cardiff.ac.uk/10.1177/1745691620966795

Scheel, A.M., Schijen, M.R.M.J., & Lakens, D. (2021). An excess of positive results: comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development, 31*(1), e2295.

Schwarcz, J. (2019). *A grain of salt: The science and pseudoscience of what we eat*. ECW Press.

Schlosser, M. D., Jennifer K. Robbennolt, Daniel M. Blumberg, & Konstantinos Papazoglou. (2021). Confirmation bias: A barrier to community policing. *Journal of Community Safety and Well-Being, 6*(4), 162-167. https://doi.org/10.35502/jcswb.219

Sijtsma, K. (2016). Playing with data – or how to discourage questionable research practices and stimulate researchers to do things right. *Psychometrika, 81*(1), 1-15. Doi: 10.1007/s11336-015-9446-0.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, S., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L. Evans, M. H., Flores Cervantes, J., … Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337-356. Doi:10.1177/2515245917747646

Silverman, C. (2011, June 17). The Backfire Effect. https://archives.cjr.org/behind_the_news/the_backfire_effect.php

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.

Simonsohn, U., Nelson, L.D., & Simmons, J.P. (2014). P-curve: a key to the file drawer. *Journal of Educational Psychology: General, 143*(2), 534-547.

Singal, J. (2021). *The quick fix: Why fad psychology can't cure our social ills*. Farrar, Straus and Giroux.

Snyder, M., & Cantor, N. (1979). Testing hypotheses about other people: The use of historical knowledge. *Journal of Experimental Social Psychology, 15*(4), 330–342.

Soderberg, C.K., Errington, T.M., Schiavone, S.R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour, 5,* 990–997. https://doi-org.abc.cardiff.ac.uk/10.1038/s41562-021-01142-4

Stefan, A. M. & Schonbrodt, F. D. (2023) Big little lies: a compendium and simulation of p-hacking strategies. *Royal Society Open Science, 10*(2), Article 220346. https://doi.org/10.1098/rsos.220346

Taber, C.S. & Lodge, M. (2006). Motivated Skepticism in the Evaluation of Political Beliefs. *American Journal of Political Science, 50*(3), 755-769

Talluri, B.C., Urai, A.E., Tsetsos, K.,Usher, M. & Donner, T.H. (2018). Confirmation bias through selective overweighting of choice-consistent evidence. *Current Biology, 28*(19), 3128-3135.e8

Tschan, F., Semmer, N. K., Gurtner, A., Bizzari, L., Spychiger, M., Breuer, M., & Marsch, S. U. (2009). Explicit reasoning, confirmation bias, and illusory transactive memory: A simulation study of group medical decision making. *Small Group Research, 40*(3), 271–300.

Van den Akker, O. (2021). Selective hypothesis reporting in psychology [conference presentation]. 9th BITSS Annual Meeting [virtual conference]. Retrieved from https://osf.io/c6rnb/

van Erkel, P.F.A. & Thijssen, P. (2016). The first one wins: Distilling the primacy effect. *Electoral Studies, 44*, 245-254. https://doi-org.abc.cardiff.ac.uk/10.1016/j.electstud.2016.09.002

Vasilev, M.R. (2013). Negative results in European psychology journals. *Europe's Journal of Psychology, 9*(4), 717-730.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Maas, H. L. J. V. D., and Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science,* 7, 632–638.

Wagge, J. R., Baciu, C., Banas, K., Nadler, J. T., Schwarz, S., Weisberg, Y., IJzerman, H., Legate, N., Grahe, J. (2019). A demonstration of the Collaborative Replication and Education Project: Replication attempts of the red-romance effect. *Collabra: Psychology*, *5*(1), 5. Doi: https://doi.org/10.1525/collabra.177

Westen, D., Blagov, P. S., Harenski, K., Kilts, C., & Hamann, S. (2006), Neural bases of motivated reasoning: An fMRI study of emotional constraints on partisan political judgment in the 2004 U.S. Presidential election, *Journal of Cognitive Neuroscience, 18*(11), 1947–1958

Wicherts, J.M., Veldkamp, C.L.S., Augusteijn, H.E.M., Bakker, M., van Aert, R.C.M. & van Assen, M.A.L. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: a checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832.

Wolf, F. M., Gruppen, L. D., & Billi, J. E. (1985). Differential diagnosis and the competing-hypotheses heuristic. A practical approach to judgment under uncertainty and Bayesian probability. *JAMA, 253*(19), 2858-62.

Wolf, F. M., Gruppen, L. D., & Billi, J. E. (1988). Use of the competing-hypotheses heuristic to reduce 'pseudodiagnosticity'. *Journal of Medical Education, 63*(7), 548-54. Doi: 10.1097/00001888-198807000-00006.

Workman, M. (2018). An empirical study of social media exchanges about a controversial topic: Confirmation bias and participant characteristics. *The Journal of Social Media in Society, 7*(1), 381-400.

Zhao, H., Fu, S., & Chen, X. (2020). Promoting users' intention to share online health articles on social media: The role of confirmation bias. *Information Processing & Management, 57*(6),102354.

Chapter 4: Comparative Analysis of Open Research Practices

4.1. Introduction

As outlined in Chapter 3, the potential for questionable research practices undermines the credibility of many published research findings. One proposed solution to questionable practices is the sharing of data and analysis code to enable re-analysis by other researchers who can validate the findings and detect selective reporting practices such as p-hacking. Furthermore, sharing other research materials such as stimuli and experimental code has been suggested in order to facilitate understanding of study procedures and aid in the replication of the original studies, as well as reducing researchers' workload by enabling re-use of existing materials for their own purposes. Finally, the sharing of preregistered protocols has been suggested as being important for bias control, revealing any deviations from the pre-planned approach, such as altered hypotheses or changes in the methods or analyses. This chapter will discuss the sharing of data, code, materials, and protocols, and why this is important for improving the transparency and quality of research. A comparative study of the availability of protocols, data, code, and materials in RRs and SRs will then be reported and discussed.

4.1.1. Open Data

Sharing and public archiving of research data has been proposed as an important solution for addressing reporting bias by allowing others to check the validity of conclusions drawn from the analysis of that data (Wicherts et al., 2011). Reanalysis can reveal questionable practices such as p-hacking and other forms of selective reporting, as well as allowing genuine mistakes to be identified. It has also been reported that the public's trust in scientists is higher when their data is made available (Pew Research Center, 2019). As a result of these potential benefits, many have called for open data policies at scientific journals (e.g., Wicherts et al., 2011; Firebaugh et al., 2007).

Furthermore, sharing data openly allows it to be re-used by others either in meta-analysis or in other forms of secondary data analysis which may explore different research questions or use analytical approaches not considered by the original researchers (Vickers, 2006). This reduces research waste and can be more efficient than other researchers having to collect their own data, as demonstrated during the COVID-19 pandemic (Acosta-Velasquez et al., 2022; Rotulo et al., 2022; Gardner et al., 2020).

Such benefits of openly sharing research data may explain why articles with open data are associated with higher rates of citations by other researchers (Piwowar & Vision, 2013; Piwowar et al., 2007). There is also increasing demand for open data from research funders (Global Innovation Fund, 2021; Economic and Social Research Council, 2021) and from scientific journals, with mandatory open data policies becoming more common (Hardwicke et al., 2018; Stodden et al., 2013). Such policies are associated with increased rates of data sharing (Kidwell et al., 2016; Hardwicke et al., 2018; Neve & Rousselet, 2021; Nuitjen et al., 2017), although it was revealed that data was not always provided in a reusable form, was often not accompanied by adequate documentation, or that their findings were not fully reproducible, suggesting that there is still work to be done on improving the way in which data is curated and shared (Kidwell et al., 2016; Hardwicke et al., 2018: Neve & Rousselet, 2021; Nuitjen et al., 2017).

Meanwhile, a variety of other initiatives have also been introduced in an effort to promote data sharing. These include open data badges which have been introduced at some journals in an effort to promote greater sharing practices (Kidwell et al., 2016; Center for Open Science, n.d.a), although the effectiveness of these has been questioned (Cruwell et al., 2022). Furthermore, the opportunity to share data reports and even publish in journals specifically focused on open datasets (e.g., The Journal of Open Psychology Data, n.d.; Geoscience Data Journal; Scientific Data, n.d.) provides another incentive for researchers to make data publicly available. The Transparency and Openness Promotion (TOP) guidelines (Center for Open Science, n.d.b) also encourage greater transparency and openness within journals from an editorial perspective, while the Peer Reviewers Openness Initiative (Morey et al., 2016) calls on peer reviewers to demand access to study data as a condition of their agreement to review a paper, unless a reason is given for why the data cannot be shared.

Despite the benefits and increasing demand for open data, rates of data availability are low in the research literature, both in other disciplines (Savage & Vickers, 2009; Stodden et al., 2018; Alsheikh-Ali et al., 2011) and in psychology (Wicherts et al., 2006; Towse et al., 2019; Martone et al., 2018). A key study from Wicherts et al. (2006) found that 73% of psychologists publishing in a selection of APA journals did not share their data for reanalysis. Savage and Vickers (2009) reported that only one of their requests for data from ten different authors was successful in obtaining the data, with many others not responding or refusing to share their data despite the open data policy of the journal they had published in. Although awareness of open data has grown in recent years, low rates of data sharing has also been

noted more recently by Stodden et al. (2018), and by Rowhani-Farid & Barnett (2016) who found rates of this at the British Medical Journal (BMJ) to be as low as 4.5% overall, although these were higher for clinical trials, at 24%.

Additionally, research by Nutu et al. (2019) showed data sharing rates as low as 2% in clinical psychology journals, while Towse et al. (2019) showed that data sharing across 15 psychology journals was as low as 4% overall, although this did vary substantially between journals in their sample. They also highlight that even when data were provided, it was very often incomplete with limited usability. Hardwicke and Ioannidis (2018b) showed that for highly-cited articles in psychology and psychiatry, between 2006 and 2011, 68% of the datasets were not available while many others were only available with restrictions. Likewise, when data was requested from authors by Vanpaemel et al. (2015), this was shared by only 38% of the authors. While this is higher than rates reported in the other studies cited, it is still far from being a universal standard. Although the exact proportions vary between studies, it is clear that rates of data sharing are low across disciplines and within psychology itself and that even when data is shared there is considerable room for improvement in the preparation of the data provided.

These low rates of data sharing may be partially explained by attitudes towards open data among researchers, including fear among authors of errors being exposed and the validity of their conclusions drawn into question. This may be a valid concern, as Bakker & Wicherts (2011) demonstrate high rates of statistical errors in the published psychology literature. However, the detection and correction of such errors is vital for correcting the scientific record and enabling progress. Interestingly, a study by Wicherts et al. in 2011 revealed that reluctance to share data was associated with more errors in the reporting of statistical results, and with weaker evidence, supporting the notion that open sharing of data may be associated with higher quality research and reporting. Alternatively, it may be that those researchers who manage and archive their data more diligently may also be more rigorous in their statistical analysis and reporting, and therefore such associations may be more reflective of the researchers themselves than of the fact that their data has been shared. In contrast, Nuijten et al. (2017) found that data sharing was not associated with inconsistencies in statistical reporting.

A lack of data sharing may also be driven by a lack of skills and training in adequately preparing and sharing the data (Chawinga & Zinn, 2019; Houtkoop et al., 2018), highlighting

the importance of ensuring that any initiatives pushing for mandatory open data also acknowledge the need for significant capacity-building and support in this area, particularly for early career researchers. The need for adequate tools and guidelines for data sharing have also been highlighted as essential in facilitating data sharing (Molloy, 2011). Similarly, ethical concerns have also been raised regarding the sharing of sensitive information and in ensuring anonymity of the data. While some data may indeed be too sensitive to share, de-identifying data appropriately where possible can help to overcome this barrier in many cases. For example, although an extensive review by Wicherts et al. (2022) did find evidence of identifiable participant information in many publicly available datasets, they also reported that most of the privacy risks detected could have been easily prevented, suggesting that training and skill development in how to properly de-identify data is needed. Ethical concerns about sensitive data also suggest the importance of systems for mediating access if necessary and for the central role of research ethics councils in addressing issues related to data sharing (Mahomed & Labuschaigne, 2023).

Although data sharing is not mandatory for RRs at most journals, many journals offering this format are likely to also encourage other open practices such as open data and code sharing. Early evidence indicates that rates of data and code sharing are higher in RRs than in the general literature and that the shared data is more re-usable. Specifically, a study by Obels et al. (2020) showed that of the 62 RRs they examined, 66% had data available and 60% had code available, although only 58% had both. Furthermore, only 58% of these articles were found to be computationally reproducible. While these rates of data and code sharing do appear to be much higher than in similar studies of standard research reports, there is still much room for improvement in this among RRs.

4.1.2. Open Analysis Code

Optimal use of open data depends on open-source analysis software and access to any associated analysis code. Analysis code can provide certainty and clarity on the exact approach taken, rather than relying on details specified in the manuscript, which may not be comprehensive. The provision of analysis scripts is also important in enabling other researchers to computationally reproduce the analysis and so verify the findings reported. Finally, having greater clarity about the analytical approaches taken may help to inform the methods for future replication studies.

Stodden et al. (2013) report that open data policies tend to lead to open code policies, providing an important step toward code transparency. However, they note that despite the journal Science's open data and code policy, only 12% of articles published there in 2011 and 2012 provided information to access both data and code without having to contact the author and that requests were often met with reluctance or confusion, with many authors seemingly unaware of the journal's policy. Requests to the authors for access to these records increased overall availability for only 44% of the articles, despite the journal's policy that they should be made available to all.

These low rates are especially concerning given the importance of analysis code for computational reproducibility. Laurinavichyute et al. (2022) compared papers before and after an open data policy was introduced at the Journal of Memory and Language and found that the presence of analysis code (which was provided for 63% of the articles that also had accessible data) was the strongest predictor of whether the published results could be reproduced, increasing successful reproducibility by 38%. Consequently, they recommend that analysis code should always be provided with the data unless there is a valid reason not to do so. Obels et al. (2020) report similar rates of code sharing in RRs as Laurinavichyute et al. (2022) report for the SRs in this study, as well as similar rates of reproducibility for papers that shared their code. However, because of the journal's new policy, the SRs in Laurinavichyute's work contained open data and code to a much greater than extent than is typical in the standard psychology-related literature.

4.1.3. Open Materials

Sharing of other research materials and resources also benefits the research community but has received less attention than the sharing of data and analysis code. Here, 'research materials' refers to resources or items used in the study that would be needed to reproduce the study's procedures and data acquisition methods (Center for Open Science, n.d. a; Bowman & Spence, 2020). This may, for example, consist of stimuli and code for running experiments, or questionnaires used in the study. It is generally recommended that research reports should contain sufficient detail about the approach that other researchers should be able to replicate the study's method based on this description. However, word limits at journals, and the difficulty of fully conveying all aspects of some experimental tasks in textual descriptions (e.g., content of videos or the appearance of all stimuli, etc.) encourage shorter descriptions than are likely to be needed for other researchers to successfully replicate that study's methods. Therefore, open provision of the materials used in the study may be

important in allowing peers to fully understand and evaluate the approach taken in the study, as well as being essential in enabling researchers to effectively replicate the methods used (Bowman & Spence, 2020). Furthermore, sharing of research materials improves efficiency and reduces research waste, as it saves other researchers from creating duplicate versions of the same tasks or tools, a process which can be time consuming and may lead to differences in the design of purportedly similar tasks. While sharing of materials may not always be possible due to legal or ethical restrictions, greater sharing of such materials is in keeping with a more open and transparent culture and is increasingly being recognised as a desirable standard where possible (Gilmore et al., 2018; Bowman et al., 2020; Center for Open Science, n.d.; Grahe, 2018).

Availability of materials, as well as other research items such as data and analysis code, are sometimes stated to be required by the journal, either publicly, or that they should be available on request (Science, n.d.; Nature, n.d.). Badges have also been created for articles containing open materials in an effort to promote this practice, similar to the badges previously mentioned for open data (Center for Open Science, n.d. a) and evidence suggests that this initiative has also been successful. Specifically, Kidwell et al. (2016) report that the use of badges resulted in approximately three times more sharing of materials and also suggest that, at baseline there was often only partial sharing of the materials, which appeared to improve following the introduction of the badges. Therefore, they suggest that the open materials badges may be useful in promoting more complete sharing of study materials as well as improving the overall sharing rates.

However, the general rates of materials sharing in psychology are unclear, both for the general research literature, and for Registered Reports. One part of the study reported in this chapter therefore seeks to examine the availability of digital, and non-digital materials in RRs and standard reports.

4.1.4. Protocol Availability

Finally, preregistering study plans offers many potential benefits including greater transparency and the potential for detecting and/or reducing reporting bias. For example, pre-specifying the hypotheses, methods and analysis plan in advance can help to clearly distinguish between confirmatory and exploratory analysis and provide a record of which hypotheses were determined before the study was conducted (Simmons et al., 2017; Logg & Dorison, 2021; Wagenmakers & Dutilh, 2016; Nosek et al., 2012). This may help in detecting

any changes to the reported hypotheses such as HARKing and may help to safeguard against p-hacking by demonstrating that a particular analysis strategy was pre-planned and not just the result of fishing for any significant findings in the data without an existing prediction. However, preregistration can only be expected to reduce such practices if researchers adhere to their preregistered plan and report the study honestly. Even so, the existence of publicly available protocols may provide some degree of accountability for researchers to be honest about their approach and may help to aid in detecting any instances where this is not the case. As mentioned in the introduction of chapter 3 (section 3.1), there has been evidence of preregistration not being a sufficient deterrent and of outcome switching and other practices still occurring, particularly in relation to clinical trials. Therefore, preregistration alone may not be sufficient to fully eliminate such practices. RRs on the other hand offer a more substantial incentive for researchers to comply with their prespecified protocol, so as not to jeopardise the in-principle acceptance of their manuscript.

Previous research has shown that protocol availability in RRs is sub-optimal, with improvements needed in both protocol transparency and in the standardisation of protocol registration (Hardwicke & Ioannidis, 2018a; Chambers & Mellor 2018). Although requiring public access to the accepted stage 1 protocol is generally recommended, many journals that adopt RRs do not require or enforce this public registration and availability of the accepted protocols, although efforts have been made to improve this situation (Chambers & Mellor 2018). As of 2018, Hardwicke and Ioannidis found that most of the RRs they studied had a publicly available protocol, although very few were considered to be fully registered, i.e., provided as a read-only time-stamped version. Furthermore, the quality of the available protocols could often have been improved at times. The current study will examine protocol availability in a slightly larger (and somewhat more recent) sample of RRs than that used in Hardwicke and Ioannidis's (2018a) study and will compare this against the SR sample.

4.1.5. Research Questions

An overview of the research questions investigated in this chapter, and their corresponding hypotheses, are outlined in Table 3 below.

Table 3

Overview of research questions and hypotheses for chapter 4

Research Questions	Hypotheses
RQ 1: Are there differences in data availability between RR and SRs?	H1a: RRs will be more likely than SRs to have shared their data. H1b: RRs will be more likely than SRs to have an explicit statement about data availability.
RQ 2: Are there differences in availability of analysis code between RRs and SRs?	H2a: RRs will be more likely than SRs to have shared their analysis code. H2b: RRs will be more likely than SRs to have an explicit statement about availability of their analysis code.
RQ 3: Where digital materials have been used, are RRs more likely than SRs to have made these digital materials available?	H3a: Where digital materials have been used, RRs will be more likely than SRs to have an explicit statement about availability of their digital materials. H3b: Where any digital materials have been used, RRs will be more likely than SRs to have made these digital materials available.
RQ 4: Where any non-digital original materials have been used, are RRs more likely than SRs to have made these materials available?	H4a: Where any non-digital original materials have been used, RRs will be more likely than SRs to have an explicit statement about the availability of these materials. H4b: Where any non-digital original materials have been used, RRs will be more likely than SRs to have made these materials available.
RQ 5: What proportion of RRs have a publicly available protocol, and how does this compare to the rate of protocol availability among the SRs?	No specific hypothesis was stated for RQ 5

4.2. Methods

4.2.1. Overview of Initial Coding Process

As described more broadly in the General Methods chapter, content analysis was used to gather the initial data, using a detailed coding protocol that was developed through an iterative process. Details about five overall characteristics were coded: availability of the preregistered protocol, availability of the data, availability of the analysis code, availability of the digital materials, and availability of the original non-digital materials. For protocol availability, this was coded as yes or no, and an accompanying link was also saved for the protocol location. For data availability, two pieces of information were coded: whether there was a statement about the availability of the data, and how available the data was. The latter characteristic had a range of possible options: publicly available, partially available, already available, gated, unclear, not available with justification, or not available without justification. For each of the other three characteristics (availability of analysis code, digital materials, and original non-digital materials), three pieces of information were coded: whether this type of item was used in the study, whether there was a statement about its availability, and how available it was, using the same range of response options as listed for the data availability characteristic. Each of these is outlined in some more detail in the following sections. In a second coding process, categorical variables were then created for each characteristic, based on the data gathered.

4.2.2. Coding of Protocol Availability

To determine protocol availability, the articles were checked to see if they contained a link to the protocol, or if the protocol was otherwise identifiable in any other way such as an ID number. Any links to repositories such as OSF were checked, as these sometimes contained the protocol along with data etc., even if they didn't always state this directly in the paper. Protocol availability was coded as either yes or no. If the authors stated that the protocol was available, but it was not actually accessible (e.g., the link provided didn't work), the protocol availability characteristic was coded as No. Based on the information gathered during this initial coding process, a categorical variable was then created for the protocol availability whereby 0 indicated 'no' and 1 indicated 'yes. The rate of protocol availability in RRs was compared against both the rate in SRs overall, and against the rate within just the SRs that actually contained any preregistered studies.

4.2.3. Coding & Analysis of Data Availability

4.2.3.1. Coding of Data Availability Statements

Initial data was gathered for whether authors provided an explicit statement about whether the data was available, and this was coded as either yes or no. It is important to note that this characteristic was for whether there was a statement about this, not about whether that statement was actually positive or negative regarding the availability of the data. The papers were checked for any statement or indication about whether the data was available, and for links to repositories such as OSF, as they often mentioned data availability when giving the OSF link. Very occasionally, a paper could state that data was available at a particular link but this was not actually available there when the linked repository was examined, in which case this was still coded as Yes for whether there was a statement but then coded as Unclear for the actual availability of data. A second round of coding was done to create variables for the analysis. Based on the data gathered during the initial round of coding, the presence of a data availability statement was coded as 1 while the absence of such a statement was coded as 0.

4.2.3.2. Coding of the Availability of the Data

For the initial coding of data availability, eight response options were used: publicly available, partially available, gated, already available, available on request, not available without justification, not available with justification, or unclear. Publicly available meant that the data, as far as I could tell, was fully available and accessible to the public without making website accounts or other gatekeeping processes to obtain the data, e.g., if it was freely available on OSF. In a small number of cases, the data was coded as 'gated'. In such cases, the data had technically been made available but those seeking to obtain it would need to register for a website account or undergo some other similar process to obtain it from a public source. In other cases, studies which made use of data that was already publicly available (e.g., secondary data analysis) would typically be coded as 'already available' since the authors may not have been at liberty to share the original data themselves, but it was often still available publicly from another source.

If there was any mention of data being available from the author, or about needing to contact the author to obtain it, this was coded as being 'available on request'. The response option 'Not available WITHOUT justification' was only used if the article explicitly stated that the data were not available, without providing any justification for why not. If any reason was given for why data was not available (e.g., not available due to confidentiality issues or if the

funders didn't permit data sharing, etc.), this would be coded as 'not available WITH justification'. Finally, if the paper made no mention of whether data was available, or the availability was otherwise unclear and didn't fit any of the other codes, this was coded as 'unclear'. If a link was given to the data but that link didn't work, and the data couldn't be found in any other way, this was also coded as 'unclear'.

Multiple categories had been used during the initial coding process, as just described. During a second round of coding, this data was condensed so that those initially coded as publicly available, gated, partially available, or already available were grouped together as being 'available' and coded as 1, while those initially coded as available on request, not available with justification, not available without justification, or unclear, were all grouped together as 'not available', and coded as 0.

The frequencies were examined for both the presence of availability statements, and for the availability of the data, and chi-square analysis was used to compare these characteristics between the article types.

4.2.4. Coding of Analysis Code Availability

4.2.4.1. Coding Whether Analysis Code Was Used

Articles were examined to determine whether authors appeared to have used analysis code in their study. This was coded as either yes, no, or unclear in order to inform the coding of the other analysis code characteristics. The use of analysis code would be coded as yes if there were specific mention of code, syntax, scripts, or software that uses code, such as R or Python. Any OSF links or other repositories were also checked in case the authors had provided the analysis code there, as this was not always mentioned explicitly in the paper even if it was actually available in the linked repository. Whether analysis code was used was coded as 'no' if there was any indication that code was not used for the analysis. If authors did not mention code being used or available and if they didn't specify the analysis software used, this characteristic was typically coded as unclear as it could not be determined whether code had been used. The use of 'unclear' was typically much more frequent than the use of 'no', given the lack of detail provided in many studies about the type of analysis software used.

4.2.4.2. Coding & Analysis of Analysis Code Availability Characteristics

Initial data was gathered for whether authors explicitly stated if the analysis code was available i.e., for the presence of a specific statement about its availability, regardless of

whether this indicated that the code was actually available. The presence of an analysis code availability statement was therefore coded as either yes, no, or N/A. 'Yes' was used if there were specific mentions of whether the analysis code was available, while 'no' was used if there was no such statement. 'N/A' was used if code had not been used for the analysis (as far as could be determined) and so this characteristic was non-applicable. In a second round of coding, in order to create data suitable for analysis, the variable for whether there was a statement about the availability of the analysis code was then coded as either 1 for yes or 0 for no.

Nine different response options were initially used to code for the availability of the analysis code. Eight of these were the same as outlined above in relation to the data availability, i.e., publicly available, partially available, gated, already available, available on request, not available without justification, not available with justification, or unclear. The additional response option, N/A, was used in cases where code had not been used for the analysis and so this characteristic was not applicable. In a second round of coding, a condensed variable was created by coding those with any kind of available analysis code as 1 and those for which it was not available as 0, as per the approach take for the creation of this variable for the data availability, i.e. Publicly available, partially available, gated, and already available were coded as 1, while available on request, not available with justification, not available without justification, and unclear, were coded as 0.

As per the approach taken to analyse the data availability variables, frequencies were examined and the characteristics were compared between the article types using chi-square analysis. This analysis was restricted to only the articles that had used (or appeared to have used) analysis code, i.e., those coded as N/A were excluded from the analysis.

4.2.4. Coding of Digital Materials Availability

4.2.4.1. Coding Whether Digital Materials Were Used

Articles were examined to determine if digital materials were used, and this information was then used to inform the coding of the other characteristics regarding the availability of digital materials. Digital materials could be very diverse but examples included code for running experimental tasks, stimuli or videos used in experimental tasks, online surveys, or any digital intervention used in the studies. Recordings and subsequent transcripts of interviews could be considered digital.

Response options for whether digital materials had been used were coded as either yes, no, or unclear. This characteristic was coded as 'yes' if there was a clear mention of digital materials being used, which was usually evident from the 'materials' or 'methods' section of the paper. 'No' was used if it was clear that no digital materials were used e.g., if the study only used in-person tasks or paper-based questionnaires. The characteristic was coded as 'Unclear' if it could not be determined whether materials were delivered digitally, although this was rarely needed because in the majority of cases this could be easily inferred.

4.2.4.2. Coding of Digital Materials Availability

Where digital materials had been used, data was gathered for whether authors provided an explicit statement of whether those digital materials were available, and this was coded as either yes, no, or N/A. 'Yes' was used when there as a specific mention of whether these materials were available, while 'No' was used if there was no statement provided about the availability of digital materials, even if materials were actually available in a repository, etc. Finally, N/A was used if digital materials had not been used and so this characteristic was non-applicable. As with the data and code availability characteristics, in a second round of coding, the variable for whether there was a specific statement about the availability of the digital materials was coded as either 1 for yes or 0 for no.

If the digital materials had been used, their availability was initially coded using the same range of response options specified previously for data and code availability (publicly available, partially available, gated, already available, available on request, not available without justification, not available with justification, unclear, and N/A). The criteria that are used to determine eligibility for 'open materials' badges (i.e., that the materials must have a persistent identifier and be provided in a format that is time-stamped and permanent; Center for Open Science, n.d. a) were not relied upon for this study and so this study's approach to this may constitute a slightly more flexible definition of availability. While many of the articles may have met this standard anyway, it was not considered necessary to meet both of the above criteria as long as the materials could be accessed by our coders based on the information provided in the article. In a second round of coding, a condensed variable was created by coding those with any kind of available digital materials as 1 and those for which these were not available as 0, as per the approach take for the creation of this variable for the data and code availability characteristics. As per the approach taken to analyse the data and code availability variables, frequencies were examined and the characteristics were compared

between the article types using chi-square analysis. This analysis was restricted to only the articles that had used (or appeared to have used) digital materials.

4.2.5. Coding of Original Non-Digital Materials Availability

4.2.5.1. Coding Whether Non-Digital Materials Were Used

As some studies could use and share non-digital materials, this was also accounted for in the initial round of coding. Non-digital materials in this case refer to any materials such as paper-based questionnaires, test booklets, or other non-digitally administered items. Additionally, however, authors may not be at liberty to share some paper-based materials that they used but did not themselves create (such as diagnostic materials, validated paper-based questionnaires created by others, etc), and so the coding of these non-digital materials was restricted where possible to only original materials created by the authors themselves, as far as this could be determined or inferred.

Articles were examined to determine if original non-digital materials appeared to have been used, and this information was then used to inform the coding of the other characteristics regarding the availability of these materials. Response options for whether original non-digital materials had been used were coded as either yes, no, or unclear, as per the approach previously described for the coding of the digital materials. This characteristic was coded as 'yes' if there was a clear mention of such materials being used, which was usually evident from the 'materials' or 'methods' section of the paper. 'No' was used if it was clear that no such materials were used e.g., if the study only used digital materials which was usually the case. The characteristic was coded as 'Unclear' if it could not be determined whether or not materials were delivered digitally. Overall, this characteristic (original non-digital materials) was rarely applicable but it still seemed worthwhile to document any instances of this, to complement the focus on the digital materials.

4.2.5.2. Coding & Analysis of Original Non-Digital Materials Availability

Where original non-digital materials had been used, data was gathered for whether authors provided an explicit statement of whether those materials were available, and this was coded as either yes, no, or N/A, as per the approach previous described for the coding of this characteristic for the digital materials. In the second round of coding, the variable for whether there was a statement about the availability of the non-digital materials was coded as either 1 for yes or 0 for no.

If the non-digital materials had been used, their availability was coded using the same range of response options specified previously for code and digital materials availability, as previously described in those sections (i.e., publicly available, partially available, gated, already available, available on request, not available without justification, not available with justification, unclear, and N/A). In the second round of coding, a condensed variable was created by coding those with any kind of available non-digital materials as 1 and those for which these were not available as 0, as per the approach taken for the creation of variables for the other availability characteristics.

As per the approach taken to analyse the other availability variables, frequencies were examined and the characteristics were compared between the article types using chi-square analysis. This analysis was restricted to only the articles that had used (or appeared to have used) original non-digital materials.

4.3. Results

4.3.1. Comparative Analysis of Protocol Availability

As previously stated, the rate of protocol availability in RRs was compared against both the rate in SRs overall, and the rate within just the SRs that contained any preregistered studies as this was a much fairer comparison. Across the total sample (N = 510, i.e., 170 RRs and 340 SRs), RRs predictably had a much higher rates of protocol availability than SRs did (75.88% vs. 9.71%), and the difference between the article types in this characteristic was highly significant: $x^2(1, N = 510) = 228.99$, $p < 0.001$. When the comparison was restricted to the SRs that had actually contained preregistered studies (n = 36), the opposite pattern was shown, with SRs having a higher rate or protocol availability than RRs (91.67% vs. 75.88%). This difference was statistically significant: $x^2(1, N = 206) = 4.41$, $p = 0.04$.

4.3.2. Comparative Analysis of Data Availability

The analysis of data availability was conducted on the full comparative sample (N = 510; 170 RRs and 340 SRs). Examination of the frequencies showed that 74.12% of RRs had a data availability statement, compared with only 43.53% of SRs. Chi-square analysis showed a statistically significant difference between the article types in inclusion of such statements: $x^2(1, N = 510) = 42.65$, $p < .001$. When the actual availability of the data was examined, RRs were found to have much higher rates of data sharing compared with SRs (70% vs. 30.29%). Chi-square analysis confirmed that there was a statistically significant difference in this between the article types: $x^2(1, N = 510) = 72.69$, $p < 0.001$

4.3.3. Comparative Analysis of Code Availability

Use of analysis code appeared to be more common in the RRs compared with the SRs, as half of RRs (50.59%) used this, compared with only 30% of SRs. The analysis of the code availability characteristics was restricted to those articles that had actually used (or appeared to have used) analysis code and so this resulted in the inclusion of 188 articles in total (n = 86 RRs and 102 SRs). There was a marked difference in how many articles included a specific statement about the availability of the analysis code. Specifically, 82.56% of RRs had a statement about this, compared with only 49.02% of the SRs. A chi-square analysis showed that this difference was statistically significant: $x^2(1, N = 188) = 22.88$, $p < .001$

Of the articles that had used analysis code, a total of 89.54% of the RRs had this available, compared with only 48.04% of SRs. A chi-square comparison showed that the difference between RRs and SRs was statistically significant: $x^2(1, N = 188) = 36.35$, $p < .001$.

4.3.4. Comparative analysis of Digital Materials Availability

The analysis for the availability of digital materials was restricted to the articles that contained (or appeared to contain) digital materials. In total this consisted of 416 articles (n = 142 RRs and 274 SRs). Within this restricted sample, RRs were more likely to have a statement regarding the availability of their digital materials than SRs (57.75% vs. 33.58%). A chi square analysis showed a statistically significant difference: $x^2(1, N = 416) = 22.46$, $p < .001$. RRs were also more likely than SRs to have their digital materials available (56.34% vs. 29.41%) and a chi-square comparison showed that this difference was statistically significant: $x^2(1, N = 414) = 28.53$, $p < .001$

4.3.5. Comparative Analysis of Original Non-Digital Materials Availability

Very few articles appeared to have used original non-digital materials (n = 54 in total i.e., 24 RRs and 30 SRs). Of these articles, RRs had higher rates of availability statements in relation to their non-digital materials than SRs did (70.83% vs. 23.33%). A chi-square analysis of non-digital materials availability statements showed that the difference between RRs and SRs was statistically significant: $x^2(1, N = 54) = 12.18$, $p < .001$. RRs also had much higher rates of availability of their non-digital materials than SRs did (70.83% vs 23.33%). A chi-square analysis conducted on this variable also showed that the difference in rates of availability of non-digital materials between RRs and SRs was statistically significant: $x^2(1, N = 54) = 12.18$, $p < .001$.

Table 4

Statements regarding availability of research materials, across article types

Article Type	Data availability statement		Analysis code availability statement		Digital materials availability statement		Non-digital materials availability statement	
	n	%	*n*	%	*n*	%	*n*	%
Article Type								
Registered Report	126	74.12	71	82.56	82	57.75	17	70.83
Standard Report	148	43.53	50	49.02	92	33.58	7	23.33

Note. N varies per characteristic. For the data availability statement results, total N = 510. For the code availability statement results, N = 188. For the digital materials availability statement results, N = 416. For the non-digital materials statement results, N = 54. Overall, the results show much higher rates of availability statements in RRs than in SRs, for all types of research materials examined.

Table 5

Availability of research materials across article types

Article Type	Data available		Analysis code available		Digital materials available		Non-digital materials available	
	n	%	*n*	%	*n*	%	*n*	%
Article Type								
Registered Report	119	70.00	77	89.54	80	56.34	17	70.83
Standard Report	103	30.29	49	48.04	80	29.41	7	23.33

Note. N varies per characteristic. For the data availability statement results, total N = 510. For the code availability statement results, N = 188. For the digital materials availability statement results, N = 414. For the non-digital materials statement results, N = 54. Overall, the results show much higher rates of availability statements in RRs than in SRs, for all types of research materials examined.

4.4. Discussion

4.4.1. Recap of Findings

In summary, RRs were more likely to share data, code, and materials than SRs were, and were more likely to contain a specific statement about the availability of these items. Although RRs were more likely than SRs *overall* to have a protocol available, when compared with only preregistered SRs, the preregistered SRs were significantly more likely than the RRs to have an available protocol. This latter approach to the protocol availability analysis was considered to be the fairer comparison of this characteristic.

4.4.2. Discussion of Results and Comparison to Previous Findings

As expected, the current study confirmed that the rates of data, code and materials sharing were higher among RRs than SRs, and these were more likely to contain specific statements indicating whether these were available. The results of the current study show that 74% of the RRs had at least some data available, compared with 44% of the SRs. Estimates of the rates of data availability in SR vary widely across different studies, from 2% (Nutu et al., 2019) to 38% (Vanpaemel et al., 2015). Data availability for SRs in the current study was slightly higher than the rates found by Vanpaemel et al. (2015), and by Hardwicke & Ioannidis (2018) who found that a total of 32% of the studies they examined provide some access to data, either with or without restrictions. The findings of this current study in relation to the rates of data sharing in SRs is therefore broadly consistent with the higher estimates of data availability in the standard literature in psychology. However, as the journals examined in the current study could all be considered to be supportive of open science practices to some degree (due to their endorsement of the RR format), this could reasonably lead us to expect that they might have had higher levels of support for open practices more broadly, including data sharing. Efforts to increase the sharing of data by standard reports could therefore be an important future direction for these journals and so other initiatives such as the TOP guidelines and badges for open practices may be important in promoting this.

Rates of data sharing in RRs was only slightly higher in the current study than that reported by Obels (70% vs. approx. 66%), suggesting that the larger sample and the inclusion of more recent RRs in the current study's sample has not had much additional impact on these rates so far. However, the rates of code sharing in RRs in the current study (89.54%) were much higher than that found by Obels (approx.. 60%). This may be explained by the sample including some more recent papers, as uptake may have become more common in recent years, or perhaps journal's procedures for checking compliance with code sharing policies

may be upheld more rigorously. However, if this was the case it would be reasonable to expect this to also influence the rates of data sharing to a greater extent.

The rate of code availability in this set of RRs was also considerably higher than that found for standard reports by Laurinavichyute et al. (2022) who reported this in 63% of cases. However, the rate of code sharing in the current study's sample of SRs was somewhat lower than in the SRs examined by Laurinavichyute's study. This difference may be due to their focus being on a single journal rather than a broad range of different journals, as well as their focus being on a journal with an open data policy, whereas this was not the case for many of the journals that supplied the current study's SRs. This proliferation of open data in Laurinavichyute's study may therefore also lead to higher rates of open code to accompany this, giving a higher than usual estimate of the rates of code sharing. This suggests that open data policies may have a knock-on effect on code availability, and that open data and code policies have an important place in journals that support open practices, to support and complement the RR format and to raise the general standard of the articles published there regardless of the format used.

The current study offers promising evidence that the rates of data and code sharing are considerably higher in RRs than in SRs, and that this availability is more clearly indicated in the RRs than in the SRs. However, the presence of specific statements regarding the availability of the data was lower for both RRs and SRs compared with the rate for SRs reported by Alsheikh-Ali et al. (2011), who found these statements in 88% of journals across disciplines. The fact that this was so much higher in their sample of standard reports, even higher than in the current sample of RRs, is surprising.

Due to the limited attention in the literature regarding the sharing of study materials other than data and code, there was limited other work to compare the current study to and so this is an area where more research would be beneficial in order to understand materials sharing practices, barriers, and areas for improvement. While Kidwell et al., (2016) have suggested that rates of open materials were higher than rates of open data, the opposite pattern was found in the current study, with only 56% of RRs sharing their digital materials whereas 70% shared their data. SRs had very similar proportions of data and materials sharing, suggesting that the pattern suggested by Kidwell et al. also did not apply to them. The RRs did show much higher rates of sharing of original non-digital materials (71%) than of their digital materials (56%), but this was a very small proportion of the overall sample, while studies

using digital materials were much more common. Also, as the non-digital materials were only considered if they appeared to be original and therefore possible for the authors to share, it is reasonable that a higher proportion of these may have been shared than among the digital materials where this distinction was not made and so more authors may not have been able to share their digital materials openly. In hindsight, this may have been a valuable consideration to add into the coding of the digital materials as well, although the challenges of identifying whether all of the digital items used were original or within the authors' rights to share may not have been feasible. In any case, the relatively low rates of digital materials sharing indicates an important area for further improvement, perhaps through the use of open materials badges or other incentives for sharing of materials, and the inclusion of specific availability statements in the articles regarding the materials to indicate whether these are available and if not, to provide a justification for not sharing them.

Finally, the rates of protocol availability in the current study were quite similar to the rates found by Hardwicke and Ioannidis (2018a), indicating that most RRs do appear to be associated with a publicly available stage 1 protocol although there is room for improvement in this and an ideal situation would involve all accepted stage 1 protocols being made available, even if they need to be temporarily embargoed while the study is being conducted. Like Hardwicke and Ioannidis, the protocols found in the current study were not always fully registered appropriately, and improved standardisation may be beneficial in ensuring consistent quality and completeness of the available protocols. However, some degree of flexibility may still be necessary due to the diversity of designs and study topics that may be used in RRs, as a rigid one-size-fits-all approach may pose challenges for disciplines and designs that are currently less well represented within RRs.

While the preregistered SRs in the current study had higher rates of protocol availability than RRs, this pattern of results is unsurprising. If the SRs preregistered their protocols it stands to reason that their protocol should be available, while any that were not actually available in this were likely to be those small number of cases in which the link to a protocol was broken, missing, or could not be reached in some way. In contrast, although all RRs must have had protocols in order to submit them to the reviewing journals, most journals did not (and many still do not) require those protocols to be made publicly available. However, calls for greater protocol availability have been made due to the need for this greater transparency and openness, as well as the potential usefulness of having access to accepted stage 1 protocols

for researchers interested in also submitting an RR to a particular journal; such examples of previous stage 1 protocols may be an important support for them in preparing this.

One key difference to much of the other literature on this topic area, is the more limited scope of the current work, as it did not seek to examine the usability of the shared items or to use them to determine the studies' reproducibility, unlike in many of the other studies conducted in this area (e.g., Obels et al., 2020). However, this limited scope was intentional due to the practicalities of conducting such a study with limited resources and within the available time frame. This does however constitute an important area for future research, because the quality and usability of the items being shared is a vital consideration to enable reproducible research and is therefore an important consideration to account for when writing and implementing policies for open practices.

4.4.3. Limitations

As the current study focused only on the actual availability of the research items e.g., whether the data was available, and not on broader aspects such as how re-usable (or reproducible) this actually was, this could be considered a limitation of the current study, because the sharing of these items alone is not a complete picture of how useable or reproducible they are. As previously mentioned, however, this was not considered a feasible addition to the current project and so this provides an opportunity for future studies to build on the current work and investigate this in more detail.

One further limitation of the approach used in this study is the compiling of all types of availability into a very simple categorical variable, thereby losing some of the nuanced detail provided in the initial coding process. This could be improved by also examining the rates of 'publicly available' data/code/materials on their own as more stringent metric of their availability.

The non-digital materials characteristic was restricted to original non-digital materials only to allow for materials that the researchers may not have had the right to share e.g., due to copyright issues. Taking a similar approach to the digital materials may have been beneficial in order to also provide this kind of nuanced consideration of the researchers' rights to share their digital materials. However, determining this may have been prohibitively difficult.

4.4.4. Implications and Future Directions

Due to the limited scope of the current study in comparison to other, similar studies of open data rates in journals, the implications and recommendations are more limited. However, it

would be beneficial for mandatory open data and code policies to be implemented in journals more widely, albeit with some flexibility for cases of highly sensitive data that cannot be anonymised or in cases where the researchers are not permitted to share their data or materials whether due to copyright issues or funder regulations. However, other sources have shown that some of the most applied disciplines such as clinical psychology (Nutu et al., 2019) and medicine (Rowhani-Farid & Barnett, 2016) also have the lowest rates of data availability and while much of this may be attributable to the challenges of de-identifying sensitive data and ensuring adequate permissions for sharing, this is an area where the reliability and trustworthiness of the findings is of paramount importance. Increased efforts to promote data sharing and train researchers in these areas in the skills necessary to fully anonymise and share sensitive data securely, could be a particularly worthwhile endeavour in terms of the potential benefits on clinical practice. Furthermore, due to the difficulties of accessing many clinical populations, the sharing of anonymised datasets in clinical areas may be particularly useful for secondary data analysis and enabling further progress in this area.

Increased sharing of the data, code and materials are of limited use if they cannot be re-used effectively and so detailed guidelines for sharing these items should be provided by journals at the point of submission. These should include specific instructions for how to format and prepare these for re-use and should clarify what types of metadata and other information are required to accompany the items shared. Checking of any shared data, code and materials by editors and reviewers, while time consuming, could be an important quality check within journals to ensure that the data submitted is both re-usable and ideally, reproducible. This does however require a huge commitment of time and effort from the journals, but the potential benefits to the scientific record could be extensive.

Open materials should be encouraged by journals wherever possible. The lack of attention to this in the literature, as well as the relatively low rates of shared digital materials in the current study suggests that this is an important area for further development.

For RRs, efforts should be made to ensure that all stage 1 protocols are fully registered and made public, as well as ensuring that high standards are maintained for the level of detail and completeness of the archived protocols. This is important in ensuring transparency and consistency for these documents.

Rates of availability statements for all items were lower than ideal, and so increased use of these should be encouraged and/or made mandatory within journals. Even if the items in

question are not actually available, having a clear statement regarding this should be easy to provide and would be very helpful in clarifying the availability, particularly if this is accompanied by some justification or explanation in cases where the items are said to not be available. Morey et al. (2016) suggest a similar requirement in their proposal of the Peer Reviewers' Openness Initiative, along with several other standards to be upheld for the openness of data and materials. Such clear statements about availability are also a requirement for level 1 adoption of the TOP Guidelines, while higher levels of this require sharing of the data, code or materials themselves.

4.5 Conclusion

In conclusion, RRs appear to be associated with higher rates of availability of data, code and materials, compared with SRs. RRs also have relatively high rates of protocol availability. Rates of these characteristics in RRs are generally comparable with rates reported in previous literature (where available) and are typically higher than the rates reported for SRs in previous studies. Although the higher rates of availability in RRs are encouraging, there is still considerable room for improvement in terms of increased rates of sharing of such items, as well as the need for complete and fully usable formatting of these items to enable reproducibility and re-use. Work should continue in order to promote and improve on these practices and help them to become standard practice.

4.6 References

Acosta-Velasquez, R., Fajardo-Moreno, W., Espinosa-Leal, L. (2022) *Predicting intensive care unit admission of COVID-19 patients with open data: analysis of the first wave in Colombia.* Preprints. https://www.preprints.org/manuscript/202212.0330/v1

Alsheikh-Ali, A. A., Qureshi, W., Al-Mallah, M. H., Ioannidis, J. P. A. (2011) Public availability of published research data in high-impact journals. *PLoS ONE 6*(9), e24357. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0024357

Bakker, M., & Wicherts, J. M. (2011) The (mis)reporting of statistical results in psychology. *Behavior Research Methods, 43*, 666–678.

Bartlett, J. (n.d.) Sharing Research Materials Online: Why You Should and How You Can. *Gorilla*. https://gorilla.sc/sharing-research-materials-online-why-you-should-and-how-you-can/ (Accessed 20th December 2022)

Bowman, N. D. & Spence, P. R. (2020). Challenges and best practices associated with sharing research materials and research data for communication scholars. *Communication Studies, 71*(4), 708-716, DOI: 10.1080/10510974.2020.1799488

Chawinga, W. D. & Zinn, S. (2019). Global perspectives of research data sharing: A systematic literature review. *Library & Information Science Research, 41*(2), 109-122. https://doi.org/10.1016/j.lisr.2019.04.004

Chambers, C. D. & Mellor, D. T. (2018). Protocol transparency is vital for registered reports. *Nature Human Behaviour, 2*, 791-792.

Center for Open Science (n.d., a). 1. View the Badges. https://osf.io/tvyxz/wiki/1.%20View%20the%20Badges/ (Accessed 12th January 2023).

Center for Open Science (n.d., b). The TOP Guidelines were created by journals, funders, and societies to align scientific ideals with practices. https://www.cos.io/initiatives/top-guidelines

Cruwell, S., Apthorp, D., Baker, B.J., Colling, L., Elson, M., Geiger, S.J., Lobentanzer, S., Monéger, J., Patterson, A., Schwarzkopf, D.S., Zaneva, M., & Brown, N.J.L. (2022). *What's in a badge? A computational reproducibility investigation of the open data badge policy in one issue of Psychological Science.* PsyArXiv. https://psyarxiv.com/729qt/ (Accessed 30th January 2023).

Economic and Social Research Council (2021). ESRC Research Data Policy. *Economic and Social Research Council*. Available from: https://www.ukri.org/wp-content/uploads/2021/07/ESRC-200721-ResearchDataPolicy.pdf (Accessed 30th January 2023).

Firebaugh, G. (2007) Replication data sets and favored-hypothesis bias. *Sociological Methods & Research, 36*, 200–209.

Funk, C., Hefferon, M., Kennedy, B., & Johnson, C. (2019, August 2). Americans say open access to data and independent review inspire more trust in research findings. Pew Research Centre. https://www.pewresearch.org/science/2019/08/02/americans-say-open-access-to-data-and-independent-review-inspire-more-trust-in-research-findings/ (Accessed 30th January 2023)

Gardner, L., Ratcliff, J., Dong, E., & Katz, A. (2020). A need for open public data standards and sharing in light of COVID-19. *The Lancet: Infectious Diseases, 21*(4), e80. https://www.thelancet.com/journals/laninf/article/PIIS1473-3099(20)30635-6/fulltext

Geoscience Data Journal (n.d.). https://rmets.onlinelibrary.wiley.com/journal/20496060

Gilmore, R. O., Kennedy, J. L., & Adolph, K. E. (2018). Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science, 1*(1), 121-130. Doi: 10.1177/2515245917746500

Global Innovation Fund (2021). Guidelines on research transparency and ethics in GIF-related impact Evaluation. *Global Innovation Fund*. Available from: https://www.globalinnovation.fund/wp-content/uploads/2021/12/Research-Transparency-and-Ethics-Guidelines-2021.pdf (Accessed 30th January 2023).

Grahe, J. (2018). Another step towards scientific transparency: requiring research materials for publication. *The Journal of Social Psychology, 158*(1), 1-6. https://www-tandfonline-com/doi/epdf/10.1080/00224545.2018.1416272

Hardwicke, T.E. & Ioannidis, J.P. (2018a). Populating the Data Ark: An attempt to retrieve, preserve, and liberate data from the most highly-cited psychology and psychiatry articles. *PLOS ONE, 13*(8), e0201856. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0201856

Hardwicke, T.E., & Ioannidis, J.P. (2018b). Mapping the universe of registered reports. *Nature Human Behaviour, 2*(11), 793-796. 10.1038/s41562-018-0444-y.

Hardwicke, T.E., Mathur, M.B., MacDonald, K., Nilsonne, G., Banks, G.C., Kidwell, M.C., Mohr, A.H., Clayton, E., Yoon, E.J., Tessler, M.H., Lenne, R.L., Altman, S., Long, B., & Frank, M.C. (2018). Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal Cognition. *Royal Society Open Science, 5*(8), 180448. http://dx.doi.org/10.1098/rsos.180448

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data sharing in psychology: A survey on barriers and preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70-85. Doi:10.1177/2515245917751886

Kidwell, M.C., Lazarević, L.B., Baranski, E., Hardwicke, T.E., Piechowski, S., Falkenberg, L.S., Kennett, C., Slowik A., Sonnleitner, C., Hess-Holden, C., Errington, T.M., Fiedler, S., & Nosek, B.A. (2016) Badges to acknowledge open practices: a simple, low-cost, effective method for increasing transparency. *PloS Biology, 14*(5), e1002456.

Logg, J. M., & Dorison, C. A. (2021). Pre-registration: Weighing costs and benefits for researchers. *Organizational Behavior and Human Decision Processes, 167*, 18-27.

Mahomed, S., & Labuschaigne, M. L. (2023). The evolving role of research ethics committees in the era of open data. *South African Journal of Bioethics and Law*, *15*(3), 80-83. https://doi.org/10.7196/SAJBL.2022.v15i3.822

Martone, M. E., Garcia-Castro, A., & VandenBos, G. R. (2018). Data sharing in psychology. *American Psychologist*, *73*(2), 111–125.

Molloy, J.C. (2011). The Open Knowledge Foundation: Open data means better science. *PloS Biology 9*(12), e1001195.

Morey R. D., Chambers C. D., Etchells P. J., Harris C. R., Hoekstra R., Lakens D., . . . Zwaan R. A. (2016). The Peer Reviewers' Openness Initiative: Incentivizing open research practices through peer review. *Royal Society Open Science*, 3(1), Article 150547.

Nature (n.d.). Reporting standards and availability of data, materials, code and protocols. https://www-nature-com/nature-portfolio/editorial-policies/reporting-standards (Accessed 12th January, 2023)

Nosek, B. A., Spies, J. R., & Moyl, M. (2012). Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631.

Nuijten, M. B., Borghuis, J., Veldkamp, C. L. S., Dominguez Alvarez, L., van Assen, M. A. L. M., & Wicherts, J. M. (2017). Journal data sharing policies and statistical reporting inconsistencies in psychology. *Collabra Psychology, 3*(1), 31.

Nutu, D., Gentili, C., Naudet, F., & Cristea, I. A., (2019). Open science practices in clinical psychology journals: An audit study. *Journal of Abnormal Psychology, 128*(6), 510-516. https://psycnet.apa.org/doi/10.1037/abn0000414

Piwowar, H. & Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ, 1,* e175, 10.7717/peerj.175.

Piwowar, H., Day, R. S., Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLOS ONE, 2*(3), e308.

Piwowar HA, Becich MJ, Bilofsky H, Crowley RS, on behalf of the caBIG Data Sharing and Intellectual Capital Workspace (2008) Towards a data sharing culture: Recommendations for leadership from academic health centers. *PloS Medicine 5*(9), e183.

PloS ONE (2011) PloS Editorial and Publishing Policies. 7. Sharing of Materials, Methods, and Data. Available from: http://www.plosone.org.abc.cardiff.ac.uk/static/policies.action#sharing.

Rotulo, A., Kondilis, E., 123he, T., Gautam, S., Torcu, O., Vera-Montoya, M., Marjan, S., Gazi, M.I., Alifa Syamantha Putri, Hasan, R.B., Mone, F.H., Rodríguez-Castillo, K., Tabassum, A., Parcharidi, Z., Sharma B., Islam, F., Amoo, B., Lemke, L., & Gallo, V. (2022*). Mind The Gap: Data availability, accessibility, transparency, and credibility during the COVID-19 pandemic, an international comparative appraisal*. MedRxiv. https://www.medrxiv.org/content/10.1101/2022.09.14.22279961v1

Rowhani-Farid, A. & Barnett, A. G. (2016). Has open data arrived at the British Medical Journal (BMJ)? An observational study. *BMJ Open, 6*, e011784. Doi:10.1136/bmjopen-2016- 011784

Savage, C. J. & Wickers A. J. (2009). Empirical study of data sharing by authors publishing in PloS journals. *PloS ONE 4*(9), e7078. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0007078

Science (n.d.). Science Journals: Editorial Policies. https://www-science-org.abc.cardiff.ac.uk/content/page/science-journals-editorial-policies#data-and-code-deposition (Accessed 12th January 2023).

Scientific Data, (n.d.). Journal Information. https://www-nature-com.abc.cardiff.ac.uk/sdata/journal-information

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2017). How to properly pre-register a study. Data Colada. http://datacolada.org/64#:~:text=A%20preregistration%20cannot%20allow%20readers,for%20the%20task%20at%20hand. (Accessed 20th December 2022).

Stodden, V., Guo, P., & Ma, Z. (2013). Toward reproducible computational research: an empirical analysis of data and code policy adoption by journals. *PloS One, 8*(6), e67111. 10.1371/journal.pone.0067111

Stodden, V., Seiler, J., & Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility. *Proceedings of the National Academies of Sciences 115*(11), 2584–2589. 10.1073/pnas.1708290115

The Journal of Open Psychology Data (n.d.). https://openpsychologydata.metajnl.com/

Towse, J. N., Ellis, D. A., & Towse, A. (2019, December 23). *Opening Pandora's Box: Peeking inside psychology's data sharing practices, and seven recommendations for change.* PsyArXiv. https://doi.org/10.31234/osf.io/k6rux. *https://psyarxiv.com/k6rux/*

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., & Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra, 1*(1), 3, pp. 1–5, DOI: http://dx.doi.org/10.1525/collabra.13

Vickers, A.J. (2006). Whose data set is it anyway? Sharing raw data from randomized trials. *Trials, 7*, 1-6.

Wagenmarkers, E.-J. & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer, 29*(9), 13–14.

Wicherts, J.M., Bakker, M., & Molenaar, D. (2011). Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. *PLOS ONE* 6(11), e26828. http://dx.doi.org/10.1371/journal.pone.0026828.

Wicherts, J. M., Klein, R. A., Swaans, S. H. F., Maassen, E., Stoevenbelt, A. H., Hartgerink, C. H. J., … Rüffer, F. F. (2022, March 29). *Privacy Protection in the Era of Open Science*. PsyArXiv. https://doi.org/10.31234/osf.io/ybzu9 *https://psyarxiv.com/ybzu9/*

Wicherts, J. M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. *The American Psychologist, 61*(7), 726-728. https://pubmed.ncbi.nlm.nih.gov/17032082/

Chapter 5: Comparative Analysis of Study Characteristics

5.1. Introduction

Recent years have witnessed rising concerns about the quality and rigour of research, and the consequent impact on validity and replicability. In turn, these concerns prompted the call for improved standards and practices in research design, including larger sample sizes and predefined sampling plans, preregistration, ensuring a clear distinction between exploratory and confirmatory analysis, replication of studies, inclusion of markers of rigour such as manipulation checks, as well as the need for sufficient detail and depth of information provided in the methods sections of research articles. This chapter will briefly consider a number of these characteristics before describing a study comparing them between RRs and SRs. Overall, the findings of this chapter indicate that RRs are associated with most such characteristics to a greater extent than SRs.

5.1.1. Sample Sizes, Underpowered Studies and Sampling Plans

The importance of increased statistical power for the replicability of research has been highlighted frequently in recent years, particularly in order to reduce false positive results (Simmons et al., 2011). Estimates of the levels of power in the literature vary slightly but have been reported to be as low as 0.35 (Bakker et al., 2012). One method of increasing the power of studies is to increase the number of observations, such as increasing the sample size (Sassenberg & Didrich, 2019; Asendorpf et al., 2013; Gelman & Carlin, 2014). However, sample sizes in the psychology literature are typically low; for example, Marszalek et al. (2011) reported a median total sample size of only 40 across four representative psychology journals. Bakker et al. (2012) explain that using several small and underpowered studies can act as a more efficient strategy to find statistically significant results than using one study with a larger and more powerful sample, thereby leading to higher rates of false positive findings and inflated effect sizes, issues which can then be further compounded by publication bias and questionable research practices (Bakker et al. 2012; Norsa, 2022). This therefore relates to the concept of the 'winner's curse fallacy', (Button et al., 2013) or the bias referred to by Loken & Gelman (2017) as 'what does not kill statistical significance makes it stronger'. These concepts are a heuristic whereby researchers may view a significant finding as particularly remarkable if it is found despite the low power of the study making this unlikely, when in fact the researcher that obtained this unlikely result can be considered to be cursed with finding an inflated estimate of that effect.

Evidence shows increasing sample sizes in psychology in recent years; Tenney et al. (2021) demonstrate some evidence of increases in the median sample sizes reported in the field of organisational behaviour between 2011 and 2019, although this increase was not statistically significant in all of the sampled journals. Sassenberg & Didrich (2019) also report finding larger samples sizes in social psychology papers published in 2016 and 2018, compared with those published in 2009 and 2011. They do however attribute this to the increase in online data collection and self-report measures as a less resource-intensive approach that may enable researchers to recruit greater numbers of participants. While this can be a helpful way to increase statistical power, such online and self-report-based approaches will not be appropriate for all studies. They also state that the journal that showed the least change over a particular time period did not have editorial guidelines that explicitly emphasised the importance of statistical power; however, after their guidelines had been changed to emphasise this, the differences in sample sizes between that journal and the others in their sample mostly disappeared (Sassenberg & Didrich, 2019). This highlights the need for these changes in standards to be specifically included in journal policies.

In addition to the sample sizes themselves, the need for a sampling plan or justification has been highlighted, in order to indicate how the study's data is expected to meet the study's inferential goals (Norsa, 2022; Asendorpf et al., 2013). In addition to statistical power calculations and other statistical approaches which are increasingly being required, justifications for the planned sample size may be based on resource constraints, such as time constraints, the costs of data collection or difficulties in accessing particular populations (Norsa, 2022; Lakens, 2021).

It has been speculated that the pre-study review process for the RR format should help to monitor for issues of low statistical power, and work to reduce this issue by requiring rigorous study plans including sample size planning (Bakker et al., 2020). However, evidence regarding this from more general preregistration or pre-study review processes has not been encouraging. Bakker et al. (2020) conducted a study of preregistrations and proposals to ethical review boards which found that being asked to conduct a formal power analysis before collecting data only increased the rates of such analyses being used but was not actually associated with an increase in the sample sizes of the studies. They also commented that there was significant room for improvement in the quality of such registrations/proposals

such as the amount of important information provided, and the facts that most of these power analyses only related to the main hypothesis test in the registration.

Evidence regarding sample sizes in RRs themselves and how they compare against sample sizes in SRs are scarce due to the relative novelty of the format. However, Soderberg et al. (2021) report that RRs had higher median sample sizes and greater inclusion of justifications for the sample sizes, indicating that RRs appear to be associated with these indicators of higher quality to a larger extent than SRs. Furthermore, sample size planning has also been examined in a sample of 46 RRs in psychology (Norsa, 2022), although they made no comparison to the standard research literature. This study found that 40% of those RRs did not specify the approach used to determine their planned sample size, although 36.95% of the RRs were found to have run a well-planned power analysis based on the existing literature and pilot studies. Given these previous findings, there appears to be some room for improvement in the use of sampling plans in RRs as well as in SRs. Further investigation of these characteristics in these articles is warranted as previous studies have relied on a relatively small sample of RR articles.

5.1.2. Preregistration and Non-Registered Studies

The need for preregistration and its potential benefits have been briefly acknowledged in chapters 3 and 4. To recap, preregistration involves the public documentation of study hypotheses, design, and analysis plans, before the study is conducted (Wicherts et al., 2016). This is thought to be important in detecting and reducing analytical flexibility and selective reporting, ensuring that analysis strategies that were pre-planned can be clearly distinguished from any exploratory or data-led approaches (Chambers, 2017; Wagenmakers & Dutilh, 2016; Wagenmakers et al., 2012; Nosek et al., 2019; Nelson et al., 2018). Toth et al. (2021) report that preregistration appeared to be associated with more transparent reporting overall, including having lower rates of statistically significant findings than non-preregistered studies (48% vs. 66%). Furthermore, preregistration may often include a sampling plan which can help to ensure the study is adequately powered or at least that there is some justification for the planned sample size and stopping point for data collection. Therefore, preregistration can be useful when adhered to faithfully and any deviations are clearly indicted, but there is still potential for degrees of freedom to be exploited and questionable practices to be engaged in even when a study has been preregistered (Mathieu et al., 2009; Ramagopalan et al., 2014). Therefore, although preregistration can't necessarily ensure research quality, it can help to increase the credibility of the research (Vazire, 2018) and may improve transparency overall.

Despite optimism about the potential of preregistration and speculation that such practices could be increasing (Nosek & Lindsay, 2018), rates of preregistration appear to be very low in the psychology literature. Rates of articles at Psychological Science earning preregistration badges were much lower than those for open data and materials badges, as reported by Lindsay (2019). Furthermore, a study of 322 empirical articles published in the journal Organisational Behavior and Human Decision Processes showed that only 10% of the papers included preregistrations, and that this practice was much less common than the sharing of data and materials (Logg & Dorison, 2021). Furthermore, they showed that the rates of preregistration have increased very little over time. Tenney et al. (2021) also showed extremely low rates of preregistration across four journals in the area of organisational behaviour research, occurring in only 0.8% of the included studies and the majority of instances were from just one journal (Organisational Behavior and Human Decision Processes). Therefore, while preregistration may be better represented in particular areas of the literature and within particular journals, it seems that it is still far from being a widespread practice. However, it has been proposed that the apparently low rates of preregistration may be partially explained by the inherent time delay between a study being preregistered, and the study being published and so metascientific findings regarding the rates of preregistrations in published articles may reflect a point slightly further back in time, making the current usage of this practice less clear. In line with this, Logg and Dorisons' (2021) survey of researchers found that although only 21% had preregistered a study prior to that point, 50% had been in the process of preregistering a study at that time, suggesting a possible increase in the uptake of this practice.

RRs offer additional benefits beyond those of standard preregistration. These benefits are due to the provision of in-principle acceptance, and thus the potential for this to further reduce questionable research practices and publication bias through increased accountability and incentives. However, additional studies such as pilot studies are often included in the final report when an RR is published (and are sometimes submitted as part of the stage 1 protocol to provide pilot data and justify aspects of the main study design). Such studies may or may not be preregistered, and it is as yet unclear to what extent stage 2 RRs include non-registered studies. For the purposes of this study, studies within each article that were RR studies or that had been preregistered were grouped together and the analysis instead examined rates of non-registered studies, comparing this between RRs and SRs.

5.1.3. Exploratory Analysis

As described in previous chapters, it is vital for researchers to clearly distinguish between confirmatory or pre-planned analyses, and exploratory analyses, in order to detect and reduce HARKing and other selective reporting. Efforts to address this therefore often rely on preregistration as this can clarify which aspects of the study were pre-planned and which were post-hoc. Post-hoc adjustments should therefore become clearer, helping to clarify and detect questionable practices (Wagenmakers et al., 2012; Wagenmakers et al., 2011; Kerr, 1998; Chambers, 2017).

However, exploratory analysis is a very important aspect of research and the over-prioritising of confirmatory non-significance hypothesis testing has been criticised by some, claiming that non-confirmatory activities are needed in order to derive a strong derivation chain between theory and tests, before well-formulated hypotheses can be adequately tested (Meehl, 1990; Scheel et al., 2020). This needs to include time and effort spent forming concepts, developing valid measures, identifying boundary conditions and auxiliary assumptions, as well as establishing causal relationships between concepts. Currently however, such efforts are lacking. Given the over-emphasis on the hypothetico-deductive approach to research and the expectations and pressures this places on researchers, this is an important area for the reform movement to consider, particularly as recent efforts have focused primarily on trying to improve how the hypothetico-deductive approach is implemented by improving the rigor of research methods (Fidler et al., 2018; Scheel et al., 2020; Spellman, 2015).

Therefore, exploratory analysis should still be encouraged, but it is essential to have clarity about whether studies, or parts of studies, are exploratory as opposed to confirmatory, in order to increase transparency and reduce questionable practices. The ongoing efforts to clarify the distinctions between confirmatory and exploratory elements of studies and to ensure that exploratory aspects are not passed off as confirmatory, are therefore an important goal. This is because when testing hypotheses, the data should only be used once and so when hypotheses are formulated based on the data, that data cannot be used to test that same hypothesis. By using the data for both purposes (known in neuroimaging as ‘double dipping’; Kriegeskorte et al., 2009; Wagenmakers et al., 2012), the type 1 error rates are inflated and p-values become less trustworthy (Wagenmakers et al. 2012; Goldacre, 2009).

Preregistration in general is thought to be important in addressing this kind of double-dipping and selective reporting (Wagenmakers et al., 2012), and RRs are believed to be particularly important in achieving this because they provide an incentive and a greater level of accountability for researchers to clearly differentiate between the different aspects of the study (Chambers, 2017). However, RRs are generally considered to be targeted towards primarily confirmatory research and a common criticism of the RRs is that they may hinder exploratory research (Chambers, 2017; Chambers & Tzavella, 2022). Although the inclusion of exploratory analysis is in fact permitted and even welcomed in RRs (Chambers 2015; Chambers et al., 2015) the perception of this being suppressed or discouraged remains. Therefore, it is important to discover to what extent exploratory analysis is actually included in RRs and how this compares with SRs. Furthermore, it is important to ascertain how clear the distinction is between confirmatory and exploratory elements of the paper. At a minimum, having a clear distinction for this is important in ensuring that RRs are being handled and published appropriately, and it would be valuable to be able to examine how this compares in SRs although actually determining this is inevitably challenging due to the lack of clarity required about this in standard articles.

5.1.4. Replication and Detailed Methods Sections

Replicating previous studies is considered vital in determining the validity of their findings, refining knowledge, and developing or refining theories (Schoenherr & Swink, 2012; Gattiker et al., 2022). Replications may be direct, whereby all aspects of a study's approach are reproduced in the same manner, or may be indirect, where certain aspects differ from the original study (Brinberg & McGrath, 1985 cited in Gattiker et al., 2022). Conceptual replications are a type of indirect approach that have received criticism due to their potential for confirmation bias. Conceptual replications involve attempting to demonstrate evidence of a particular phenomenon using different methods and measures that are considered to be similar to that phenomenon. While conceptual replication can be important in understanding the generalisability of phenomena, an over-reliance on conceptual replication instead of on direct replication can be concerning, since there is considerable subjectivity involved in the decisions made in a conceptual replication, regarding whether the concepts and measures are similar enough. The question of whether researchers are actually measuring the same concept can lead to conflicting views about whether particular findings have actually been replicated successfully or not, and this provides space for confirmation bias to influence researchers'

judgements of the replicability of particular findings based on their own views (Chambers 2017).

Increased attention has been paid to the topic of replication since the replication crisis came to light, but rates of replication efforts remain low, suggesting that this approach is not popular among researchers and there may be a lack of incentives to engage in this. For example, some reviewers and editors, as well as researchers, may consider replication research to not be as interesting as original studies and that this may reduce the likelihood of their replication efforts being published, read or cited (Chambers, 2017). Efforts to increase the replication of previous studies have made some promising progress, particularly within psychology. For example, the Psychological Science Accelerator is an international network that enables multi-lab replications of key findings in psychology (Chartier et al. 2018). Other multi-lab replication projects include the Many Labs projects (e.g., Klein et al., 2014; Klein, 2018; Ebersole et al., 2020), as well as the Pipeline Project (Schweinsberg et al., 2016). Promotion of replication studies within education of students has also been successful with examples such as the Collaborative Replications and Education Project (CREP; Wagge et al., 2019). Despite such high-profile initiatives, however, replication attempts are still relatively infrequent in psychology overall. For example, Tenney et al. (2021) found that replication was rarely included as the main purpose of studies and that this had not increased over time.

Due to the way in which RRs change the incentives within scientific publishing, the format may encourage researchers to pursue replication of previous studies to a greater extent than usual. Specific formats that encourage the submission of replication studies as RRs include the Registered Replication Report format (e.g., Simons et al., 2014), as well as several special issues that have focused on RRs based on replication studies (e.g., Carsten et al., 2023; Nosek & Lakens, 2014). Despite the frequent challenges of agreeing on replication study design and on expectations regarding the meaning of various outcomes (Nosek & Errington, 2020), the earlier review process for RRs may facilitate discussions around what various outcomes will mean and what design choices will be meaningful, potentially avoiding confusion and disagreements later on in the process. Therefore, Nosek and Errington call for authors of original studies and independent replicators to reach a ‘precommitment’ agreement prior to replication attempts, to agree on a suitable replication experiment approach. They specify that this should be clearly documented using preregistration or, ideally, an RR.

Early evidence also indicates that rates of replication are reasonably high among RRs, with Scheel et al. (2021) reporting that 58% of their sample of psychology RRs were replication studies, compared to only 3% of their SRs. Norsa (2022) found that of the 46 psychology RRs they examined, slightly more of these were original studies (n=27, 58.69%) rather than replications, such as Registered Replication Reports of a particular study (n=14, 30.43%) or multi-lab replication studies (n=5, 10.86%). While this still makes replication studies the minority of those in the sample, the rates are still much higher than is typical in the general psychology literature. They also found that the replication attempts appeared to be less likely to find a significant effect than the RRs that reported on original studies (59.26% successful vs. 21.05%). The current study aims to build on these findings in the current study by comparing the rates of replications of various kinds, in a larger sample of articles.

In order to achieve a successful replication of a previous study, there must be sufficient detail provided about the methods used (APA, 2010 cited in Brenninkmeijer et al., 2019). However, the level of detail provided in methods sections varies considerably depending on where it is published and the journals' policies regarding word limits. Furthermore, the level of specificity that researchers believe is necessary or customary can vary considerably (Wilson, 2014; Srivastava, 2016). This is an important issue because the lack of particular details could lead to unknown moderators that influence the success of subsequent replication attempts (Brenninkmeijer et al., 2019). For example, an experiment by Friedman (1967 cited in Brenninkmeijer et al., 2019) demonstrates variations in researchers' behaviours when following the same set of procedural details, and that these differences also partially influenced the participant's behaviours. Some of these differences and omitted details have been attributed to the concept of 'lab lore', or informal practices about study procedures that are passed down within a lab or research group but are not formalised and reported more widely (Buskist & Johnston, 1988; Srivastava, 2016). Brenninkmeijer et al. (2019) interviewed 22 experimental psychologists in the Netherlands in an attempt to understand such informal laboratory practices and revealed that experimental psychologists do frequently neglect to mention various aspects of their methods or practices in their articles, such as the specific instructions used for participants. Attitudes towards such informal practices varied between participants, with some not seeing them as a problem, but many others feeling that making such practices explicit would be helpful especially if they are likely to influence the outcome of an experiment, although it was sometimes recommended that this should be done with caution. Such aspects included practices such as the use of (mild) deception, the order

of measurements, the use of brief vs. elaborate instructions, and how standardized the experimental protocol was. Other aspects noted as frequently not being reported were often those that can be considered to be related to professionalism in interacting with participants and in preparation of suitable settings to conduct the research in. However, this was typically considered to be more of an ethical consideration than a methodological issue, although some did consider this to be important in ensuring participants took the study seriously and so provided good data. A number of other considerations important for producing good data were also highlighted.

It is unclear whether the amount of detail reported regarding the study methods has changed in recent years in line with increased attention to methodological rigour, although Lindsay (2019) does acknowledge that overall article lengths at Psychological Science have increased considerably from a mean length of 6.6 pages in 2010 to 11.2 pages in 2018 and that these tended to include more technical and statistical details. However, it is uncertain whether this trend is also apparent in other journals and subdisciplines and if so, to what extent. It is also currently unclear whether RRs have longer or more detailed methods sections than SRs do, although it seems possible since the protocols submitted for peer review need to be sufficiently detailed for the reviewers to evaluate the study based on the proposed methods. However, no empirical evidence yet exists regarding this and so, the word counts of the methods sections were examined in the current study in order to compare this between the RRs and the SRs.

5.1.5. Manipulation Checks

Manipulation checks are a diverse set of measurements that, generally speaking, are designed to ensure that a manipulation has had the intended effect (Hauser et al., 2018). While others adopt a narrow view of what constitutes a manipulation check (e.g., Chester & Lasko, 2021) such as measures of the actual construct the manipulation is meant to affect, others also include instructional manipulation checks such as attention checks or comprehension checks (Oppenheimer et al., 2009; Hauser et al., 2018).

Checks like these are important to help determine whether experimental manipulations have had the intended effects and to verify whether participants were even paying attention and reading instructions, particularly as online data collection becomes increasingly common. The importance of such checks therefore should not be understated; Oppenheimer et al., (2009) found that 30% of their participants failed a simple attention check (known as an instructional manipulation check) and that only those who had passed this check were

affected by the experimental manipulation, indicating that those who did not read the attention check question carefully enough may also have not read the actual task question(s) carefully enough. Experimental manipulation checks are also considered important, particularly when attempting to manipulate a concept such as the researcher's emotional state or other intangible construct. The more recent use of manipulation checks for examining possible mediating variables between the independent and dependent variable, can also help to clarify potential processes through which the effect may occur (Hauser et al., 2018; Lench et al., 2014). Therefore, manipulation checks have generally been viewed as desirable additions to experimental study design.

However, some contest the view that manipulation checks are always a positive addition and speculate that these may in fact influence the participants, either by implying something about the researchers' hypotheses, or by influencing the participant's behaviour. In this way, a manipulation check could potentially constitute another (unintended) manipulation and could even undo or interact with the effects of the intended manipulation (Fayant et al., 2017; Hauser et al., 2018; Ejelov & Luke, 2020). This could therefore undermine the validity of the study's conclusions and may also alter the procedure sufficiently to undermine replication attempts if the manipulation check was not part of the original study, as has been suggested regarding Wagenmakers et al.'s (2016) failure to replicate the effect of the 'pen in mouth' study originally reported by Strack et al., (1988) although their added manipulation check has not been conclusively demonstrated as being responsible for this (Hauser et al., 2018). Despite these criticisms, much of the discourse around manipulation checks is favourable, with these typically being viewed positively and as a desirable addition to studies, particularly if they are used as properly and unobtrusively as possible (Hauser et al., 2018; Ejelov & Luke, 2020).

Manipulation checks appear to be relatively common in certain areas of the psychology literature. Hauser et al. (2018) examined articles in five psychology journals in 2015 and early 2016 and showed that manipulation checks were present in 33% of the articles, although they were much more common in social psychology than general psychology journals. However, they do highlight that many did not manipulate the order in which these appeared, leading to the potential for order effects. Likewise, Ejelov and Luke (2020) report rates of manipulation checks as high as 57.25% within key psychology journals, although a number of problematic practices and deficiencies were detected in how these were used.

It is unclear how widespread manipulation checks are in the literature and particularly within RRs, so the current comparative study will attempt to investigate this within the current comparative sample of RRs and SRs in an effort to clarify this.

5.1.6. Registered Reports as a Potential Solution

It was hoped that RRs may help to improve the overall quality of studies conducted, increasing their rigour and transparency (Chambers, 2017). Furthermore, concern has been expressed that in a system plagued by publication bias, reviewers' attention to clean and significant results may overshadow their attention to the rigour and quality of the research methods, or so that these may be unduly influential in the publication decision at the expense of the methods (Soderberg et al. 2021). Therefore, as RRs require that peer review focuses on the research questions and methods by conducting the reviews before the study takes place, it is expected that this may reduce publication bias and ensure that the reviewers' evaluation is focused on the methods without the potential to be distracted by the results. The incentive for researchers to then comply with the proposed methods if they receive in-principle acceptance holds them accountable for adhering to the approach they pre-specified.

The existing evidence, though limited, appears to support these speculations. The evidence discussed in chapter 3 regarding the much lower rates of positive findings in RRs than SRs (Scheel et al., 2021) may suggest more rigorous approaches. Furthermore, a recent observational study by Soderberg et al. (2021) demonstrates that RRs were considered to have much more rigorous methodology, higher quality methodology, better alignment of the research question and methodology, more rigorous analysis strategy, and higher overall paper quality, compared with SRs. This study involved expert review of matched stage 2 RRs and SRs and mimicked a typical peer review process, although the blinding of the reviewers to the article type they were evaluating could have benefitted from being stronger, as just over half of them correctly identified which article was an RR and which was an SR. Furthermore, they reported slightly greater differences in rating from the reviewers who reported believing that the RR format improves rigour and quality. Given the suspicions of the reviewers regarding the format of the articles they were evaluating, this raises the possibility of confirmation bias affecting their results (Higgs & Gelman, 2021).

Importantly, Soderberg's study mainly investigated reviewers' perceptions of the quality or rigour of the studies, but not more detailed aspects of the methods themselves. This may be a concern as researchers may differ on what characteristics constitute high quality or greater

credibility or which are the most important characteristics for this. These differing standards or conceptualisations may therefore lead to variations in how they evaluate the studies (Higgs & Gelman, 2021). While some limited evidence does now exist regarding the comparative rates of particular practices and standards (as outlined in the previous sections of this introduction), additional research would be helpful to verify these findings. The current study reported in this chapter therefore aims to investigate a variety of characteristics of the studies reported within these different formats in order to compare how these different article types perform on particular characteristics that are often associated with greater rigour and transparency, in addition to some other general characteristics regarding the methods used, in order to gain a better understanding of what the studies reported in these different article formats consist of. Additionally, previous studies done in this area used a much smaller sample size than that used in the current study. For example, Soderberg et al.'s sample of articles consisted of only 29 RRs and 57 SRs. Therefore, the current study aims to investigate the rates of a range of specific methodological characteristics in this much larger sample of RRs and SRs in order to understand how these differ between the two article types.

5.1.6. Research Questions

The research questions and hypotheses investigated in this chapter are outlined in Table 6 below.

Table 6

Overview of research questions and hypotheses for chapter 5

Research Question	Hypothesis
RQ 1: Are there differences in sample sizes between RRs and SRs?	H1: Sample sizes will be larger in RRs than in SRs.
RQ 2: Are there differences in the use of sampling plans between RRs and SRs?	H2: RRs are more likely to use a pre-defined sampling plan compared with SRs
RQ 3: Do the number of studies included in a paper differ between RRs and SRs?	H3: RRs will, in general, contain more studies per paper than SRs do.[6]

[6] This hypothesis was initially suggested with the assumption that authors of RRs may be less likely to engage in other questionable practices such as salami slicing data from a study into many separate publications in order to maximise the number of outputs, and/or that authors and reviewers of RRs might be more likely to require more evidence/replication of the findings in order to be convinced before publishing, compared with authors of a standard report. In hindsight, however, any potential difference may lie in the opposite direction, as conducting many smaller studies and reporting these in one paper, rather than conducing one larger and more rigorous

RQ 4: What proportion of RRs include non-preregistered/non-RR studies in their papers (e.g., unregistered pilot studies), and how does this compare to the SRs?	No specific hypothesis was stated for RQ 4.
RQ 5: Are there differences in the inclusion of exploratory analysis between RRs and SRs?	H5: RRs are more likely than SRs to have included exploratory analysis.
RQ 6: Are there differences between RRs and SRs in the proportion of papers that clearly distinguish confirmatory from exploratory analysis?	H6: RRs are more likely than SRs to have clearly distinguished between confirmatory and exploratory analysis.
RQ 7: Are there differences in the nature of the exploratory analysis included in RRs and SRs?	No specific hypothesis was stated for RQ 7.
RQ 8: Are there differences between RRs and SRs in the proportion of replication studies reported?	H8: RRs will be more likely than SRs to contain replication studies.
RQ 9: How do the lengths of the methods sections differ between RRs and SRs?	. H9: The word counts of the methods sections will be larger in the RRs than in the SRs.
RQ10: Are there differences in the use of manipulation checks or other equivalent data quality checks etc. between RRs and SRs?	H10: RRs will be more likely than SRs to include manipulation checks or other equivalent data quality checks.

5.2 Methods

5.2.1 Overview of Initial Coding Process

The process of matching the samples and developing the coding protocol were the same as described in previous chapters. The overall difficulty of the initial coding process for a range of characteristics was then coded for and analysed as outlined in the online Supplementary Appendix 1. Comparisons between the coding of characteristics for the two article types showed that overall there was no significant difference between the article types in the overall level of coding difficulty, or in the difficulty of coding most of the characteristics. There was however a greater level of difficulty when coding whether manipulation checks were passed in the RRs than in the SRs (12.28% vs. 1.61%), and this difference was statistically

study, may be more likely among SR authors, particularly without the influence of stage 1 reviewers to mitigate this.

significant: $x^2(1, N = 119) = 5.39$, $p = 0.02$. The difficulty of coding the number of studies was also more common in the RRs than in the SRs (4.12% vs. 0.88%), which was also statistically significant: $x^2(1, N = 510) = 6.17$, $p = 0.01$.

5.2.2. Coding of Sample Sizes

5.2.2.1. Initial Coding of Sample Sizes

The initial coding of sample sizes was done at each applicable level of the three coding levels, depending on where this information was presented in the paper and based on the structure of the individual paper and its hypothesis structure. At each level, the sample size was gathered for three different stages: intended or planned, pre-exclusion, and post-exclusion.

Planned or intended sample size was the sample the authors reported intending to gather, typically determined by some kind of sampling plan. This might have been based on the result of a power calculation, or some other intended number for recruitment e.g., based on the number it was thought possible to recruit given resource constraints. The pre-exclusion sample size was the sample size recruited prior to any exclusions from analysis, while the post-exclusion sample size was the final sample after any participants have been excluded from the analysis. If intent-to-treat analysis had been used this could be considered as the post-exclusion sample size if necessary and appropriate. Response options for each of these three characteristics were recorded as open responses in digits, not words, or if appropriate this could be coded as either ‘unclear’ or N/A. Information about the unit of the sample was also coded alongside this, to help inform the coding of the numerical data. An open response was used for this but while also trying to keep the words used consistent e.g., ‘people’, ‘rats’, ‘articles’, ‘dyads’, ‘organisations’, etc. If no sample data was appropriate for a particular level this was coded as N/A.

These characteristics were only coded at article level (as opposed to study level) if the same overall sample was used across all of the studies in a multi-study article, i.e., if they reported these altogether without differentiating between different studies. The information was coded at study level if this was the sample size used for that entire study (i.e., if it was not broken down separately per hypothesis within the study). If the sample was sub-divided further for particular hypotheses, the sample size could be coded at hypothesis level instead. For example, this would occur if there was a different sample per hypothesis within the study, or

if there is only one hypothesis within the study so that the methodological details used to test that one hypothesis would all need to be coded at hypothesis level rather than study level.

5.2.2.2. Analysis of Sample Sizes

In order to create an overall-level variable from the data that had been gathered across the three coding levels, the mean sample size was calculated for each paper, whether that sample size was initially coded at study or hypothesis level. If the sample size for a multi-study paper was only given at article level, this was divided by the number of studies in the paper in order to give a reasonable estimate of an average size per study. If the sample size had been sub-divided up for each hypothesis within a particular study, the mean of these was taken as the mean for that study. Descriptive statistics were examined. Mann Whitney U tests were undertaken to compare the samples sizes between the article types, and this was done for each of the three stages of the sample i.e., planned, pre-exclusion, and post-exclusion.

5.2.3. Coding of Sampling Plans

5.2.3.1. Initial Coding of Sampling Plans

Initial data was gathered for what kind of sampling plan was used and this was coded at each of the three coding levels that was applicable based on how the information was presented. For example, if the overall sample size had been given at article level in a multi-study paper, then the sampling plan was likely to have also been coded at that level. Response options used at this stage were as follows: Frequentist power, Bayesian, Other, Unclear, and N/A. Frequentist power typically involved some kind of power calculation and in some cases the authors may have used G*Power for this. Power calculations were often informed by the samples used in previous studies. ‘Other’ approaches to the sampling plan would include being based on the sample size from a previous study (when this has not been used to inform a power calculation), or if recruitment plans were limited by resource constraints. Such constraints might include limited funding to pay participants, or working with a difficult-to-reach population that is limited in number and so if the researchers aimed to recruit as many participants as possible within these limits, this could be considered as ‘Other’.

5.2.3.2. Variable Creation & Analysis of Sampling Plans

Seven categorical variables were created based on the information gathered during the initial coding process. These were created at the overall level of the paper rather than being subdivided per sub-level. Two of these variables were broader overarching variables – one for whether there was any sampling plan, and one for whether there was any kind of

statistical sampling plan. The other categorical variables documented whether there was a power calculation, a Bayesian approach, an 'other' approach. Frequencies were examined for each of these categorical variables in order to see how common each type of sampling plan was in both the RRs and the SRs. Chi-square analysis was conducted to determine if differences between the article types were statistically significant.

5.2.4. Coding and Analysis of the Number of Studies

During the initial coding process, the number of studies reported in each paper or inferred from the description within the paper, was documented. If the overall structure of the paper was not very clear or if multiple studies were grouped together when described in the paper, it was sometimes necessary to infer the exact number of studies. For example, in some papers, studies may describe a pilot study in the same section as they describe a main study but if this pilot study was sufficiently detailed and used a different sample than that of the main study, these could be considered as separate studies.

No further changes were needed to the data to create a variable as the numerical data gathered for this during the first coding stage was sufficient. This data was checked before use to ensure that there were no apparent issues with it. The mean and median scores were examined but the median was prioritised as the measure of choice for this. Mann Whitney U tests were run to compare the number of studies between RRs and SRs.

5.2.5. Coding of Preregistration Status

5.2.5.1. Initial Coding of Studies' Preregistration Status

Data was coded for whether each study had been pre-registered or not and whether it is an RR, because many RR papers also include unregistered pilot studies, or studies that have been preregistered (e.g., on OSF or as clinical trials) but were not part of the accepted stage 1 RR. Initial response options for this were RR, 'preregistered non-RR', 'non-preregistered non-RR', or unclear. The RR status of individual studies could be somewhat unclear at times if not explicitly stated. A single paper could contain both preregistered and non-preregistered studies, so each study was checked carefully for indications of this. If a protocol was available, this sometimes needed to be checked to see which studies within the paper were preregistered and whether they were part of the accepted RR.

5.2.5.2. Variable Creation & Analysis of Preregistration Status (i.e., Inclusion of Non-Registered Studies)

For the purposes of the analysis, RR and preregistered studies were both considered to be preregistered since, even if the RRs were not publicly preregistered, they had been effectively preregistered with the journal during the review process. With this in mind, the focus was instead on the studies in the paper that were not preregistered and not RRs, (i.e., those that had been coded as 'non-preregistered non-RR' during the initial coding process). Two variables were created for this. One was a categorical variable for whether a non-registered study had been included, and one was a proportion score to indicate the proportion of studies within each paper that was not preregistered in any way.

For the categorical variable for whether non-registered studies had been included, frequencies were examined and chi-square analysis was run to investigate whether there were any significant differences in this between the two article types. For the proportion score, the median was examined, and a Mann-Whitney U test was used to check if there was a significant difference in this between the article types.

5.2.6. Coding of Exploratory Analysis

5.2.6.1. Initial Coding of Exploratory Analysis

At each of the three levels, as applicable, information was gathered for whether there was, or appeared to be exploratory analysis. This was initially coded for regardless of whether or not the rest of that paper or study also contained any hypotheses i.e., whether or not there was also a confirmatory aspect of the paper. Response options used for this were 'yes', 'no', 'unclear', or N/A.

Unless there was a clear distinction between confirmatory and exploratory/additional parts, it could sometimes be difficult to tell whether and where there was a difference in this and so this sometimes needed to be inferred. An additional characteristic was also coded for alongside this, for whether there was a clear distinction between the confirmatory and exploratory analysis. Response options for this were 'yes', 'no', and N/A.

In a multi-study paper, there may be overall exploratory analysis at the article level, rather than exploratory analysis conducted just within a specific study. In a paper with multiple hypotheses at hypothesis level, exploratory analysis characteristics would be coded as yes at hypothesis level only if we could attribute the exploratory analyses to particular hypotheses at that level. If this couldn't be attributed to specific hypotheses at hypothesis level or if the

exploratory analysis was more of an overall, study-level exploration, it would instead be coded at study level.

An attempt was then made to code whether the exploratory analysis that was included was general exploration, or whether it followed up on stated hypotheses, etc. However, it is important to acknowledge that the exact distinction between general and follow up exploration was often difficult to determine and this judgement was largely very subjective. Sometimes the exploratory analysis given, especially at a more overarching article or study level, might include aspects of both of these and so it would then be coded as 'Both'. N/A was used if exploratory analysis had not been included at this level and so this characteristic was not applicable. 'Other' was included as a possible response option in case of instances that didn't fall correctly into any of these categories although this was rarely used, if ever. The nature of the exploratory analysis was coded for at whichever of the three levels was applicable within each paper.

5.2.6.2. Variable Creation & Analysis for Use of Exploratory Analysis

Two overall-level variables were created based on this initial data that was gathered. A categorical variable was used to reflect whether exploratory analysis had been included. For whether there was a clear distinction between the confirmatory and exploratory elements, a proportion score was calculated for each paper based on all of the exploratory analysis that had been documented in each paper, and the proportion of this that was clearly distinguished.

Two categorical variables were created for the nature of the follow up analysis; one of these indicated whether the exploratory analysis was follow-up, and another indicated whether this was general exploration. This was coded based on the information coded during the initial coding process.

Frequencies were examined for the categorical variable for whether exploratory analysis had been included in the paper. Chi-square analysis was used to check whether there was a statistically significant difference in the inclusion of this characteristic between the article types. To examine the proportion that had a clear distinction between exploratory and confirmatory elements, the median proportion was examined for RRs and SRs, and a Mann-Whitney U test was run to determine whether there was a statistically significant difference between the article types. Frequencies and chi-square comparisons were also used to examine the nature of the exploratory analyses and compare this between the article types.

5.2.7. Coding of Replication Status

5.2.7.1. Initial Coding of Replication Status

Studies within each paper were coded for whether the study was (or appeared to be) original or a replication attempt. Replications could be direct or indirect and could be either a replication of a different study (an external replication), or an internal replication of their own earlier study in the same paper. Response options used were as follows: original, direct replication, indirect replication, direct internal replication, and indirect internal replication.

Indirect replications were sometimes referred to as 'conceptual replications' in the papers. The difference between direct and indirect replications could be quite subjective at times. Typically, a direct replication (or 'close' replication) used the same study methods (or a slightly altered version of the same methods) as the study they were replicating. An indirect or 'conceptual' replication would have differences in the method, such as different measurement tasks, different population, etc. but aimed to replicate the same general finding as a previous study. However, it was often challenging to determine whether minor differences in the approach were substantial enough to merit the study being classified as an indirect rather than direct replication and the coder's best judgement of this had to be given.

It could also be difficult to be certain if a study was original since most research is based on or informed by previous studies to some extent and so the distinction between original studies and conceptual replications in particular could sometimes be challenging to determine. Therefore, unless the paper explicitly stated that they were attempting to replicate some aspect of a previous study or that their study was based on one conducted by a certain author, it was generally considered appropriate for the study to be coded as original instead. Some papers mentioned vaguely in the discussion section that their findings replicated the findings of a previous study. However, unless the authors stated or seem to very clearly imply from the beginning of the paper that they were deliberately trying to replicate this previous study's finding, such statements were considered likely to be post-hoc, meaning it was not a true replication attempt and could instead be coded as 'original'.

5.2.7.2. Variable Creation & Analysis of Replication Status

Five categorical variables were created based on the data gathered during the initial coding process. One of these was used to indicate whether any kind of replication was included in each paper. Four categorial variables were then created to indicate the presence of each kind of replication specifically: direct (external) replications, indirect (external) replications, direct

internal replications, and indirect internal replications. Frequencies were examined for each of these categorical variables, and chi-square analysis was used to compare the rates of these between the two article types.

5.2.8. Coding & Analysis of Methods Sections' Word Counts

The word count of the methods section(s) within the paper were coded at article and study level as these were the levels at which methods sections were considered most likely to occur. This involved finding the number of words in the methods sections for each study, excluding tables and footnotes. To get the word count, each methods section was copied and pasted into a word document, making sure any tables, footnotes or article page headers/footers (e.g., name of the journal or page numbers), were not included in this; to avoid this it was necessary to copy and paste this content page by page. The heading 'Methods' was not included when copying/pasting, but any subheadings within the section were retained. It was also ensured that the spacing between words copied over correctly, as sometimes some of the spaces or hyphens from the original document could be missing, which would cause errors in the word count given by the Word document.

It is important to acknowledge that not all elements are included the same across papers and journals, as some have details of analysis techniques included in the results section rather than methods section. In other cases, there could be a lack of detail in the methods sections of later studies of a multi-study paper if the same methods or measures were used and had already been described in detail in the earlier studies. Despite the potential for such variations and inconsistencies, whatever authors had presented as the methods section for that study was included in the word count. Since the RRs and SRs were from the same journals, it was considered less likely that there would be important levels of journal-based variations between the two article types in how methods sections were presented and what specific components they include, even if this could occur within the total sample of each article type.

While the word count was typically coded at study level, in some cases a communal methods section was given at article level in certain papers that described the methods across multiple studies and which could not be easily disentangled for each specific study and so this was coded at article level in such cases.

The mean word count was then calculated for each paper based on the numerical data collected during the initial coding process. Where a communal methods section had been

presented at article level in a multi-study paper, this was divided by the number of studies in the paper in order to get an approximation of what the mean score per study might have been.

The mean and median were both examined but the median score was considered more important. A Mann-Whitney U test was used to compare this characteristic between the two article types.

5.2.9. Coding of Manipulation Checks

5.2.9.1. Initial Coding of Manipulation Checks

During the initial coding process, the presence of a manipulation check was coded at each of the three levels, as applicable, and at each applicable level this was included three times in order to allow for the possibility of multiple manipulation checks at each of these levels. Response options used to code for this characteristic were 'yes', 'no', 'unclear' and N/A.

Manipulation checks may be referred to by other terms e.g., validation check, attention checks, comprehension check, or positive control. In line with the conceptualisation of a manipulation check used by Oppenheimer et al., (2009), as well as by Hauser et al., (2018), instructional manipulation checks such as attention and comprehension checks were also included in our definition when searching for these measurements in the papers. Due to lack of familiarity with many of the specialised research topics of the different papers, it was often difficult to identify whether manipulation checks had been used unless it was clearly stated by the authors, as they may have described them without referring to them explicitly as manipulation checks or a similar term, assuming readers who were experts in that particular topic area would recognise these tasks as such.

After coding whether there was a manipulation check, whether this check had been passed was also coded using the response options 'yes', 'no', 'unclear' or N/A. Wherever a manipulation check had been coded as existing (at a particular level and however many times), whether this check was passed was also coded. If the manipulation check was successful, this was coded as Yes. Also, if those who failed the manipulation check were excluded from the analysis, this could also be coded as 'Yes'. It would be coded as N/A if no manipulation check was used (i.e., if the previous column was coded as 'No'). This was only coded as 'Unclear' if absolutely necessary.

5.2.9.2. Variable Creation & Analysis of Manipulation Checks

Variables were created based on the data gathered during the initial coding process, and these were created at a broader overall level rather than the three specific sub-levels used during

the initial coding process. For the presence of the manipulation checks, a categorial variable was created for whether a manipulation check was included in the study. A proportion score was created to indicate what proportion of the manipulation checks in the paper were passed, and another for whether this was failed.

For the categorical variable for the presence of manipulation checks, frequencies were examined and chi-square analysis was used to compare this characteristic between the two article types. For the proportions of manipulation checks that were passed or failed the mean and median scores were examined but the median score was considered more important for this. Mann-Whitney U tests were used to compare these characteristics between the two article types.

5.3. Results

5.3.1. Sample Sizes

As outlined in section 5.2.3. of this chapter's methods section, samples sizes were calculated for intended, pre-exclusion, and post-exclusion stages of the studies.

5.3.1.1. Differences in Intended Sample Size

Data had been available for the intended sample size in only 163 articles (95 RRs and 68 SRs). Although the mean intended sample size was considerably higher among SRs (M = 8294.30, *SD* = 59042.83) than among the RRs (*M* = 841.10, *SD* = 3012.21), the medians were considered a more appropriate measure of central tendency for the comparison due to the non-normal distribution of the data. Descriptively, the difference between the medians was in the expected directions, with the RRs having a higher median for the intended sample size (144.0) than the SRs (63.17). Comparative analysis using the Mann-Whitney U test showed a borderline significant difference (*U* = 3822, *p* = 0.047) between the article types in the intended sample size.

5.3.1.2. Differences in Pre-Exclusion Sample Size

As with the intended sample size, the mean pre-exclusion sample size was much higher for the SRs than for the RRs (*M* = 14078.48, *SD* = 204578.68, vs. *M* = 1564.22, *SD* = 5909.09). However, the medians appeared to be in the expected direction, with the RRs having a higher pre-exclusion sample size than the SRs (146.67 vs. 91.0). A Mann-Whitney U test found a significant difference between the article types (*U* = 20192.0, *p* = 0.04).

5.3.1.3. Differences in Post-Exclusion Sample Size

As with the intended and pre-exclusion sample size analyses, the means were in the opposite direction than expected, with the SRs having a much higher mean (M = 23632.15, SD = 233710.73) than the RRs (M = 846.2, SD = 2140.272). The medians, however, were in the expected direction, with the RRs having a higher median sample size than the SRs (140.80 vs. 96.15). The Mann-Whitney U test did not find a statistically significant difference between the article types for this variable (U = 16627.0, p = 0.13).

5.3.2. Sampling Plans

5.3.2.1. Differences in Use of <u>Any</u> Sampling Plans

The full comparative sample (N = 510, i.e., 170 RRs and 340 SRs) were used for this overall analysis of the sampling plans. Frequencies showed that 80.59% of RRs had some kind of sampling plan, compared with only 47.06% of SRs. Chi-square comparison on the use of sampling plans between the two article types showed a statistically significant difference between the article types for the use of any kind of sampling plan: $x^2(1, N = 510) = 52.39$, $p < 0.001$

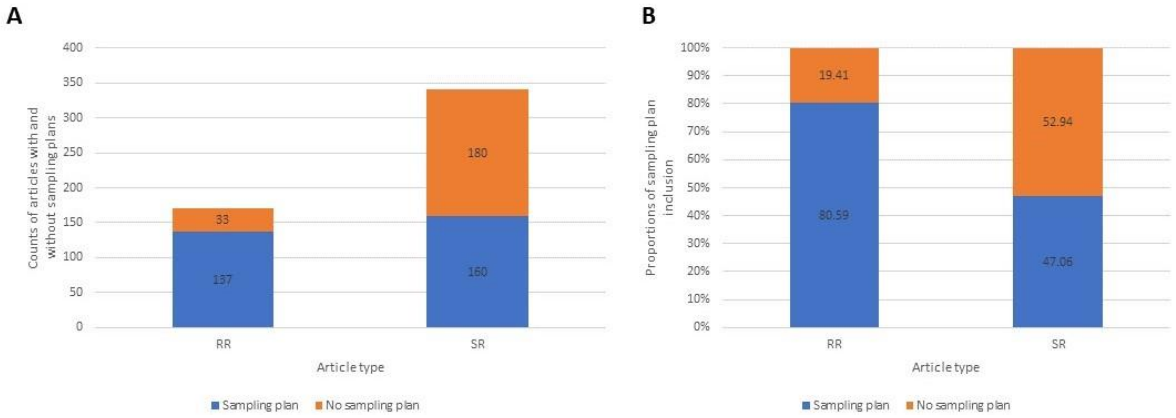

Figure 12: Inclusion of any kind of sampling plan, per article type. ***A****. Counts of articles with and without sampling plans, in RRs and SRs.* ***B.*** *Proportions of inclusion and non-inclusion of sampling plans in RRs and SRs.*

5.3.2.2. Differences in Use of <u>Statistical</u> Sampling Plans

Like the overall sampling plan analysis, this analysis was conducted on the full sample of 510 articles. Frequencies showed that statistical sampling plans were more common in RRs (50.59%) than in SRs (13.53%). A chi-square analysis on this between the two article types showed a statistically significant difference: $x^2(1, N = 510) = 81.14$, $p < 0.001$.

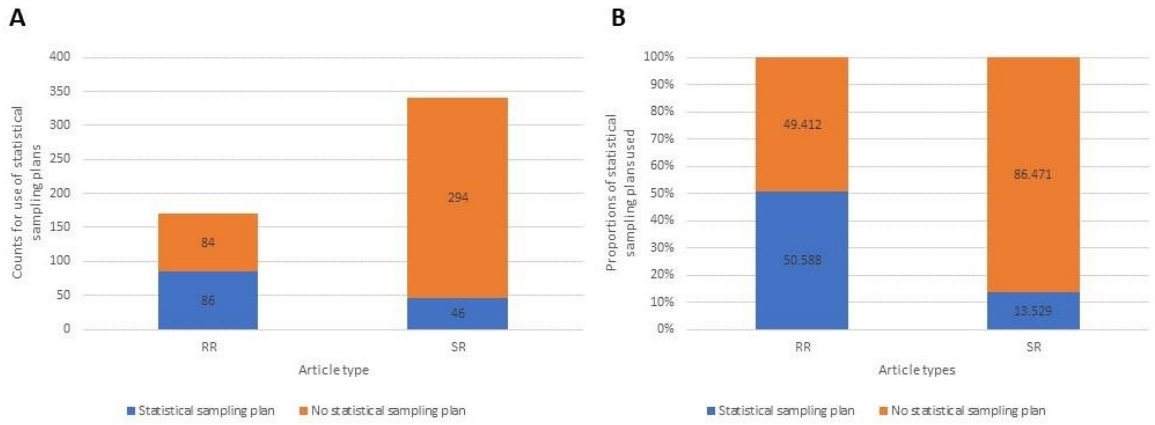

Figure 13: Inclusion of statistical sampling plans, per article type. ***A****. Counts of articles with and without statistical sampling plans in RRs and SRs.* ***B****. Proportions of inclusion and non-inclusion of statistical sampling plans in RRs and SRs.*

5.3.2.3. Differences in Types of Sampling Plans.

Both article types had a mode of 0 for the use of power analysis. However, frequencies showed that RRs had greater use of power analysis (42.94%) than SRs (12.94%). A chi-square comparison of the use of power analyses between the two article types showed a significant difference: $x^2(1, N = 510) = 57.70$, $p < 0.001$.

Frequencies showed that RRs had greater use of Bayesian sampling plans (8.24%) than SRs (0.59%), and a chi-square analysis showed the following significant difference in this between the article types: $x^2(1, N = 510) = 21.81$, $p < 0.001$.

The mode for the use of any other kind of sampling approach was 0 for both article types and the frequencies showed similar rates for both article types, with 33.82% of SRs and 32.94% of RRs including these. A chi-square analysis showed no significant difference: $x^2(1, N = 510) = 0.04$, $p = 0.84$.

5.3.3. Number of Studies Per Paper

The full comparative sample of 510 articles was used for this analysis. The median for the number of studies was 1 in both article types. The mean was only slightly higher in the RRs than in the SRs and this had a larger SD (*M* for RRs was 1.78, *SD* = 2.42, vs. *M* for SRs which was 1.64, *SD* = 1.33), but regardless, the median was preferred because of the non-normality of the distribution. An independent-samples Mann-Whitney U test was used, and this showed no significant difference between the article types in the number of studies (U = 29227.50, p = 0.80).

5.3.4. Inclusion of Non-Preregistered/non-RR Studies

For the inclusion of non-registered studies within the paper the mode for SRs was 1 whereas the mode for RRs was 0. Examining the frequencies revealed that 91.47% of the SRs contained at least one non-registered study, compared with only 9.41% of the RRs. A comparative analysis using chi-square found a statistically significant difference in the expected direction: $x^2(1, N = 502) = 325.34, p < 0.001$.

Similarly, the proportion of non-registered studies included showed a similar pattern. The median proportion of non-registered studies included was 1 for SRs, compared with a median of 0 for RRs, while the mean for SRs was also higher than that of the RRs ($M = 91$, $SD = 0.28$ vs. $M = 0.06$, $SD = 0.18$). Comparative analysis using a Mann-Whitney U test showed a significant difference between the article types in the proportion of non-registered studies included, and this was in the expected direction ($U = 2564.0$, $p < 0.001$).

5.3.5. Use of Exploratory Analysis

5.3.5.1. Differences in Inclusion of Exploratory Analysis

The mode for whether exploratory analysis was included was 1 for both SRs and RRs, when the full comparative sample (n=510) was examined. Within this same full comparative sample, frequencies were examined for whether exploratory analysis was included in each article type. SRs appeared to be less likely to include exploratory analysis than RRs were, with 57.06% of SRs having exploratory analysis, compared with 75.29% of RRs. The chi-square comparison between the two article types showed a statistically significant difference between the two article types for whether exploratory analysis was included in the paper: $x^2(1, N = 510) = 16.19, p < 0.001$.

The analysis was also run after filtering out the papers that only had exploratory analysis without any suggestion of confirmatory aspects, so that the analysis sample only consisted of papers that had at least some confirmatory component (n = 381). Within this sample, the results were very similar to the previous analysis, as considerably more RRs contained exploratory analysis (71.43%) than SRs did (43.15%). Chi-square analysis showed a statistically significant difference between the article types in the rate of inclusion of exploratory analysis, in the restricted sample: $x^2(1, N = 381) = 28.46, p < 0.001$.

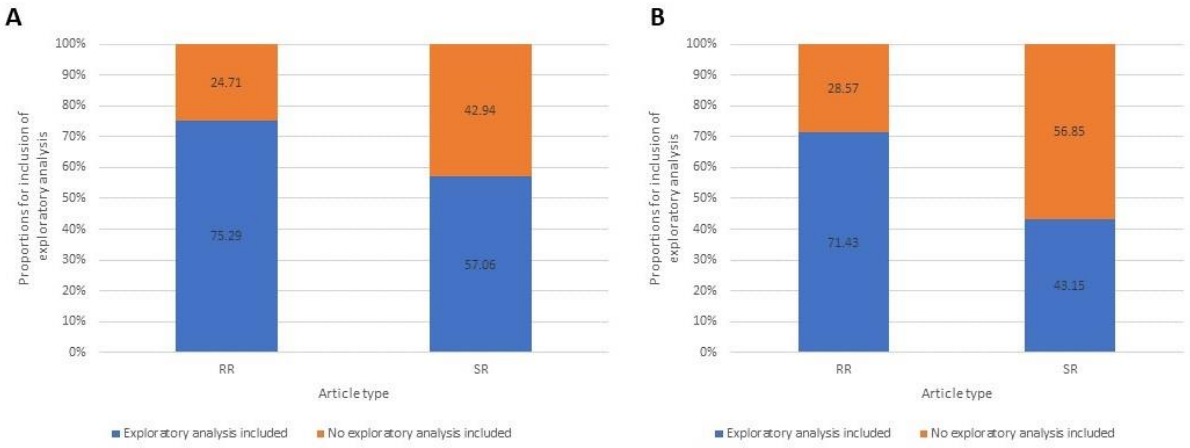

Figure 14: ***A.*** *Proportions of articles that did and did not include exploratory analysis in SRs and RRs, in full comparative sample.* ***B****. Proportions of articles that did and did not include exploratory analysis in SRs and RRs, in restricted sample of articles that only contained exploratory analysis. Both graphs show that a higher proportions of RRs than SRs included exploratory analysis.*

5.3.5.2. Additional Exploratory Analysis: Differences in Whether Papers ONLY Contained Exploratory Analysis

The mode for whether papers contained ONLY exploratory analysis was 0 (no) for both the SRs and RRs. Frequencies showed that a relatively small proportion of the articles had only exploratory analysis, but that this was more common among the SRs than among the RRs (29.12% vs. 17.65%). A chi-square comparison between the article types in the rate of exploratory-only analysis showed a statistically significant difference, in the expected direction: $x^2(1, N = 510) = 7.89$, $p = 0.005$

5.3.5.3. Differences in Proportion of Clear Distinctions between Exploratory and Confirmatory Aspects

Although this was also run for the full comparative sample, the results reported here are only for the comparative sample with the papers that only contained exploratory analysis excluded. These have also been excluded from the subsequent analyses about the nature of the exploratory analysis. The median proportion of articles that had a clear distinction between confirmatory and exploratory analysis was higher for the RRs (1), than for the SRs (0.83). A Mann-Whitney U test showed that there was a significant difference between the article types in the proportion that clearly distinguished between confirmatory and exploratory ($U = 6931.0$, $p < 0.001$).

5.3.5.4. Differences in Nature of Exploratory Analysis

The median proportion of follow-up exploratory analysis was considerably higher among RRs (0.79) than in SRs (0.25). However, the difference between the article types was not statistically significant: $U = 5447.0$, $p = 0.51$. The median proportions for the inclusion of general exploration were similar between SRs and RRs (0 vs. 0.03), and comparative analysis using a Mann-Whitney U test showed that there was not a statistically significant difference between the article types in the proportion of general exploration included ($U = 5102.50$, $p = 0.80$). Additionally, the median proportion of exploratory analysis coded as Unclear was 0 for both SRs and RRs. Comparative analysis using a Mann-Whitney U test showed no statistically significant difference between the article types in the proportion of unclear exploratory analysis: $U = 5055$, $p = 0.28$

5.3.6. Inclusion of Replication Studies

5.3.6.1. Differences in Inclusion of <u>Any</u> Replications

The full comparative sample was again used to compare rates of replications ($n = 510$). The mode for whether any replications had been included in the paper was 0 for both the SRs and the RRs. Frequencies showed that the SRs were less likely to include replication studies (31.18%), compared with the RRs (50%). Comparative analysis using chi-square showed that the difference between the article types in the inclusion of replications, was statistically significant and in the expected direction: $x^2(1, N = 510) = 17.14$, $p < 0.001$

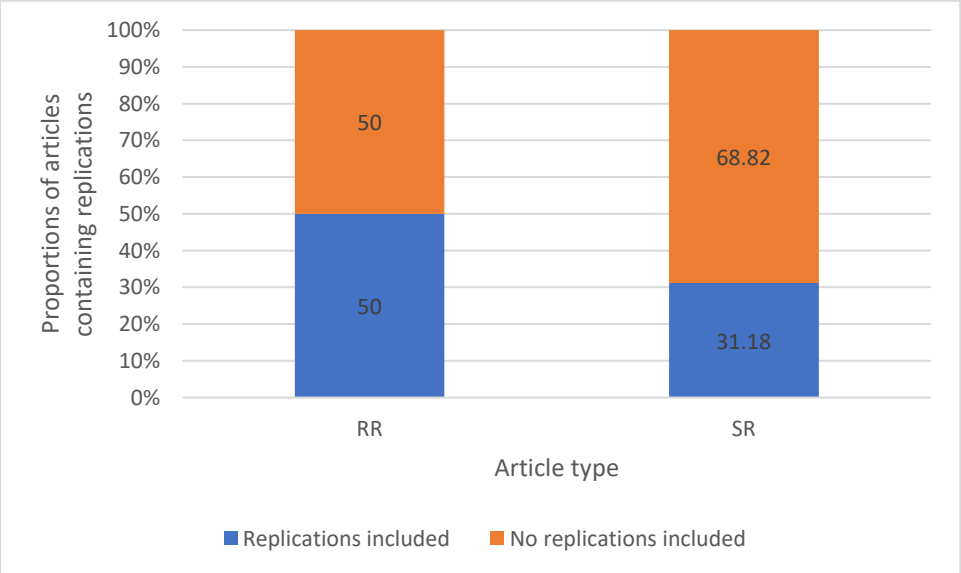

Figure 15: Bar chart showing the proportions of articles that contain replications, per article type. The figure shows that while RRs were as likely as not to contain some form of replication study, less than one-third of SRs included such studies.

Also considered was the type of replication study (direct external, indirect external, direct internal, and indirect internal) and whether inclusion of each of these kinds differed between the two article types.

5.3.6.2. Differences in Types of Replication Studies Included

Frequencies showed that only 2.94% of the SRs contained a direct external replication, compared with 28.24% of the RRs. Chi-square comparison showed that the difference between the article types in whether direct external replications were included, was statistically significant: $x^2(1, N = 510) = 71.94$, $p < 0.001$. Similarly, only 13.24% of SRs included an indirect external replication, compared with 21.77% of RRs and the difference between the article types was also statistically significant: $x^2(1, N = 510) = 6.11$, $p = 0.01$. However, the result for whether direct internal replications were included was found to be similar between the two article types, with 6.18% of SRs and 5.29 % of RRs containing a direct internal replication, and chi-square analysis confirmed that there was no significant difference between the article types in whether they contained any direct internal replications: $x^2(1, N = 510) = 0.16$, $p = 0.69$. Finally, rates for indirect internal replications were examined. Contrary to the pattern in the other replication-related analyses, SRs were more likely to contain this type of replication study than RRs were: 15% of SRs contained an indirect internal replication, compared with just 3.53% of RRs. Comparative analysis using chi-square showed that the difference between article types in the rate of indirect internal replications was statistically significant: $x^2(1, N=510) = 15.02$, $p < 0.001$

5.3.7. Methods Sections Word Counts

Descriptively, the mean word count of the RRs ($M = 1556.06$, $SD = 1069.79$) was higher than the SRs ($M = 1222.74$, $SD = 789.68$), as was the median (1330.67 vs 1039.0). A Mann-Whitney test found a statistically significant difference between the two article types ($U = 34028.50$, $p < 0.001$).

5.3.8. Use of Manipulation Checks

5.3.8.1. Differences in Inclusion of Manipulation Checks

Frequencies showed that the majority of SRs (81.77%) did not contain any manipulation checks, compared with 66.47% of RRs which did not contain any. A chi-square analysis showed that there was a statistically significant difference between article types in whether manipulation checks had been used: $x^2(1, N = 510) = 14.82$, $p < 0.001$.

5.3.8.2. Differences in Proportion of Manipulation Checks Passed and Failed

The median proportion of manipulation checks passed was 1 for both RRs and SRs. The mean proportion was higher among SRs than in RRs ($M = 0.95$, $SD = 0.20$ vs. $M = 0.77$, $SD = 0.37$) but as the distribution was highly skewed, the median was preferred as the measure of central tendency. The valid sample for this analysis only consisted of those articles that actually had at least one manipulation check included i.e., 62 SRs and 57 RRs. A Mann-Whitney U test showed that there was a significant difference between the article types in the median proportion of manipulation checks that had been passed ($U = 1288.50$, $p < 0.001$).

The median proportion of manipulation checks failed was 0 for both SRs and RRs. The SRs had a mean of 0.02 ($SD = 0.09$), while the RRs had a mean of 0.10 ($SD = 0.23$). A Mann-Whitney U test showed that there was a significant difference in the proportion of failed manipulation checks between the two article types: $U = 2078.0$, $p = 0.003$.

5.4 Discussion

5.4.1. Recap of Results

The median sample size was higher in the RRs than the SRs, although this difference was only statistically significant for the intended and pre-exclusion sample sizes, not for the post-exclusion sample sizes. RRs were significantly more likely than SRs to have a sampling plan and this was true for both the use of any sampling plan, and for the use of a statistical sampling plan. This same pattern was also observed for both power analysis and Bayesian sampling, but not for ‘other’ sampling plans, for which there was no significant difference between the article types. There was also no significant difference between the article types in the average number of studies included in each paper. RRs were more likely than SRs to include exploratory analysis, and this was more likely to be clearly distinguished from confirmatory analysis in RRs than in SRs. However, there was no significant difference between the article types in the types of exploratory analysis included.

Overall, RRs were more likely than SRs to contain replication studies. When broken down by type of replication only external replications (direct or indirect) were more common in RRs than in SRs. There was no significant difference in the inclusion of direct internal replications, while indirect internal replications were more likely in SRs than in RRs. RRs had a higher average word count than SRs, and this difference was statistically significant. Finally, RRs were significantly more likely than SRs to include manipulation checks (or equivalent), but these were more likely to be passed in SRs than in RRs.

5.4.2. Discussion of Results

Like the results reported by Soderberg et al. (2021), means for the sample sizes were also much higher for SRs than RRs, but medians were considered the more indicative measure since the wide variation in sample size led to a non-normal distribution. The findings of the current study broadly support those reported by Soderberg et al. in terms of the median sample size in RRs being larger than that of SRs. However, in the current study this was only true for the planned and pre-exclusion sample size but not for the post-exclusion sample size, and it is unclear whether Soderberg's measurement of the sample size is more equivalent to this current study's use of pre-exclusion or post-exclusion sample sizes. In either case the median sample size reported by Soderberg for the RR studies was larger than that found for RRs in the current study. This may be because the results from Soderberg's study were based only on the last study reported in the paper, whereas the current study included all studies reported in each paper in the calculation of the average sample sizes. Therefore, earlier studies such as pilot studies with much smaller sample sizes may have brought this average score down considerably compared to the final study only. The median (post-exclusion) sample size found for the SRs was slightly larger in the current study than in Soderberg et al.'s. They did report higher sample sizes for their author-matched SRs, but their journal-matched SRs were much more comparable to the SR sample in the current study. It is noteworthy that Soderberg et al. found a larger sample size in the author-matched SR (118) compared with the journal-matched SR (78), as this may suggest that standards and practices around sample size determination are influenced by the author involved, not just the article type being used, as particular authors may give more attention or care to the sample size regardless of the format, especially if they are particularly aware of the need for sufficiently powered studies.

The current study also noted a clear difference between the article types regarding the use of a sampling plan. Rates of the use of any sampling plans were slightly lower among the RRs

in this study compared with the rates reported by Soderberg et al. (80.59% in the current study compared with 90% in their study). Interestingly the rates of this for the SRs was higher in the current study than that reported by Soderberg for either of their SR comparators (47% in the current study compared with only 17% for their journal-matched SRs and 29% for their author-matched SRs). Due to the relative flexibility of an 'other' sampling plan and how this can be interpreted, it may be that the coding of this in the current study was more inclusive than in the approach used by Soderberg, thus potentially explaining the differing findings. However, without sufficient details about their process of coding this characteristic, this is impossible to be sure of.

The rates of statistical sampling plans in the current study were much lower overall than the use of any kind of sampling plan. Rates of statistical sampling plans for the RRs are slightly higher than those reported by Norsa (2022) who found that although 40% of those RRs did not specify the approach used to determine their planned sample size, 36.95% of the RRs they examined were found to have run a well-planned power analysis based on the existing literature and pilot studies. According to Norsa (2022) the majority of the RRs (65.22%) did not specify what software had been used for their power analysis, an issue that may hinder efforts to reproduce those analyses if needed. This also reflects the challenges experienced in the current study, as it was frequently not clear how sample size planning had occurred or (where a calculation had been done) what software was used for this.

The current study found higher average word counts in the methods sections of RRs than of SRs, which appears to be a novel finding which has not been previously investigated. Likewise, the similar numbers of studies in RRs and SRs does not seem to have been previously investigated, though this was slightly lower that than reported by Sassenberg & Didrich (2019) who found a mean of 2.81 studies per article (range 1 to 10) within the general psychology literature, with suggestions that the number of studies may have increased over time in their sample, although this did vary slightly between different journals.

The extent to which RRs include non-registered studies also does not appear to have been previously examined, and although the finding that SRs are more likely than RRs to have included these is perhaps obvious, the finding that 9.41% of the RRs contained such studies helps to clarify exactly what RRs typically consist of and to what extent such studies are included in the final stage 2 report. Likewise, the finding that the majority of RRs include some form of exploratory analysis, compared to only around half of SRs, does not appear to

have been previously reported, but helps to indicate what a typical RR can look like while also demonstrating that perceptions of RRs hindering exploratory analysis are not justified. These findings may also be reflective of the greater clarity RRs enforce about the distinction between confirmatory and exploratory elements of the paper, whereas the rates for the SRs may be undermined by selective reporting of exploratory analysis as if it were confirmatory. It is therefore possible that exploratory analysis is not in fact actually higher among the RRs than the SRs but that as far as could be determined, the reporting in the papers indicated that this was so.

Rates of replications reported by Scheel indicate that these were present in 57.75% of RRs and only 2.63% of SRs. However, they did not include indirect (conceptual) or internal replications and so their findings are most comparable to the rates of direct external replications found in the current study. The current study found a similar rate of replications in the SRs, at only 2.94%, but the rate of this type of replications in RRs were much lower than that found by Scheel, at only 28.24%. This may potentially be explained by the fact that their sample was more limited in scope than that used in the current study. For example, the current study's sample includes many articles from the Journal of Medical and Internet Research (JMIR) which, because of its focus on applied healthcare research rather than experimental studies, does not appear to publish many replication studies. It is possible that other included journals happen to have a similar scope in terms of their articles. This may have led to the lower overall proportion of replication studies within this larger sample compared with Scheel's more focused sample.

Finally, manipulation checks were found in approximately 33.53% of the RRs and only 18.24% of the SRs. Previous studies have reported rates of 33% (Hauser et al., 2018) and even up to 57.25% for the use of manipulation checks in SRs although this latter figure was acknowledged to be higher than would be typical for many other journals (Ejelov & Luke, 2020). It is somewhat surprising that these findings from the general experimental psychology literature are similar and/or higher than the rates of manipulation checks found for the RRs in the current study, as RRs might have been expected to make greater use of manipulation checks than the general literature. This difference could indicate an underestimation of the manipulation checks in the current study, particularly given the high levels of subjectivity felt when coding these. However, as outlined above regarding replication studies, the wider scope of the current sample and the inclusion of a greater diversity of topics and designs than just experimental research may also help to explain why

manipulation checks were found in a relatively smaller proportion of the overall sample than might be expected based on the previous studies that focused more on experimental psychology. While efforts were made to include any equivalent type of check when coding this characteristic, it may be that such measures are just less common in other types of research and so the overall proportions were influenced by this.

5.4.3. Limitations

It is important to acknowledge that many of the characteristics included in this chapter were highly subjective to interpret and to code. In many cases, coding decisions may have been influenced by the fact that the coding was being done by someone not intimately familiar with the minutiae of the designs and topics for many of the studies and so, there is considerable potential for misinterpretation and misunderstanding. With this in mind, these results should be considered as an initial exploration and more expert confirmation of the accuracy of this coding would be beneficial. However, clearer or more explicit communication about such methodological characteristics within the articles would have lessened the uncertainty around the steps taken by the authors, and such clarity should be considered a reasonable expectation for a published journal article.

Furthermore, many of these characteristics are based on the traits of all of the studies in the papers rather than only the specific RR study and a comparable (e.g., last) study within the SR paper which was the approach used in other studies. This means that the characteristics as they are coded here reflect all studies within each article type, not just the unique RR reviewed studies. Arguably, it may be more appropriate to only consider the specific studies within an RR paper that have actually be reviewed and were subject to in-principle acceptance. However, as readers will likely read and evaluate most papers as a whole, determining the overall traits of the paper across all the different studies included in it may be a more useful approach. Understandably though, the use of this approach may make comparison of the findings with other studies of RRs more challenging because they tend to only use the specific RR-study (or last study in the paper) for their work.

The fact that exploratory analysis is common in RRs and that this is typically clearly distinguished from the confirmatory analysis is encouraging and hopefully provides reassurance that RRs are not, as have been suggested by some, likely to inhibit exploration. The fact that this is indeed more common in RRs than in SRs (at least as far as it can be judged based on what is reported) particularly undermines this criticism. It is important to

acknowledge though that the comparatively low rates of exploratory analysis in the SRs could be driven by selective reporting such as HARKing so that analyses that were actually exploratory were presented in the papers as being confirmatory and so, based on the information provided in those papers, this would have been coded as not having exploratory analysis (as far as could be determined) when this may not in fact be the case. This is a possibly inevitable challenge when coding articles from the general research literature due to the ever-present potential for biased reporting.

Finally, identifying what did and did not count as a manipulation check was often challenging, and so I aimed to be as inclusive as possible. However, some degree of subjectivity seemed to be unavoidable. It is also possible that some studies may have used such checks but not reported them, considering them as one of the kinds of informal practices described by Brenninkmeijer et al. (2019).

5.4.4. Implications and Future Directions

These results demonstrate that although RRs typically outperform SRs in their use of practices and standards associated with greater rigour and/or transparency, there is still considerable room for improvement in this. For example, the inclusion of some kind of sampling plan or justification should reasonably be possible to include in any study even if a statistical plan has not been created. The freedom to use an 'other' sampling plan based on relevant constraints or other causes, makes it particularly open for authors to use as even a loose justification for their sampling approach and it is reasonable to expect that this should be possible to specify before the study is conducted.

Furthermore, the lower rates of some characteristics within the RRs compared to what might be expected may reflect the increased diversity of RRs beyond experimental research designs. While this is only speculation, future research could explore these changing uses in more depth.

5.5. Conclusion

Overall, the evidence presented in this chapter indicates that RRs are mostly associated to a greater extent with characteristics indicating greater rigour, quality, and transparency, than SRs are. This is particularly the case for characteristics that have received considerable attention in the context of the replication crisis, such as sample size and use of sampling plans, clarity of the distinction between confirmatory and exploratory analyses, and the inclusion of replication studies. While there is still some room for improvement and

furthermore, some of these findings are based on highly subjective judgements by a non-expert, the evidence presented here constitutes a promising insight into the prevalence of these practices in both RRs and SRs, and it is hoped that this may provide a useful foundation for future research to replicate and extend these findings in a more robust manner.

5.6. References

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., Fiedler, S., Funder, D. C., Kliegl, R., Nosek, B. A., Perugini, M., Roberts, B. W., Schmitt, M., van Aken, M. A. G., Weber, H., & Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108-119. https://doi-org/10.1002/per.1919

Bakker, M., van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi-org.abc.cardiff.ac.uk/10.1177/1745691612459060

Bakker, M., Veldkamp, C. L. S., Akker, O. R. van den, Assen, M. A. L. M. van, Crompvoets, E., Ong, H. H., & Wicherts, J. M. (2020). Recommendations in pre-registrations and internal review board proposals promote formal power analyses but do not increase sample size. *PLOS ONE, 15* (7), e0236079.https://doi.org/10.1371/journal.pone.0236079

Brenninkmeijer, J., Derksen, M., & Rietzschel, E. (2019). Informal laboratory practices in psychology. *Collabora: Psychology*, *5*(1), 45. https://doi.org/10.1525/collabra.221

Buskist, W. & Johnston, J. M. (1988). Laboratory lore and research practices in the experimental analysis of human behavior. *The Behavior Analyst*, *11*(1), 41-42.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14* (5), 365–376. https://doi.org/10.1038/nrn3475

Carsten, M., Clapp-Smith, R., Haslam, S. A., Bastardoz, N., Gooty, J., Connelly, S., & Spain, S. (2023). Doing better leadership science via replications and registered reports. *The Leadership Quarterly*, *34*(4), 101712, https://doi.org/10.1016/j.leaqua.2023.101712.

Chambers, C. D. (2015). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 10-11.

Chambers, C. D., Dienes, Z., McIntosh, R. D., Rotshtein, P., & Willmes, K. (2015). Registered reports: Realigning incentives in scientific publishing. *Cortex, 66,* A1-A2.

Chambers, C. (2017). *The seven deadly sins of psychology*. Princeton University Press.

Chester, D. S., & Lasko, E. N. (2021). Construct validation of experimental manipulations in social psychology: Current practices and recommendations for the future. *Perspectives on Psychological Science, 16*(2), 377–395.

Ebersole, C. R., Mathur, M. B., Baranski. E., Bart-Plange, D.-J., Buttrick, N. R., Chartier, C. R., Corker, K. S., Corely, M., Hartshorne, J. K., IJzerman, L. B., Lazarevic, L. B., Rabagliati, H., Ropovik, I., Acxel, B., Aeschbach, L. F., Andrighetto, L., Arnal, J. D., Arrow, H., Babincak, P., & Nosek, B. A. (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. Doi:10.1177/2515245920958687

Ejelov, E. & Luke, T. J. (2020). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology, 87,* 103937. https://doi.org/10.1016/j.jesp.2019.103937

Fayant, M.-P., Sigall, H., Lemonnier, A., Retsin, E., & Alexopoulos, T. (2017). On the limitations of manipulation checks: an obstacle toward cumulative science. *International Review of Social Psychology, 30,* 125–130. Doi: 10.5334/irsp.102

Fidler F., Singleton Thorn F., Barnett A., Kambouris S., Kruger A. (2018). The epistemic importance of establishing the absence of an effect. *Advances in Methods and Practices in Psychological Science, 1*, 237–244. 10.1177/2515245918770407

Gattiker, T. F., Hartmann, J., Wynstra, F., Pagell, M., Cantor, D., Yan, T., & Tate, W. (2021). Testing the shoulders of giants—Replication research using registered reports. *Journal of Supply Chain Management, 58*(3), 89-94.

Gelman, A. & Carlin, J. (2014). Beyond power calculations: Assessing type S (Sign) and type M (Magnitude) errors. *Perspectives on Psychological Science, 9*(6), 641-651.

Goldacre, B. (2009). *Bad science*. Fourth Estate.

Hauser, D. J., Ellsworth, P. C., & Gonzalez, R. Are manipulation checks necessary? *Frontiers in Psychology, 9,* 998. Doi: 10.3389/fpsyg.2018.00998. PMID: 29977213; PMCID: PMC6022204.

Higgs, M.D., & Gelman, A. (2021) Research on registered report research. *Nature Human Behaviour, 5,* 978–979. https://doi-org.abc.cardiff.ac.uk/10.1038/s41562-021-01148-y

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196–217. https://doi-org.abc.cardiff.ac.uk/10.1207/s15327957pspr0203_4

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., 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.abc.cardiff.ac.uk/10.1027/1864-9335/a000178

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., 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.

Kriegeskorte N., Simmons W. K., Bellgowan P. S. F., Baker C. I. (2009). Circular analysis in systems neuroscience: The dangers of double dipping. *Nature Neuroscience, 12,* 535–540.

Lakens, D. (2021). *Sample Size Justification* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/9d3yf

Lench, H. C., Taylor, A. B., & Bench S. W. (2014). An alternative approach to analysis of mental states in experimental social cognition research. *Behavior Research Methods, 46,* 215–228. 10.3758/s13428-013-0351-0

Lindsay, D. S. (2019). Swan song editorial. *Psychological Science, 30*(12), 1669–1673. https://doi-org.abc.cardiff.ac.uk/10.1177/0956797619893653

Logg, J. M. & Dorison, C. A., (2021). Pre-registration: weighing costs and benefits for researchers. *Organizational Behaviour and Human Decision Processes, 167,* 18-27.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*(6325), 584–585.https://doi.org/10.1126/science.aal3618

Marszalek, J. M., Barber. C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331-348. Doi: 10.2466/03.11.PMS.112.2.331-348.

Mathieu, S., Boutron, I., Moher, D., Altman, D. G., & Ravaud, P. (2009). Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA*, *302*(9), 977-84. Doi: 10.1001/jama.2009.1242.

Meehl P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66,* 195–244. 10.2466/pr0.1990.66.1.195

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534

Norsa, R. (2022). *Do Registered Reports Improve Sample Size Planning in Psychology? An exploratory study*. Thesis. Padua Thesis and Dissertation Archive. https://thesis.unipd.it/handle/20.500.12608/32381

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., van 't Veer, A. E., & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences, 23*(10), 815-818.

Nosek, B. A. & Errington, T. M. (2020) What is replication? *PloS Biology, 18*(3), e3000691.

Nosek, B. A., & Lakens, D. (2014). Registered Reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137-141.

Nosek, B. A. & Lindsay, D. (2018). Preregistration becoming the norm in psychological science. *APS Observer*. Retrieved from: https://www.psychologicalscience.org/observer/preregistration-becoming-the-norm-in-psychological-science

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872. https://doi.org/10.1016/j.jesp.2009.03.009.

Ramagopalan, S. V., Skingsley, A. P., Handunnetthi, L., Magnus, D., Klingel, M., Pakpoor, J., & Goldacre, B. (2015). Funding source and primary outcome changes in clinical trials registered on ClinicalTrials.gov are associated with the reporting of a statistically significant primary outcome: a cross-sectional study. *F1000Research, 4*, 80. Doi: 10.12688/f1000research.6312.2. PMID: 26069729; PMCID: PMC4431380.

Sassenberg, K. & Didrich, L. (2019). Research in social psychology changed between 2011 and 2016: Larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science, 2*(2), 107-114. Doi:10.1177/2515245919838781

Scheel, A.M., Schijen, M.R.M.J., & Lakens, D. (2021). An excess of positive results: comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Scheel AM, Tiokhin L, Isager PM, Lakens D. (2020). Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science, 16*(4):744-755. Doi: 10.1177/1745691620966795.

Scheel, A. M. (2022). Why most psychological research findings are not even wrong. *Infant and Child Development, 31*(1), e2295.

Schoenherr, T., & Swink, M. (2012). Revisiting the arcs of integration: Cross-validations and extensions. *Journal of Operations Management, 30*(1), 99–115. https://doi-org.abc.cardiff.ac.uk/10.1016/j.jom.2011.09.00

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S. A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L. L., Diermeier, D., Heinze, J. E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J., Davis-Stober, C. P., Plessis, C., Gronau, Q. F., Hafenbrack, A. C., Eko Yi Liao, E. Y., … & Uhlmann, E. L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology*, *66*, 55-67, https://doi.org/10.1016/j.jesp.2015.10.001.

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.

Simons, D. J., Holcombe, A. O., & Spellman, B. A. (2014). An introduction to Registered Replication Reports at Perspectives on Psychological Science. *Perspectives on Psychological Science, 9*(5), 552–555. http://www.jstor.org/stable/44290039

Soderberg, C.K., Errington, T.M., Schiavone, S.R., Bottesini, J., Thorne, F. S., Vazire, S., Esterling, K. M. & Nosek, B. A. (2021). Initial evidence of research quality of registered reports

compared with the standard publishing model. *Nature Human Behaviour, 5*, 990–997. https://doi-org/10.1038/s41562-021-01142-4

Spellman B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*, 886–899. https://doiorg.abc.cardiff.ac.uk/10.1177/1745691615609918

Srivastava, S. (2016, August 18). Lots of us, probably all of us, have 'lab lore' [Tweet]. Retrieved from https://twitter.com/hardsci/status/766317950059945985

Strack, F., Martin, L. L., and Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*, 768–777. Doi: 10.1037/0022-3514.54.5.768

Tenney, E. R., Costa, E., Allard, A., & Vazire, S., (2021). Open science and reform practices in organizational behavior research over time (2011 to 2019). *Organizational Behavior and Human Decision Processes, 162*, 218-223, https://doi.org/10.1016/j.obhdp.2020.10.015.

Toth, A. A., Banks, G. C., Mellor, D., O'Boyle, E. H., Dickson, A., Davis, D. J., DeHaven, A., Bochantin, J., & Borns, J. (2021). Study Preregistration: An evaluation of a method for transparent reporting. *Journal of Business and Psychology, 36*, 553-371.

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: The case of psi: Comment on Bem (2011). *Journal of Personality and Social Psychology*, *100*(3), 426–432. https://doi-org.abc.cardiff.ac.uk/10.1037/a0022790

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 632-638

Wagenmakers, E.-J. & Dutilh, G. (2016). Seven selfish reasons for preregistration. *APS Observer*, *29*(9), 13–14.

Wagge Jordan R., Brandt Mark J., Lazarevic Ljiljana B., Legate Nicole, Christopherson Cody, Wiggins Brady, Grahe Jon E. (2019). Publishing research with undergraduate students via replication work: The Collaborative Replications and Education Project. *Frontiers in Psychology, 10*, 247. https://doi.org/10.3389/fpsyg.2019.00247

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and

reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832. 10.3389/fpsyg.2016.01832

Wilson, A. (2014, May 26). Psychology's real replication problem: our Methods sections. Retrieved from https://psychsciencenotes.blogspot.com/2014/05/psychologys-real-replication-problem.html (Accessed 02/02/2023).

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science, 13*(4), 411-417. Doi: 10.1177/1745691617751884. PMID: 29961410.

Chapter 6: Comparative Analysis of Author Demographics

6.1. Introduction

6.1.1. Author Collaboration

In recent years there has been a move away from the idea of research as being the work of individual scientists or the 'lone genius' model of science. Instead, there have been calls for increased collaboration, and greater recognition of this collaborative approach (Frith, 2020; Kiser, 2018; Forscher et al., 2022; Ledgerwood et al., 2022; Patel et al., 2021; Myers et al., 2021; Green et al., 2019; Clark, 2017). This is in line with general trends of greater numbers of authors per paper in many disciplines (Nabout et al., 2015; Wang et al., 2016; Wuchty et al., 2007). Potential benefits of collaboration can include more diverse and interdisciplinary perspectives, greater sharing of resources and skills, broader distribution of the workload, greater transparency, increased inclusivity, access to larger sample sizes, and potentially a greater chance of catching any errors in the paper. It can also allow researchers to conduct research that is more impactful, more ambitious, and more frequently cited (Wuchty et al., 2007; Uhlmann et al., 2019).

Efforts to increase big team science and crowdsourcing have been noted in other disciplines such as biomedical science (Errington et al., 2021; Budge et al., 2015; Clancy et al., 2013; Saez-Rodriguez et al., 2016), pre-clinical animal research (Multi-PART, n.d.) and clinical research (Custovic et al., 2015) as well as through multiple initiatives in psychology and related disciplines. For example, the Framework for Open and Reproducible Research Training (FORRT) community provides a platform for researchers and educators interested in topics such as research transparency, reproducibility, rigor, ethics, and open scholarship approaches to collaborate on research and creation of teaching resources (Azevedo, 2019). Elsewhere, the Pipeline Project (Schweinsberg et al., 2016) used a 'Pre-Publication Independent Replication' approach in which 25 research groups conducted replications of effects regarding moral judgements, before these findings had been published, thereby facilitating an international team approach.

Many such initiatives use the RR format to publish their work. For example, the Psychological Science Accelerator (PSA) is an international collaborative network of psychology researchers and labs which conducts large multi-site studies in psychology (Chartier et al., 2018) and this has relied on the RR format to register and publish their

studies. Their approach enables them to obtain much larger and more diverse samples, thereby increasing the power of their analyses, while also maintaining an open and inclusive approach to collaboration. Similarly, The Registered Replication Report format at Advances in Methods and Practices in Psychological Science attempts to use the RR format to enable collections of replications studies to be conducted using the same approved protocol, thereby encouraging multi-site/multi-team approaches (APS, n.d.). The Many Labs project (Klein et al., 2014) and its various sequels (e.g., Klein et al., 2018; Ebersole et al., 2020), also took this approach using the RR format for their replication initiative in a special issue of the journal Social Psychology in 2014.

As some of these team science initiatives use the RR format, the publications arising from their own collaborations will undoubtedly contain large numbers of authors. However, these constitute only a small proportion of all of the stage 2 RRs that currently exist and it is unclear whether stage 2 RRs as a whole are associated with larger numbers of authors than the standard research literature.

6.1.2. Author Seniority

Whether RRs are likely to increase collaboration or not, a widespread concern regarding the format has been whether the RR approach is inaccessible to early career researchers (ECRs), such as PhD students, postdoctoral researchers, or those on short-term contracts (Allen & Mehler, 2019; Morey & Tzavella, 2018). Such concerns typically relate to the perceived length of time required to complete the RR process and the high minimum power requirements for RRs at many journals which may be challenging to achieve given the limited time and resources available to such ECRs to complete their projects (Parker et al., 2019; Maizey & Tzavella, 2019; Allen & Mehler, 2019). Recent innovations such as Peer Community In (PCI) Registered Reports (PCI Registered Reports, n.d.; Eder & Frings, 2021) and their scheduled review track have helped to address this by accelerating the stage 1 review process. Expedited peer review has also been used previously for stage 1 RRs such as the COVID-19 Registered Reports initiative (Chambers, 2020; Chambers & Dunn, 2022). Furthermore, careful planning and time management may help to ensure that many projects fit into the required timescale although this may not be possible in all cases (COS, n.d.; Parker et al., 2019; Chambers, 2019; Ludwig, 2019; Kathawalla et al., 2021; Kiyonaga & Scimeca, 2019). Additionally, anecdotal evidence from some early career researchers has indicated that the format can be feasible for those at this early stage of their careers (Henderson, 2019). However, another potential issue is the importance of being familiar and

experienced with the research paradigm before submitting a RR for a study. The challenges of achieving this within the typical length of a PhD in the UK may hinder the ability of students to use this format effectively (Morey & Tzavella, 2018; Maizey & Tzavella, 2019).

An additional barrier that has been highlighted is the need for early career researchers to publish in the most high-profile journals possible and that most of these journals in psychology and other fields do not offer the RR format (COS, n.d.). As increasing numbers of journals adopt RRs, it is hoped that this barrier may become less of a concern but in the meantime, it may impact on ECRs' willingness to use the format.

However, it has been suggested that ECRs are generally more likely to be open to, and advocate for, the use of such a novel publishing format, as well as other innovative and open approaches generally (O'Brien et al., 2019; Toribio-Florez et al., 2021; DeHaven et al., 2019). If so, ECRs may in fact be more likely than more senior researchers to use or at least be supportive of the RR format. In support of this view, Chambers (2019) reported that at Cortex, 78% of first authors of RRs were PhD students or postdocs, compared with only 67% in a control sample of standard articles. This finding is significant in providing clear evidence that RRs are being published by ECRs and that they are not any less likely to publish RRs than more senior authors are. However, these figures are based on a limited sample at only one journal and so it is unclear whether this is also true more generally.

Conflicting with this view, evidence suggests that ECRs typically have little awareness of open science practices, and that when they are aware of them, their desire to engage in new practices is constrained by their environments and circumstances, particularly by the 'publish or perish' culture (Nicholas, Rodríguez-Bravo et al., 2017; Nicholas, Watkinson, et al., 2017). Therefore, access to training, resources and support is essential for early career researchers to increase their awareness and skills in relation to open science practices, but larger-scale institutional barriers and academic incentive structures may still hinder their uptake of such practices.

Overall, the matter of how accessible RRs are to ECRs is currently unclear, with concerns mainly based on speculation rather than empirical evidence. As this is one of the most common criticisms of the RR format, it is important to clarify the actual proportions of early career researchers publishing RRs and how this compares to rates within the standard research literature.

6.1.3. Lack of Geographic Diversity in Authorship

Major concerns have been expressed regarding the lack of geographic diversity among authors of the literature in psychology and other disciplines, with the US dominating the field, closely followed by Europe (particularly the UK), while countries in regions such as Africa, South America, and Asia (often referred to as the Global South, or Majority World) are vastly underrepresented (Lin & Li, 2022; O'Gorman et al., 2012; Piocuda et al., 2015; Thalmayer et al., 2021; Cheek, 2017; Naidoo et al., 2021). For example, analysis of research published in leading medical and global health journals since 2018 showed that first and last authors from high-income countries were 19 times as prevalent as authors from low-income countries (Merriman et al., 2021). Although the geographic diversity of authors in psychology has increased in recent years, including greater international collaboration, there remains much more room for improvement (Lin & Li, 2022; O'Gorman et al., 2012; Piocuda et al., 2015; Haslam & Kashima, 2010; Adair et al., 2012; Webster et al., 2021; Lund, 2022).

This lack of representation of countries from the Global South or 'Majority World' is concerning as greater cultural and geographic diversity of authors and journal editors has been highlighted as being important in addressing the WEIRD sampling problem in psychology, whereby the overwhelming majority of research participants are Western, Educated, Industrialised, Rich and Democratic and so their views and thought processes are not universal (Henrich et al., 2010; Cheon et al., 2020; Apicella et al., 2020). This issue is also linked to the need to decolonise psychology (Dege & Strasser, 2021; Adams et al., 2017), and to the issues and inequalities resulting from English being the lingua franca of science (Suzina, 2020; Montgomery, 2009; O'Neil, 2018). There is, however, criticism of the concept of WEIRD sampling, including the fact that is used as a catch-all term to indicate a lack of diversity despite the fact that it does not explicitly acknowledge important considerations such as ethnicity or religion. Furthermore, it is not made clear how the original creators of the WEIRD acronym identified these five dimensions as being the most important domains, and its lack of specificity has led to it being over-applied without sufficient nuance given (Dutra, 2021; Syed, 2021). Nevertheless, critics of the term WEIRD do express deep concern for the lack of diversity in sampling and authorship and acknowledge Henrich et al.'s work in creating this acronym as being important in raising awareness of this critical issue in psychological research.

While many open science approaches advocate for increased accessibility and representation, including for greater geographic diversity (e.g., Barch, 2021), there has also been concern

that some open science efforts have in fact reinforced such inequalities, as seen in the open access movement (Burgman, 2019; Ross-Hellauer, 2022). For example, Smith et al. (2021) reported that the geographic diversity of authors of open access articles was significantly lower than that of articles that had not been published open access and that most open access articles were written by authors in high-income countries. Their results supported the view that article publishing charges (APCs) are a barrier to open access publishing for researchers from the Global South, a finding which is also supported by Mekonnen et al. (2021). Although alternatives such as preprints are valuable in enabling researchers with little funding to share their work publicly, these do not carry the same benefits for researchers in terms of hiring and promotion.

Open science approaches appear to have had a mixed impact at best on the representation and inclusion of authors from non-Western countries. It is unclear as yet how RRs may have impacted on this issue, and so it is important to understand who is publishing RRs and whether (and to what extent) these include authors based in non-Western countries.

6.1.4. Research Questions

In summary, it is unclear who is publishing RRs and whether this format has been associated with greater rates of collaboration and representativeness. This is important to understand to determine how accessible this format is to researchers of different demographics, and to inform efforts to improve in the future. The research described in this chapter attempts to answer these questions by coding and comparing the number and characteristics of the authors of RRs and their matched standard articles. The research questions are outlined in Table 7 below.

Table 7

Overview of research questions and hypotheses for chapter 6

Research Question	Hypothesis
RQ 1: How does the seniority of the authors differ between RRs and SRs? - RQ1a: Are there differences between article types in the rates of authors that have a PhD? - RQ1b: Are there differences between article types in the rates of authors that are early career researchers?	No specific hypothesis was stated for RQ1.

RQ 2: How does the geographic diversity of authors differ between RRs and SRs? - RQ2a: Are there differences between article types in whether the articles included any authors from non-Western countries? - RQ2b: Are there differences between the article types in the number of non-Western countries represented in the authors' affiliations? - RQ2c: Are there differences between the article types in the proportion of non-Western countries represented in the authors' affiliations?	No specific hypothesis was stated for RQ2.
RQ 3: Does the number of authors per paper differ between RRs and SRs?	H3: It was tentatively hypothesised that RRs may have a larger number of authors than SRs as they may be associated with greater collaboration.

6.2. Methods

6.2.1. Methods for Coding and Analysis of Author Seniority

6.2.1.1 Initial Coding of Author Seniority

As described in previous chapters, content analysis was used to code the data, this time on the author demographics for each of the included articles. Due to the excessive time required to gather data regarding all of the authors on each paper, the gathering of data about the authors' job title was initially restricted to the first three authors and the last author, with the first and last author being of most interest. While the meaning of last authorship is not always consistent across disciplines or sub-disciplines, its common use as an indicator of seniority or of a particularly important supervisory role in the work justified focusing on this. For each of these authors, their job title at the time the article was published was documented in the initial coding sheet. To gather this information, I searched online for the author, particularly in combination with the name of the institution given as their affiliation in the article. I especially sought out profile pages on institution websites, LinkedIn profiles, and CVs, and frequently used the term 'CV' or 'resume' in the search. In general, sites such as LinkedIn or the authors' personal websites most often contained the required details. University profile pages tended to be less informative, unless they contained the author's CV as a downloadable file, or if they contained a very detailed bio with specific years stated for their positions which was much less common than more generic or outdated bios.

One issue I anticipated in this process was that authors may have left their previous institution or moved onto a different role since working on the published project, especially for older papers. For coding purposes, I only focused on their role at the time of publication, not their role when the work was done or their current role, if these roles were different. If no information for that timeframe could be found, this was coded as 'unclear'.

If, however, I couldn't find authors' information for that time period but could find information for their current role, this was initially included in the database in red font with the word 'currently', (e.g. Currently – Assistant Professor) and specified in the notes column that this is the person's current role but their role title in the year of publication was unclear. This was thought to be potentially useful in case I could verify their information in the future, particularly in cases where papers were published quite recently and the author's current role might still be the same as their role when the paper was published. Due to time constraints during the later stages of the project, these instances were not followed up and the seniority level of these authors was instead considered to be unclear and was coded as such in the subsequent process of variable creation.

Where the author held multiple positions at the same time, efforts were made to prioritise academic job titles rather than industry or clinical roles, and if possible to ascertain whether one of their multiple roles was their primary occupation. Where there were multiple roles, the institution(s) documented as their affiliation in the published paper often helped in deciding while role to code for.

The authors' job title was initially collected as a free text entry. However, as the coding protocol was refined in the first year of this project, a number of standard response categories were created for commonly used job titles such as 'Postdoctoral researcher', 'PhD student' or 'Senior Lecturer'. These codes were used wherever possible during the vast majority of the coding process, and earlier entries were re-coded to reflect these categories wherever possible. However, in cases where the authors' job title did not fit into one of these categories, it was entered as a free text entry. This was often the case with authors based in clinical or industry roles if they did not have an additional academic affiliation clearly indicated.

6.2.1.2. Variable Creation for Whether Authors had a PhD

Based on the data gathered during the initial coding process, two new variables were created for each of the four authors included. The first of these variables attempted to provide a

measure of whether the author had a PhD. This had not been captured specifically during the initial coding process. However, it could generally be inferred from the job titles involved (e.g. undergraduate student, PhD student, Postdoctoral researcher) and in cases where this was not obvious this was checked for again. In some cases, when gathering the initial free text responses for the first round of coding, information was also recorded for whether those with industry or clinical roles also had completed a PhD in the past. However, in many other cases this information was unclear.

Where academic job titles were clear, typical free text responses gathered in the initial coding process tended to be as follows: Undergraduate student, research assistant, masters student, PhD student, Postdoctoral researcher, Research Associate, Research Fellow, Lecturer, Assistant Professor, Associate Professor, Senior Lecturer, Professor, Director, Dean, Chair. For the purposes of this analysis, undergraduate students, research assistants, masters students and PhD students were coded as 0 to indicate that they did not hold a PhD. The other job titles listed that were considered to require a PhD were typically coded as 1 for this variable unless there was some suggestion from the available information that this was not appropriate. Those for whom it was not clear whether they had a PhD were coded as N/A and were excluded from the analysis in order to focus on those with clear data for this.

As previously stated, these variables were created for the first three authors and last author, but not all papers have this many authors. In cases where the second or third author was the last author, they were coded as being the last author. When taking this approach, or when a second or third author just did not exist for the paper, the variable for the second/third author was coded as N/A and so it would later be excluded from the analysis.

On papers where the last author was a consortium authorship, the relevant variables were initially created both with and without taking into account the authors involved in the consortium, as the decision about whether to include those authors affected both the 'last author seniority' variables, and the number of authors per paper. However, as none of the articles in this chapter's comparative sample appear to have had a consortium authorship in their authorship list, this did not need to be taken into account for the analysis after all.

6.2.1.3. Variable Creation for Whether the Author was an ECR

Based on the data gathered during the initial coding process, another categorical seniority variable was created for whether the author was an ECR. Using the list of common job titles specified above, students and those on shorter research contracts (e.g. Research Affiliate,

Research Associate) were considered ECR and coded as 0, while those who were documented as being a Lecturer or above (e.g. Assistant Professor, Associate Professor, Senior Lecturer, Professor, Director, Dean, Chair.) were considered non-ECR and coded as 1. While some Lecturers and Assistant Professors can be considered to be ECRs, it was usually difficult to distinguish between this for individual cases and so a blanket approach needed to be taken instead. As per the approach taken with the PhD variable, where there were less than four authors on the paper, the final author was coded as the last author rather than as second or third, and where there were no second or third authors (e.g. for single-author papers) these were coded as N/A and so were not included in the analysis.

Despite efforts to prioritise academic job titles over clinical or industry roles during the initial coding process, in some cases it was not possible to do so, and so some industry or clinical titles had been included in that initial process. This presented a challenge when re-coding the data as it was often difficult to determine seniority within these contexts. Therefore, authors with only industry or clinical titles were coded as 99 and then were excluded from the analyses where needed. This affected a relatively small proportion of the overall sample.

6.2.1.4. Overview of Analysis of Author Seniority

The comparative sample of 170 RRs and 340 SRs described in previous chapters was used to investigate differences between the article types in this seniority of the authors. Due to the exclusion of some data from some of these analyses, the actual sample size varied per analysis and this is reported in more detail for each specific result. Details of the rates of first-author seniority in the full RR-only sample are given in chapter 9.

As the data were categorical, the descriptive statistics examined for each article type were the frequencies of responses, and the mode. Chi-square tests were used to determine whether there were differences in author seniority between the two article types.

6.2.2. Methods for Coding and Analysis of Geographic Diversity

6.2.2.1. Initial Coding of Authors' Countries

The countries of the authors were coded based on the information provided in the article about each authors' institutional affiliation. In cases where there were multiple countries (e.g., an international collaboration), all of the countries were included, in the order in which they were listed in the article (except when countries re-occurred multiple times for different authors, as each country was only coded as being included once per paper). Countries were

included for all authors, not just the maximum of four authors that were used for the seniority variables.

6.2.2.2. Variable Creation for Geographic Diversity

Based on the information gathered during the initial coding process, three variables were created. First, a categorical variable was created for whether there was a non-Western country represented in the paper, and this was coded as either 0 (no) or 1 (yes). 'Western countries' was defined as the USA, Canada, Australia, New Zealand, and European countries except for Eastern Europe and Turkey. Within the papers that included an author based in a non-Western country, the number of non-Western countries represented in that paper was coded as an additional variable. Finally, a proportion score was created to show what proportion of the countries represented in each paper were non-Western countries. Due to the limited number of papers that included any non-Western countries, the number of papers that actually had relevant data for the latter two variables was very small.

The creation of these variables did not take into account the number of authors from each country; whether there was a single author or multiple authors from a particular country, this country was only represented once within the list and subsequent calculation. Such an adjustment to the variable creation could have given a more detailed view of how well represented non-Western countries are within the samples. It is anticipated that this could have decreased the overall proportion scores for the non-Western countries even more, because if any countries are represented more than once, it is likely that this would most often be true of non-Western countries.

Furthermore, it is worth bearing in mind that the country in which authors are based at the time of publication does not necessarily reflect the countries that authors are originally from and so the variables created are a limited measure of the authors' true cultural and geographic diversity. However, due to the difficulty of finding such detailed information, particularly from publicly available information only, a more nuanced measure of author diversity was not considered feasible for this project.

6.2.2.3. Overview of Analysis of Differences in Geographic Diversity

For the analysis of whether there was a difference in whether any author was based in a non-Western country (a categorical variable), the mode and frequencies were examined. Chi-square analysis was used to compare this characteristic between the two article types. For the comparison of the number of non-Western countries represented, the mean and median were

examined, and a Mann-Whitney U test was used to compare this between the two article types. The Mann Whitney U test was used as the relevant data were not normally distributed. For the proportion of non-Western countries, the median proportion was examined, and a Mann Whitney U test was run to compare this characteristic between article types.

6.2.3. Methods for Coding and Analysis of the Number of Authors

The number of authors was coded for each paper by checking the article and counting the number of authors included in the author list. The data that had been gathered for the number of authors per paper was checked and no additional changes were needed before analysing this. The mean and median for the number of authors per paper were examined and a Mann Whitney U test was run to compare this between the article types.

6.3 Results

6.3.1. Results for Differences in Author Seniority

6.3.1.1. Differences in Seniority of the First Author

In order to analyse whether the first authors appeared to have a PhD, a total of 103 authors (26 RR authors and 77 SRs authors) had to be removed from this analysis due to lack of clarity regarding their job title and/or PhD status. Following the exclusion of these unclear responses, analysis of the comparative sample (n = 407, i.e., 144 RRs and 263 SRs) suggested that RR authors were slightly more junior, with slightly fewer first authors of RRs appearing to have a PhD compared with authors of SRs (64.58% vs. 71.48%). However, chi-square analysis showed no statistically significant difference between the article types in whether the first author had a PhD: $x^2(1, N = 407) = 2.07$, $p = 0.15$.

In order to analyse whether the first authors were ECRs, a total of 115 authors (27 RR authors and 88 SRs authors) had to be removed from this analysis due to lack of clarity regarding their job title at the time of publication. Following the exclusion of these unclear responses, analysis of the comparative sample (n = 395 i.e., 143 RRs and 252 SRs) suggested that both article types had similar proportions of first authors that were ECRs, including 52.38% of SRs and 55.94% of RRs. Chi-square analysis showed no statistically significant difference between the article types in whether the first author was an ECR: $x^2(1, N = 395) = 0.47$, $p = 0.50$.

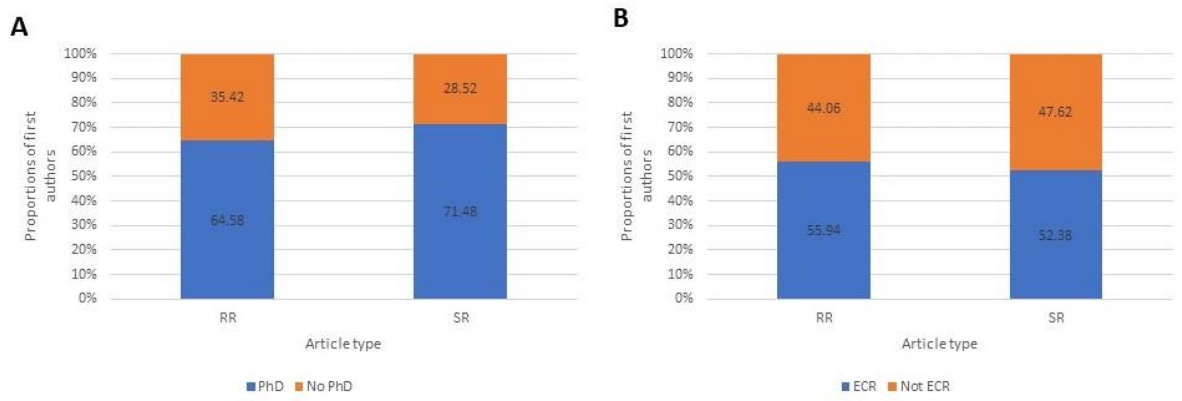

Figure 16: Proportions of first authors' seniority per article type. ***A****. Proportions of first authors with and without PhD, per article type.* ***B.*** *Proportions of first authors who were and were not ECRs, per article type.*

6.3.1.2. Differences in Seniority of the Last Author

In order to analyse whether the last authors appeared to have a PhD, a total of 121 authors (36 RR authors and 85 SR authors) had to be removed from this analysis due to lack of clarity regarding their job title and/or PhD status. Following the exclusion of these unclear responses, analysis of the comparative sample (n = 389 i.e., 134 RRs and 255 SRs) suggested that there was little difference between the article types, with 92.55% of SRs and 91.79% of RRs having last authors that appeared to have PhDs. Chi-square analysis also showed no statistically significant difference between the article types in whether the last author had a PhD: $x^2(1, N = 389) = 0.07$, $p = 0.79$. As all cases where the last author was a consortium authorship had been excluded from the comparative sample, these findings would not change based on whether the authors within consortium authorships were or were not included in the analysis.

In order to analyse whether the last authors were ECRs, a total of 131 authors (35 RR authors and 96 SRs authors) had to be removed from this analysis due to lack of clarity regarding their job title at the time of publication. Following the exclusion of these unclear responses, analysis of the comparative sample (n = 379, i.e., 135 RRs and 244 SRs) slightly more last authors were ECRs in RRs (14.82%), compared with SRs (8.20%). Chi-square analysis showed only a borderline significant difference between the article types in whether the last author was an ECR: $x^2(1, N = 379) = 4.03$, $p = 0.05$. As with the analysis of whether the last author had a PhD, authors from consortium authorships did not affect these results as papers with consortium authorships were not included in this sample.

The seniority of the second and third authors were also considered and these findings are reported in the online Supplementary Appendix 3.[7]

In summary, no substantial differences were found between RRs and SRs in the seniority of the authors, although a trend was noted for whether the last author was an ECR, with RRs being slightly more likely than SRs to have an ECR as the last author.

6.3.2. Results for Differences in Geographic Diversity in Authorship

6.3.2.1. Representation of Any non-Western Countries

The sample for the analysis of whether any non-Western countries were represented used the full comparative sample of 510 articles (170 RRs and 340 SRs). Descriptively, SRs were more likely to include at least one non-Western country (14.12%) compared with RRs (8.82%), but chi-square analysis showed no statistically significant difference: $x^2(1, N = 510) = 2.93$, $p = 0.09$.

6.3.2.2. Number and Proportion of non-Western Countries Represented

Due to the low numbers of articles that actually included any authors based in non-Western countries, the sample for this analysis consisted of only 63 articles in total (48 SRs and 15 RRs). The median number of non-Western countries was 1 for both article types, while the mean number was higher in RRs ($M = 3.73$, $SD = 6.72$) than in SRs ($M = 1.15$, $SD = 0.46$). An independent-samples Mann-Whitney U test showed no significant difference between the article types in the number of non-Western countries represented although there was a slight trend ($U = 424$, $p = 0.09$). When the proportions of authors from non-Western countries within each paper were examined, the SRs had higher average proportions of non-Western countries represented (median 1, $M = 0.80$, $SD = 0.29$) than the RRs (median 0.5, $M = 0.57$, $SD = 0.34$). The independent-samples Mann-Whitney U test showed that there was a statistically significant difference between the article types in the proportions of non-Western countries represented ($U = 221.5$, $p = 0.01$).

6.3.3. Results for Differences in Number of Authors per Paper

The sample for this analysis was the full comparative sample of 510 articles in total (consisting of 170 RRs and 340 SRs). The median number of authors per paper was 4 for both article types, while the mean number was higher in RRs ($M = 8.02$, $SD = 22.00$) than in SRs ($M = 4.49$, $SD = 4.60$). An independent-sample Mann-Whitney U test showed no

[7] Supplementary Appendices are available on the project's OSF page: https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

significant difference between the article types in the number of authors per paper (U = 31103.5, p = 0.16).

6.4. Discussion

The results found low average rates of collaboration with no significant difference between RRs and SRs in the number of authors per paper, indicating that RRs are not associated with significantly greater collaboration than standard research reports. Low rates of geographic diversity were also found and the only significant difference between the article types showed that SRs included higher average proportions of non-Western authors than RRs did, indicating that this may be an important area for improvement of the RR format and its accessibility. Finally, the rates of ECR authors suggest that the format is not wholly inaccessible to junior researchers, and that there were no substantial differences between the article types in the seniority of the authors.

6.4.1. Comparison with Other Research

Although the study generally did not reveal significant differences between the article types, it did show very low rates of representation of authors based in non-Western countries which has also been reported in a number of other studies. In particular, this is broadly in line with the results reported by Lin and Li (2022) who examined geographic diversity of authors, editors and journal owners within psychology and found that the US was overwhelmingly overrepresented compared with other countries, particularly countries in the Majority World. Their approach to calculating this diversity was more detailed than the approach used in the current study, as they used Simpson's diversity index, a calculation which allowed them to consider both the number of countries, and the distributions within these countries. The index ranges from 0 to 1, with a larger number representing greater diversity. This is a useful alternative to the simpler approach taken to coding the rates of non-Western representation in the current study as the method used doesn't account for the number of authors from each country. This approach is therefore limited and more nuanced coding of this is a potential direction for future research, as this is likely to show even lower rates of representation from non-Western countries than was found in the current study, since countries represented more than once in the author list are likely to often be Western countries.

Additionally, the current study showed descriptively lower rates of ECRs publishing RRs, compared with findings reported by Chambers (2019) and Chambers and Tzavella (2022). Specifically, Chambers (2019) reports that 78% of first authors of RRs at Cortex were PhD

students or postdoctoral researchers, compared with 67% of first authors of SRs. Similarly, Chambers and Tzavella (2022) examined 141 RRs across 4 journals and found that 77% of stage 1 manuscripts submitted had PhD students or postdoctoral researchers as their first authors. These rates are much higher than the 56% of ECR authors found in the current study. However, this difference may potentially be explained by the fact that their study focuses on stage 1 reports, while the current study examined this only in completed stage 2 reports. It may be that junior researchers did not previously make as much use of the format as they currently do, but that greater numbers of ECRs are now submitting protocols as the format has become more widely known. Additionally, as their study considered submissions regardless of whether these were later accepted or rejected, this may account for some of the differences in rates between the two studies if some of these submissions were then rejected and did not lead to final manuscripts. It is possible that there may also be disciplinary differences, as the journals examined in these other studies focused primarily on neuroscience, whereas the current study includes a broader range of journals and disciplines within the area of psychology, neuroscience and related disciplines. Furthermore, whereas the current study includes those on short term research contracts like Research Affiliates or Research Associates in the definition of ECRs, Chambers and Tzavella (2022) mention only PhD students and postdoctoral researchers and it is unclear whether they considered these other types of roles as being ECRs.

6.4.2. Limitations

Some of the analyses relied on relatively small sample sizes due to the number of exclusions that were necessary. For example, authors frequently had to be excluded from the author seniority analysis as their role or seniority level was unclear. This raises the question of whether the seniority variables could be coded in a different way, but as these were heavily dependent on the amount of information available online about the authors, a lack of clarity for many of these individuals is inevitable. Furthermore, the authors that seemed most likely to be unclear were the second and third authors rather than the first or last. This is worth acknowledging as the seniority of the first and last authors were the key analyses of interest for this research question, and these also seemed to be the least affected by this lack of clarity.

As the variables that were eventually created for the author seniority analyses were not planned specifically from the beginning of the initial coding process, the data gathered about the authors' job titles often did not specify whether the author had a PhD. Therefore, these

details had to be re-checked when creating the PhD variables, which added extra work on top of what had already been a very time-consuming process. This is a clear limitation of the current coding approach, that could have been avoided if the specific analysis approach for this had been planned in advance. It does however, provide a clear lesson for the improvement of the coding approach going forward. In future the initial coding process should require that this information is also clearly documented in the first instance.

An additional concern regarding the coding of the author seniority was that finding information about authors based outside of Western countries (particularly in China) was particularly challenging, with this information frequently not being available. While the exact extent of this problem wasn't recorded while coding, it does mean that such authors are underrepresented in the seniority samples. Although this issue does not affect the analysis of the representation of non-Western countries, the fact that the author seniority analyses contain fewer authors from such countries is an unfortunate limitation of the data collection process. It is possible that this information may sometimes be available but not in English in which case it could have been missed while collecting the initial data.

Finally, while first and last authors were used for the seniority analyses, the meaning of last author varies across disciplines. While it is often taken to represent a more senior or supervisory role in the work, this is not universal (Bhattacharya, 2010; Einav & Yariv, 2006). In particular if co-authors are at similar career stages then the last authorship position may mean little and so it may be more meaningful only when the last author is a primary investigator. Therefore, any variation in the meaning of last authorship of included papers may undermine the conclusions drawn here. However, given the importance that is often attributed to last authorship positions, it was still considered an important piece of information to investigate.

6.4.3. Future Directions and Implications

In addition to the potential changes to the coding approach alluded to above, the findings have implications for the RR format more generally. For example, the findings highlight the need for greater accessibility of the format internationally. This may require advocates of the RR approach to work with journals based in other countries, to focus on raising awareness of the RR format internationally, and to work on ensuring that open access fees and other barriers to publishing do not unduly hinder authors from majority world countries using this format. For example, many journals offer fee waivers or reductions for authors from low or

middle income countries, and ensuring that this option is available (and advertised widely) may help to increase uptake of the format internationally. However, evidence does exist to indicate that even when authors from low and middle-income countries are eligible for such waivers they are rarely used, and that even large reductions in publishing fees are often still not sufficient to enable researchers from lower-income regions to publish open access (Smith et al., 2021), so such approaches may not be particularly effective after all. Furthermore, a larger barrier may be obtaining funding to carry out the research itself, particularly since this needs to be in place for RRs when they are first submitted, whereas SRs are submitted after the research is done and so any challenges in accessing funding and resources would already have been resolved by the point of submission. Targeted RR-funding partnerships may be one approach to consider in order to support researchers from lower-income countries to conduct and publish their research as RRs.

Although the RR format has been used by ECRs, it could be more widely used in this population. Efforts should be made to increase awareness among ECRs in relation to the RR format, as well as also increasing awareness among senior researchers who supervise ECRs, as this may help to make senior researchers more open to their students and colleagues using this approach. Efforts are also needed to increase practical support for ECRs using this format. This may include increased training and resources being made available to ECRs in relation to this format such as the skills necessary to write an RR, as well as related skills such as preparing data for sharing.

It is also important to fully understand the reasons why ECRs are not engaging more with the format, in order to identify common barriers and perceptions of the format that may affect their uptake by junior researchers. However, as many criticisms of the RR format as being inaccessible to ECRs centre around the time and resources required to complete a full RR, existing initiatives that aim to address these concerns should be promoted and developed further. For example, PCI-RR and its scheduled review track helps to reduce the time required for the stage 1 review process and so this should be promoted to ECRs who have been reluctant to use the format due to the time commitment required for the editorial process. ECRs' involvement in large-scale collaborative RR projects such as the Psychological Science Accelerator rather than only in individual projects, may help to overcome skill and resource constraints that prevented their use of the RR format.

The integration of RR-like stages into undergraduate student projects (Button, 2018) may also help to normalise this approach to conducting research and so may help to build more positive attitudes and familiarity with this type of approach among students. This may in turn help to ensure that the next generation of researchers is more amenable to this approach to conducting and publishing research.

6.5. Conclusion

In conclusion, this study found little difference in authorship characteristics between article types, with no significant differences found in the number of authors per paper or the seniority of the authors. The proportion of non-Western countries included was higher for SRs than for RRs, although rates of this were low for both article types, indicating that this an important area for future improvement both for RRs and for the standard literature.

6.6. References

Adair, J. G., & Huynh, C.-L. (2012). Internationalization of psychological research: Publications and collaborations of the United States and other leading countries. *International Perspectives in Psychology: Research, Practice, Consultation, 1*(4), 252–267. https://doi-org.abc.cardiff.ac.uk/10.1037/a0030395

Adams G, Gómez Ordóñez L, Kurtiş T, Molina LE, Dobles I. (2017). Notes on decolonizing psychology: from one special issue to another. *South African Journal of Psychology*, 47(4), 531-541. doi:10.1177/0081246317738173

Allen, C. & Mehler, D.M.A. (2019). Open science challenges, benefits and tips in early career and beyond. *PLoS Biology 17*(5), e3000246. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pbio.3000246

Apicella, C., Norenzayan, A., & Henrich, J. (2020). Beyond WEIRD: A review of the last decade and a look ahead to the global laboratory of the future. *Evolution and Human Behavior, 41*(5), 319-329. https://doi.org/10.1016/j.evolhumbehav.2020.07.015.

Association for Psychological Science (n.d.) Registered replication reports. *Association for Psychological Science*. Available from: https://www.psychologicalscience.org/publications/replication (Accessed 3rd November 2022).

Azevedo, F., Parsons, S., Micheli, L., Strand, J., Rinke, E., Guay, S., Elsherif, M., Quinn, K., Wagge, J.R., Steltenpohl, C., Kalandadze, T., Vasilev, M., Ferreira de Oliveira, C., Aczel, B., Miranda, J., Galang, C.M., Baker, B.J., Pennington, C.R., Marques, T., Laverty, C., Liu, M., Weisberg, Y., Evans, T.R., Pownall, M., Clark, K., Albayrak-Aydemir, N., Westwood, S., Hartmann, H., & FORRT. (2019, December 13). *Introducing a Framework for Open and Reproducible Research Training (FORRT).* OSF Preprints. https://doi.org/10.31219/osf.io/bnh7p

Barch, D.M. (2021). Biological Psychiatry: Global Open Science—Supporting Open Science and Global Diversity in Research. *Biological Psychiatry: Global Open Science, 1*(1), 2. https://doi.org/10.1016/j.bpsgos.2021.04.004

Bhattacharya, S. (2010). Authorship issue explained. *Indian Journal of Plastic Surgery, 43*(2), 233-234.

Budge, E.J., Tsoti, S.M., Howgate, D.J., Sivakumar, S., & Jalali, M. (2015). Collective intelligence for translational medicine: Crowdsourcing insights and innovation from an interdisciplinary biomedical research community. *Annals of Medicine, 47*(7), 570-5. https://doi.org/10.3109/07853890.2015.1091945

Burgman, M. (2019). Open access and academic imperialism. *Conservation Biology, 33*(1), 5-6. https://doi-org.abc.cardiff.ac.uk/10.1111/cobi.13248

Button, K. (2018). Reboot undergraduate courses for reproducibility. *Nature, 561*, 287. https://doi-org.abc.cardiff.ac.uk/10.1038/d41586-018-06692-8

Center for Open Science (n.d.) Registered Reports FAQs. Center for Open Science. https://osf.io/gha9f

Chambers, C. (2020, March 16). CALLING ALL SCIENTISTS: Rapid evaluation of COVID19-related Registered Reports at Royal Society Open Science. NeuroChambers. http://neurochambers.blogspot.com/2020/03/calling-all-scientists-rapid-evaluation.html

Chambers, C. (2019, June 18). Registered Reports and PhDs – What? Why? How? An interview with Chris Chambers. Center for Open Science. https://www.cos.io/blog/rrs-phds-what-why-how-interview-chambers

Chambers, C. (2019). What's next for Registered Reports? *Nature, 573*, 187-189. https://doi-org.abc.cardiff.ac.uk/10.1038/d41586-019-02674-6

Chambers, C. & Dunn, A. (2022). Rapidly reviewing Registered Reports: A retrospective. *Royal Society Open Science*. https://royalsociety.org/blog/2022/09/registered-reports/

Chartier, C., McCarthy, R., & Urry, H. (2018, Feb 28). The Psychological Science Accelerator. *Association for Psychological Science*. Available from: https://www.psychologicalscience.org/observer/the-psychological-science-accelerator

Cheek, N. N. (2017). Scholarly merit in a global context: The nation gap in psychological science. *Perspectives on Psychological Science*, *12*(6), 1133–1137. https://doi-org.abc.cardiff.ac.uk/10.1177/1745691617708233

Cheon, B. K., Melani, I., & Hong, Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science*, *11*(7), 928–937. https://doi-org.abc.cardiff.ac.uk/10.1177/1948550620927269

Clancy, C.M., Margolis, P.A., & Miller, M. (2013). Collaborative networks for both improvement and research. *Pediatrics, 131*(Suppl 4), S210-4. https://doi.org/10.1542/peds.2012-3786h

Clark, K. (2017, November 21). Myth of the genius solitary scientist is dangerous. *The Conversation*. https://theconversation.com/myth-of-the-genius-solitary-scientist-is-dangerous-87835

Custovic, A., Ainsworth, J., Arshad, H., Bishop, C., Buchan, I., Cullinan, P., Devereux, G., Henderson, J., Holloway, J., Roberts, G., Turner, S., Woodcock, A. & Simpson, A. (2015). The Study Team for Early Life Asthma Research (STELAR) consortium 'Asthma e-lab': team science bringing data, methods and investigators together. *Thorax, 70*(8), 799-801. http://dx.doi.org/10.1136/thoraxjnl-2015-206781

Dege, M., Strasser, I. (2021). Decolonize psychology. In: Strasser, I., Dege, M. (eds) *The psychology of global crises and crisis politics. Palgrave Studies in the Theory and History of Psychology*. Palgrave Macmillan. https://doi-org.abc.cardiff.ac.uk/10.1007/978-3-030-76939-0_17

DeHaven, A. C., Graf, C., Mellor, D. T., Morris, E., Moylan, E., Pedder, S., & Tan, S. (2019, September 17). *Registered Reports: views from editors, reviewers and authors.* MetaArXiv https://doi.org/10.31222/osf.io/ndvek

Dutra, N. B. (2021). Commentary on Apicella, Norenzayan, and Henrich (2020): Who is going to run the global laboratory of the future? *Evolution and Human Behavior, 42*(3), 271-273. https://doi.org/10.1016/j.evolhumbehav.2021.04.003.

Eder, A.B. & Frings, C. (2021). Registered Report 2.0: The PCI RR Initiative. *Experimental Psychology, 68*(1), 1-3. https://doi.org/10.1027/1618-3169/a000512

Einav, L. & Yariv, L. (2006). What's in a surname? The effects of surname initials on academic success. *Journal of Economic Perspectives, 20*, 175–188.

Errington, T.M., Denis, A, Perfito, N., Iorns, E., & Nosek, B.A. (2021) Reproducibility in Cancer Biology: Challenges for assessing replicability in preclinical cancer biology. *eLife, 10,* e67995. https://doi.org/10.7554/eLife.67995

Forscher, P.S., Wagenmakers, E.J., Coles, N.A., Silan, M.A., Dutra, N.,Basnight-Brown, D., & IJzerman, H. (2022). The benefits, barriers, and risks of big-team science. *Perspectives in Psychological Science*. Advance online publication. https://doi-org.abc.cardiff.ac.uk/10.1177/17456916221082970

Frith, U. (2020). Fast lane to slow science. *Trends in Cognitive Science, 24*(1), 1-2. https://doi-org.abc.cardiff.ac.uk/10.1016/j.tics.2019.10.007

Green, S. (2019, August 24). The truth behind the lone genius myth. Wiley. https://www.wiley.com/en-us/network/publishing/research-publishing/trending-stories/the-truth-behind-the-lone-genius-myth

Haslam, N. & Kashima, Y. (2010). The rise and rise of social psychology in Asia: A bibliometric analysis. *Asian Journal of Social Psychology, 13*(3), 202-207. https://doi-org.abc.cardiff.ac.uk/10.1111/j.1467-839X.2010.01320.x

Henderson, E. (2019 July 22). Why you'll love writing a Registered Report (a guest blog). Prolific. https://www.prolific.co/blog/why-youll-love-writing-a-registered-report

Henrich, J., Heine, S. & Norenzayan, A. (2010). Most people are not WEIRD. *Nature* 466, 29. https://doi-org.abc.cardiff.ac.uk/10.1038/466029a

Kathawalla, U., Silverstein, P., & Syed, M. (2021). Easing into open science: A guide for graduate students and their advisors. *Collabra: Psychology, 7*(1), 18684. https://doi.org/10.1525/collabra.18684

Kiser, G.L. (2018). No more first authors, no more last authors. *Nature, 561,* 435. https://doi-org.abc.cardiff.ac.uk/10.1038/d41586-018-06779-2

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.abc.cardiff.ac.uk/10.1027/1864-9335/a000178

Kiyonaga, A., & Scimeca, J. M. (2019). Practical considerations for navigating registered reports. *Trends in Neurosciences, 42*(9), 568–572. https://doi-org.abc.cardiff.ac.uk/10.1016/j.tins.2019.07.003

Ledgerwood, A., Pickett, C., Navarro, D., Remedios, J.D., & Lewis, N.A. (2022). The unbearable limitations of solo science: Team science as a path for more rigorous and relevant research. *Behavioral and Brain Sciences, 45*(81). https://doi.org/10.1017/s0140525x21000844

Lin, Z., & Li, N. (2022). Global diversity of authors, editors, and journal ownership across subdisciplines of psychology: current state and policy implications. *Perspectives on Psychological Science*. Early online publication. https://doi-org.abc.cardiff.ac.uk/10.1177/17456916221091831

Ludwig, R. (2019). Registered Reports for ECRs: Enabling slow science on tight timelines. *Innovation in Aging, 3*(Suppl 1), S24. https://doi.org/10.1093%2Fgeroni%2Figz038.089

Lund, B.D. (2022). Is academic research and publishing still leaving developing countries behind? *Accountability in Research, 29*(4), 224-231. https://doi-org.abc.cardiff.ac.uk/10.1080/08989621.2021.1913124

Maizey, L. & Tzavella, L. (2019). Barriers and solutions for early career researchers in tackling the reproducibility crisis in cognitive neuroscience. *Cortex, 113*, 357-359. https://doi.org/10.1016/j.cortex.2018.12.015

Mekonnen, A., Downs, C., Effiom,E., Kibaja,M., Lawes, M., Omeja, P., Ratsoavina, F., Razafindratsima, O., Sarkar, D., Stenseth, N., & Chapman, C. (2021). Can I afford to publish? A dilemma for African scholars. *Ecology Letters, 25*(4), 711-715. https://doi-org.abc.cardiff.ac.uk/10.1111/ele.13949

Merriman R, Galizia I, Tanaka S, Sheffel A, Buse K, Hawkes S. (2021). The gender and geography of publishing: a review of sex/gender reporting and author representation in leading general

medical and global health journals. *BMJ Global Health, 6*(5), e005672. https://doi.org/10.1136%2Fbmjgh-2021-005672

Montgomery, S.L. (2009). English and Science: realities and issues for translation in the age of an expanding lingua franca. *The Journal of Specialised Translation, 11, 6-16.*

Morey, C.C. & Tzavella, L. (2018, August 17). Going for a Registered Report? The Mnemonic Lode. https://www.candicemorey.org/?p=265

Moshontz, H., Ebersole, C.R., Weston, S.J., & Klein, R.A. (2021). A guide for many authors: Writing manuscripts in large collaborations. *Social and Personality Psychology Compass, 15,* e12590. https://doi.org/10.1111/spc3.12590

Multi-PART, n.d. Multicentre Preclinical Animal Research Team. Available from: https://cordis.europa.eu/project/id/603043 (accessed 3rd November 2011).

Myers, J.S., Lane-Fall, M., & Soong, C. (2021). No one left behind: a case for more inclusivity in authorship for quality improvement and implementation research. *BMJ Quality & Safety, 30,* 779-781. http://dx.doi.org/10.1136/bmjqs-2021-013067

Nabout, J.C., Parreira, M.R., Teresa, F.B., Carneiro, F.M., da Cunha, H.F., de Souza Ondei, L, Caramori, S.S., & Soares, T.N. (2015). Publish (in a group) or perish (alone): the trend from single- to multi-authorship in biological papers. *Scientometrics* 102, 357–364. https://doi-org.abc.cardiff.ac.uk/10.1007/s11192-014-1385-5

Naidoo A.V., Hodkinson P., Lai King L., & Wallis, L.A. (2021). African authorship on African papers during the COVID-19 pandemic. *BMJ Global Health, 6,* e004612. http://dx.doi.org/10.1136/bmjgh-2020-004612

Nicholas, D., Rodríguez-Bravo, B., Watkinson, A., Boukacem-Zeghmouri, C., Herman, E., Xu, J., Abrizah, A., & Świgon, M. (2017). Early career researchers and their publishing and authorship practices. *Learned Publishing, 30,* 205–217. https://doi-org.abc.cardiff.ac.uk/10.1002/leap.1102

Nicholas, D., Watkinson, A., Boukacem-Zeghmouri, C., Rodríguez-Bravo, B., Xu, J., Abrizah, A., Świgon, M., & Herman, E. (2017). Early career researchers: Scholarly behaviour and the prospect of change. *Learned Publishing, 30,* 157-166.

O’Brien, A., Graf, C., & McKellar, K. (2019). How publishers and editors can help early career researchers: Recommendations from a roundtable discussion. *Learned Publishing, 32*(4), 383-393. https://doi-org.abc.cardiff.ac.uk/10.1002/leap.1249

O'Gorman, J., Shum, D. H. K., Halford, W. K., & Ogilvie, J. (2012). World trends in psychological research output and impact. *International Perspectives in Psychology: Research, Practice, Consultation, 1*(4), 268–283. https://doi-org.abc.cardiff.ac.uk/10.1037/a0030520

O’Neil, D. (2018). English as the lingua franca of international publishing. *World Englishes, 37*(2), 146-165. https://doi-org.abc.cardiff.ac.uk/10.1111/weng.12293

Parker, T., Fraser, H., Nakagawa, S. (2019). Making conservation science more reliable with preregistration and registered reports. *Conservation Biology, 33*(4), 747-750. https://doi.org/10.1111/cobi.13342

Patel, M.M., Moseley, T.W., Nia, E.S., Perez. F., Kapoor, M.M., & Whitman, G.J. (2021). Team science: a practical approach to starting collaborative projects. *Journal of Breast Imaging, 3*(6), 721–726. https://doi.org/10.1093/jbi/wbab034

PCI Registered Reports (n.d.) Guide for Authors. https://rr.peercommunityin.org/help/guide_for_authors#h_61998243643551613309672490

Piocuda, J. E., Smyers, J. O., Knyshev, E., Harris, R. J., & Rai, M. (2015). Trends of internationalization and collaboration in U.S. psychology journals 1950–2010. *Archives of Scientific Psychology, 3*(1), 82–92. https://doi-org.abc.cardiff.ac.uk/10.1037/arc0000020

Ross-Hellauer, T. (2022). Open science, done wrong, will compound inequities. *Nature, 603,* 363. https://doi-org.abc.cardiff.ac.uk/10.1038/d41586-022-00724-0

Saez-Rodriguez, J., Costello, J.C., Friend, S.H., Kellen, M.R., Mangravite, L., Meyer, P., Norman, T., & Stolovitzky, G. (2016). Crowdsourcing biomedical research: leveraging communities as innovation engines. *Nature Reviews: Genetics, 17*(8), 470-86. https://doi.org/10.1038/nrg.2016.69

Syed, M. (2021, 10 June). WEIRD Times: Three Reasons to Stop Using a Silly Acronym. Get Syeducated. Available from: http://getsyeducated.blogspot.com/2021/06/weird-times-three-reasons-to-stop-using.html

Schweinsberg, M., Madan, N., Vianello, M., Sommer, S.A., Jordan, J., Tierney, W., Awtrey, E., Zhu, L.L., Diermeier, D., Heinze, J.E., Srinivasan, M., Tannenbaum, D., Bivolaru, E., Dana, J.,

Davis-Stober, C.P., du Plessis, C., Gronau, Q.F., Hafenbrack, A.C., Liao, E.Y., … Uhlmann, E.L. (2016). The pipeline project: Pre-publication independent replications of a single laboratory's research pipeline. *Journal of Experimental Social Psychology, 66*, 55-67. https://doi.org/10.1016/j.jesp.2015.10.001.

Simpson E. H. (1949). Measurement of diversity. *Nature, 163*(4148), 688–688.

Smith, A.C., Merz, L., Borden, J.B., Gulick, C.K., Kshirsagar, A.R., Bruna, E.M. (2022). Assessing the effect of article processing charges on the geographic diversity of authors using Elsevier's "Mirror Journal" system. *Quantitative Science Studies*, 2(4), 1123–1143. https://doi.org/10.1162/qss_a_00157

Suzina, A. C. (2021). English as lingua franca. Or the sterilisation of scientific work. *Media, Culture & Society*, *43*(1), 171–179. https://doi-org.abc.cardiff.ac.uk/10.1177/0163443720957906

Thalmayer, A. G., Toscanelli, C., & Arnett, J. J. (2021). The neglected 95% revisited: Is American psychology becoming less American? *American Psychologist, 76*(1), 116–129. https://doi-org.abc.cardiff.ac.uk/10.1037/amp0000622

Toribio-Flórez, D., Anneser, L., de Oliveira-Lopes F.N., Pallandt, M., Tunn, I., & Windel, H. (2021). Where do early career researchers stand on open science practices? A survey within the Max Planck Society. *Frontiers in Research Metrics and Analytics, 5*, 1-20. https://doi.org/10.3389/frma.2020.586992

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

Wang, D., Yan, K.-K., Rozowsky, J., Pan, E., & Gerstein, M. (2016). Temporal dynamics of collaborative networks in large scientific consortia. *Trends in Genetics: Trends in Genetics, 32*(5), 251–253. https://doi.org/10.1016/j.tig.2016.02.006

Webster, G. D., Mahar, E. A., & Wongsomboon, V. (2021). American psychology is becoming more international, but too slowly: Comment on Thalmayer et al. (2020). *American Psychologist, 76*(5), 802–805. https://doi-org.abc.cardiff.ac.uk/10.1037/amp0000747

Wuchty, S., Jones, B. F., & Uzzi, B. (2007). The increasing dominance of teams in production of knowledge. *Science*, *316*(5827), 1036–1039. https://doi.org/10.1126/science.1136099

Chapter 7: Comparative Analysis of Article Citation Rates and Journal Impact Factor

7.1 Introduction

7.1.1. Article Citation Rates and Journal Impact Factor

Article citation counts represent the number of other publications that reference a particular article (Kulkarni et al., 2007). Citations counts are therefore a commonly used metric considered to signify the importance, quality and 'impact' of published research, with a higher citation rate taken as a sign of more important or more successful research. Citation counts are also often considered to be an important marker of a researcher's performance and prominence (Hanel & Haase, 2017; Asknes, 2005) and so, are influential in hiring and promotion decisions (Carpenter et al., 2014; Holden et al., 2008).

A number of factors affect how frequently cited an article is, including use of open practices. For example, whether an article is available open access has been proposed as a possible contributing factor to higher citation rates (Ottaviani, 2016; Langham-Putrow et al., 2021; Sotudeh, 2020; McKiernan et al., 2016). Furthermore, it has been reported that data availability is associated with approximately 97 more citations per article than articles that do not make their data available (Christensen et al., 2019). Colavizza et al. (2020) also show that providing a link to available data is associated with up to a 25% higher citation impact, on average. Other factors affecting citation rates appear to include the author's gender, with men typically being cited more often than women although this is somewhat controversial (Dion et al., 2018; Bendels et al., 2018; Huang et al., 2020). Meanwhile, other sources show that certain characteristics of the article's title and/or keywords appear to be associated with higher citation rates (Rostami et al., 2014; Subotic & Mukherjee, 2014). Excessive self-citation rates can also bias this measure (Szomszor et al., 2020; Ioannidis, 2015). This has led to calls for more qualitative consideration of the ways in which the source is discussed in the paper citing it, as this gives greater context to the citation and its impact (Jha et al., 2017; Hernández-Alvarez et al., 2017).

Citation counts are also thought to be biased by statistically significant results. Clinical studies reporting positive findings receive far more citations than those reporting more negative outcomes (Jannot et al., 2013; Misemer et al., 2016). For example, Duyx et al.'s (2017) review suggests that medical studies with statistically significant results are cited about twice as frequently as those with non-significant results. This citation bias contributes

to the existing problems of positive findings being overrepresented due to publication bias, further distorting the literature. This chain of bias, from publication bias, outcome reporting bias and spin, to citation bias, is outlined by de Vries et al. (2018) in their examination of antidepressant trials. This showed how reporting and citation biases have damaging cumulative effects on the numbers of negative results published, and on the difficulty of finding any negative results that are published.

Another concerning metric is journal impact factor (JIF), which represents the average citation count per article within a journal from the previous two-year period (Hegarty & Walton, 2012). JIFs thereby attempt to quantify a journal's prominence, with higher JIFs suggesting more eminent journals (Brembs et al., 2013). Publishing in journals with higher JIFs is considered an important marker of a researcher's success and, like citation counts, can be influential in hiring and promotion decisions (Hicks et al., 2015; McKiernan et al., 2019), although their influence has been disputed (Abbott et al., 2010), and alternative initiatives such as the Declaration on Research Assessment aim to disrupt such influences.

Due in part to the aforementioned citation biases, JIF for the journal in which an article is published is another potentially misleading indicator of research validity. Furthermore, evidence suggests that methodological rigour may be only average or lower than average in higher-ranking journals (Brembs, 2018; Tressoldi et al., 2013). Impact factor bias has also been found whereby studies reporting positive outcomes have been published in higher-ranking journals than those reporting more negative findings (Tang et al. 2014). Given the greater degree of respect and attention paid to these higher-ranking journals, this may also distort the literature and the representation of positive vs. null or negative findings. Therefore, concerns have been raised regarding the over-reliance on a journal's JIF as a metric of a study's quality. Alternative suggestions have included focusing on the N-pact Factor (NF), i.e., a measure of the statistical power of the empirical studies the journals publish in order to detect typical effect sizes, thereby shifting the focus to the methodological quality and statistical power of the included studies (Fraley & Vazire, 2014). However, such suggestions have not been widely adopted and instead, the system continues to rely heavily on citation counts and JIF.

Reliance on such flawed metrics is a concern not only for its own sake but also because of the considerable influence they wield on researchers' future opportunities and career trajectories. For example, Cabezas-Clavijo et al. (2013) claims that researchers' bibliometric rankings

influence the future allocation of grant funding. This is also supported by Safer and Tang (2009). While this approach may seem sensible given the importance of prioritising truly impactful and promising research areas for funding allocation (Sandström, 2009; van Leeuwen et al., 2001), the biases inherent in these metrics and particularly their suggested associations with lower quality research, may lead to increased funding and support for more biased and less rigorous research. Furthermore, given the extent to which questionable research practices such as HARKing and p-hacking are thought to be associated with positive findings in general, these concerning practices might also be contributing to these more widely cited positive results (Atkinson et al., 1982; Kiai, 2019; Mahoney, 1977). If this is so, the greater likelihood of these positive findings being published and cited to a much greater extent than null or negative results would lead to considerably more attention being paid to findings that may be less trustworthy (Mlinarić et al., 2017), which is in keeping with the culture's emphasis on significant and novel results over robustness and transparency (Duyx et al., 2017; Head et al., 2015; Ioannidis et al., 2014; Nosek et al., 2012; Simmons et al., 2011). This may therefore bias future research and practice, due to greater belief in the findings that are mostly widely reported and cited (Asknes et al., 2019; Carlsson, 2009).

Despite the flaws inherent in these metrics, as long as they are considered to be influential in hiring and promotion decisions, researchers are incentivised to publish in journals with higher JIF, and to try to ensure that their work is highly citable. Given the apparent importance of positive findings for higher citation counts, this may lead researchers to prioritise publishing positive findings and may make them more likely to engage in the kinds of questionable practices discussed in previous chapters, undermining the credibility and trustworthiness of research.

7.1.2. RRs and Citation Rates

As outlined in previous chapters, the RR publishing format has safeguards in place to reduce bias and questionable practices by researchers. However, the perceived costs and benefits of using the RR format may be an important factor in determining whether researchers choose to adopt this approach. There has, for example, been some concern expressed that many higher-ranking journals do not offer the RR format and so researchers who are interested in using this approach may be restricted to publishing in lower-ranked journals that may not be as well respected in their field. Furthermore, researchers may want assurance that publishing RRs will not cost them citations compared to publishing their work as an SR. Likewise, journal

editors may be reluctant to adopt a format that might lead to lower citations and JIF within their journal.

As mentioned above, previous findings suggest that articles employing open practices were cited more frequently, thereby providing some encouraging support for open publishing practices more generally (McKiernan et al., 2016). One previous study has also demonstrated that RRs had similar and possibly higher citation counts than SRs (Hummer et al., 2017). This provides some promising initial evidence of RRs not only not leading to a citation cost, but potentially having a citation boost. However, further research is needed to verify this finding in a larger sample and also to further explore associations between citation rates and the extent to which hypotheses are supported. As RRs aim to reduce the emphasis on positive findings (as demonstrated in chapter 3), their citation rates should, hypothetically, not be associated with positive findings.

7.1.3. Research Questions

This study seeks to investigate three main research questions within the comparative sample of RRs and their matched SRs.

- Question 1: Do article citation rates differ between RRs and SRs?
    - Hypothesis 1: RRs are expected to be cited, on average, at a similar or higher rate than SRs.
- Question 2: Is there an association between article citation rates and rates of supported hypotheses?
    - Hypothesis 2a: Within the SRs, higher article citation rates will be associated with these articles having a higher proportion of supported hypotheses.
    - Hypothesis 2b: Within the SRs, lower article citation rates will be associated with these articles having a higher proportion of unsupported hypotheses.
    - Hypothesis 2c: If RRs break the link between the results of research and the article citation impact, then RRs should not show a positive correlation between higher citation rates and higher rates of supported hypotheses.
    - Hypothesis 2d: If RRs break the link between the results of research and the article citation impact, then RRs should not show a negative correlation between lower citation rates and higher rates of unsupported hypotheses.[8]

[8] As outlined in chapter 3, supported and unsupported were not the only possible coding states for the hypothesis support data - there was also the option of the support being considered unclear. Therefore, the supported and

- Question 3: Are articles that reported greater support for hypotheses published in journals with higher impact factors?
    - Hypothesis 3a: It is expected that within the SRs, there will be a positive correlation between the proportion of supported hypotheses, and the JIF of the journals they are published in[9].
    - Hypothesis 3b. It is expected that within the SRs, there will be a negative correlation between the proportion of unsupported hypotheses, and the JIF of the journals they are published in.
    - Hypothesis 3c. If RRs break the link between journal prominence and research outcomes, then there should be no relationship between rates of supported hypotheses, and JIF.
    - Hypothesis 3d: If RRs break the link between journal prominence and research outcomes, then there should be no relationship between rates of unsupported hypotheses, and JIF.

7.2 Methods

7.2.1. Coding of Citation Rates

The full comparative sample described in most of the previous chapters was used (n = 510, i.e., 170 RRs and 340 SRs) to gather citation rates. Article citation rates for each article were found using three different search engines due to the differences in their measures of this. Google Scholar, Web of Science, and SCOPUS were used to search for these. Citations rates were sought for each of these by searching for the title of the article. If this could not be found, the author names were searched for and the results refined as needed to determine whether the article could still be accessed. The citation rate for each article was then documented in the coding sheet. However, there were a number of articles that could not be found on either Web of Science or Scopus, or occasionally that could not be found on either of these. This was not an issue for any of the articles on Google Scholar which is unsurprising as this has been shown to be the most comprehensive source (Yang & Meho, 2006). To

unsupported hypotheses are not necessarily the inverse of each other (although they would still be highly correlated). This lack of a binary response to whether hypotheses were supported led to the need for separate correlations to assess the relationships of citation rates with both supported and unsupported hypotheses. Likewise, there needed to be separate correlations to assess the relationships of JIF with both supported and unsupported hypotheses.

[9] It was later considered that this prediction may not be the case if there is too little variation here in the data, e.g., if most SRs have quite high rates of support.

account for the differences between the three sources, the citation rates for each article were averaged across the sources to get an overall mean citation score for each article.

Journal impact factor was also sought for each of the included journals, using the most recent report by Clarivate. However, a number of journals were not accessible through Clarivate and so this data could not be universally obtained. The remaining journals were instead searched for through Google in an attempt to find the journal's JIF from other sources but this was often unsuccessful and so the articles whose journals did not have JIF data (N = 14, i.e., 18.67%) were not included in this analysis.

7.2.2. Analysis

Descriptive statistics were examined for the samples. To test hypothesis 1, an independent-samples Mann Whitney U test was used to compare RRs and SRs on citation rates. To test hypothesis 2, one-tailed Spearman's rho correlations were examined between citation rates and the proportions of supported or unsupported hypotheses, for both the SRs and the RRs. This was also conducted controlling for year of publication, to account for any influence that this might have on the results. One-tailed Spearman's rho correlations were also used to examine whether there was a relationship between the rates of supported or unsupported hypotheses, and the journal impact factors. Where needed, the correlational analyses were repeated with outliers removed.

7.3. Results

7.3.1. Comparison of Citation Rates between SRs and RRs

The mean and median citation rates showed similar results for SRs and RRs, as outlined in table 8 below. For example, the overall median scores were 12.5 for RRs and 12.58 for SRs, while for just the Google Scholar citations, RRs had a median citation score of 17.5 compared with 17.0 in SRs. A Mann-Whitney U test on the overall citation score showed that there was no significant difference between the article types in the rate at which they had been cited ($U = 29144.5$, $p = 0.88$). This lack of evidence for a citation cost within RRs supports hypothesis 1, in that the RRs and SRs appear to receive similar numbers of citations, although it does not support the prediction that RRs might be cited at a slightly *higher* level than SRs. Therefore, these results provide partial support for hypothesis 1 overall.

Table 8

Descriptive statistics for citation rates per article type and source

	Mean of citations		Google Scholar citation rates		Web of Science citation rates		SCOPUS citation rates	
	RR	**SR**	**RR**	**SR**	**RR**	**SR**	**RR**	**SR**
Valid	170	340	170	340	157	312	165	304
Missing	0	0	0	0	13	28	5	36
Median	12.500	12.583	17.500	17.000	8.000	10.000	10.000	10.000
Mean	27.900	24.297	36.753	34.735	22.682	17.821	24.842	17.987
Std. Deviation	60.752	45.139	78.291	64.835	61.739	31.759	64.992	30.713
IQR	19.667	19.250	25.750	26.250	15.000	14.000	16.000	16.000
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Maximum	548.333	470.000	832.000	668.000	622.000	337.000	659.000	405.000

7.3.2. Results for Correlations between Citation Rates and Hypothesis Support

7.3.2.1 Relationship Between Citation Rates and Supported Hypotheses in SRs

When examining the relationship between citation rates and supported hypotheses, partial correlations were used in order to control for the year of publication. These showed that the fully supported hypotheses had a positive relationship with citation rates: $r(239) = 0.14$, $p = 0.01$. A similar pattern was found when the more general 'supported hypotheses' variable was used, which combined both fully and partially supported hypotheses as one variable: $r(239) = 0.12$, $p = 0.03$. These correlations remained significant when outliers (3 *SD*s above or below the mean) were removed: fully supported hypotheses again showed a significant one-way positive correlation with citation rates: $r(222) = 0.15$, $p = 0.01$. This was also the case for the combined fully and partially supported hypothesis variable which maintained a significant positive correlation with citation rates when outliers were removed: $r(222) = 0.14$, $p = 0.02$. All of these same analyses were also run without controlling for year, and these can be seen in the online Supplementary Appendix 4.[10] Overall, these findings support hypothesis

[10] The online Supplementary Appendices are available on this project's OSF page: https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

2a that SRs with a higher proportion of supported hypotheses are associated with higher citation rates.

7.3.2.2. Relationship between Citation Rates and Unsupported Hypotheses in SRs

There was initially a significant negative correlation between lower citations and higher rates of non-supported hypotheses: $r(239) = -0.13$, $p = 0.02$, thereby indicating support for hypothesis 2b. However, this relationship became non-significant when the partial correlation that controlled for the year was examined: $r(239) = -0.1$, $p = 0.06$. On balance, the results therefore provide mixed support, at best, for hypothesis 2b. The same patterns were observed when outliers were removed, whereby the analysis that did not control for year showed the following a significant result, but this again became non-significant in the partial correlation that controlled for year: $r(222) = -0.10$, $p = 0.07$.

7.3.2.3 Relationship between Citation Rates and Supported Hypotheses in RRs

Before controlling for year, the results found using the combined hypothesis support variable provided some tentative evidence for a correlation between citation rates and supported hypotheses, but this was not demonstrated using the more important 'fully supported' variable in this initial analysis. However, partial correlations were also examined in order to control for the influence of year. This showed that the relationship between fully supported hypotheses and citation count became significant; $r(138) = 0.15$, $p = 0.04$, and the correlation of the citation count with the combined supported variable was highly significant: $r(138) = 0.25$, $p = 0.00$. With outliers removed, the partial correlation between citation rates and fully supported hypotheses was also still significant: $r(126) = 0.27$, $p = 0.001$, as was the partial correlation that used the combined support variable: $r(126) = 0.31$, $p < 0.001$. Overall, then, the evidence does not appear to support hypothesis 2c (i.e., that there should not be a significant relationship between these variables in the RRs), although the uncontrolled analysis provides some conflicting results.

7.3.2.4. Relationship between Citation Rates and Unsupported Hypotheses in RRs

The result of the initial one-tailed correlation between citation rates and rates of unsupported hypotheses was initially not significant (see online Supplementary Appendix 4 on the project's OSF page[11]) but the partial correlation controlling for year was significant: $r(138) = -0.22$, $p = 0.01$. When outliers were excluded, this showed significant negative correlations between citation rates and the proportion of unsupported hypotheses, whether or not the year

[11] https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133

was controlled for. Specifically, when the year was controlled for the results were as follows: $r(126) = -0.27$, $p < 0.001$. Overall, then, the evidence does not appear to support hypothesis 2d (i.e., that there should not be a significant relationship between citation rates and unsupported hypotheses in the RRs), although the uncontrolled analysis provides some conflicting results.

7.3.3. Results for Correlations between Hypothesis Support and Journal Impact Factor

Valid JIF data could only be obtained for 149 of the RRs and 298 of the SRs, with the remaining items excluded from the analysis. The journal impact factors ranged from 1.53 to 24.25 ($M = 5.02$, $SD = 4.33$; median 3.70).

7.3.3.1. Relationship between Hypothesis Support and JIF in SRs

Within the SR sample, the relationship between JIF and the proportion of fully supported hypotheses was not significant: $r(219) = 0.10$, $p = 0.06$. This was also not significant when using the combined (fully and partially) supported measure: $r(219) = 0.05$, $p = 0.25$. There was also no significant relationship between the JIF and the proportion of unsupported hypotheses: $r(219) = 0.00$, $p = 0.51$. The relationships remained non-significant when outliers were excluded. These findings therefore do not support hypotheses 3a or 3b (i.e., that for SRs, there should be significant relationships between JIF and supported and unsupported hypotheses, respectively).

7.3.3.2. Relationship between Hypothesis Support and JIF in RRs

The relationship between JIF and the proportion of fully supported hypotheses was also not significant in the RR sample: $r(126) = 0.01$, $p = 0.48$. This was also not significant for the combined supported variable: $r(126) = 0.05$, $p = 0.28$, or the unsupported hypotheses variable: $r(126) = -0.04$, $p = 0.35$. Within the RR sample, outliers were excluded and the correlations all remained non-significant. These findings therefore support hypotheses 3c and 3d (i.e., that for RRs, there should not be a significant relationship between JIF and supported hypotheses, and unsupported hypotheses, respectively).

Table 9

Overview of support for chapter 7 hypotheses

Hypotheses	Conclusions
H1: RRs are expected to be cited, on average, at a similar or higher rate than SRs.	H1 was partially supported. There was no significant difference in citations rates between SRs and RRs, supporting the prediction that they would be cited at similar rates, but not supporting the prediction that the RRs might be cited at higher rates than the SRs.
H2a: Within the SRs, higher article citations rates will be associated with these articles having a higher proportion of supported hypotheses.	H2a was supported. There was a positive correlation between higher citation rates and higher rates of supported hypotheses, within the SRs.
H2b: Within the SRs, lower article citation rates will be associated with these articles having a higher proportion of unsupported hypotheses.	H2b was partially supported. There was a negative association between lower citation rates and higher rates of unsupported hypotheses, within the SRs, but only before year was controlled for. When year was controlled for, this relationship became non-significant and this controlled analysis was considered more informative.
H2c: If RRs break the link between the results of research and the article citation impact, then RRs should not show a positive correlation between higher citation rates and higher rates of supported hypotheses.	H2c was not supported, overall. Overall, there was a significant positive relationship between higher citations and higher rates of supported hypotheses within the RRs. One of the uncontrolled analyses did find a non-significant relationship but the analyses controlling for year were considered more informative.
H2d: If RRs break the link between the results of research and the article citation impact, then RRs should not show a negative correlation between lower citation rates and higher rates of unsupported hypotheses.	H2d was not supported, overall. Overall, there was a significant negative relationship between lower citations and higher rates of unsupported hypotheses within the RRs. One of the uncontrolled analyses did find a non-significant relationship but the analyses controlling for year were considered more informative.
H3a: It is expected that within the SRs, there will be a positive correlation between the proportion of supported hypotheses, and the JIF of the journals they are published in.	H3a was not supported. There was no significant relationship between the rates of supported hypotheses and JIF within the SRs.
H3b: It is expected that within the SRs, there will be a negative correlation between the proportion of unsupported hypotheses, and the JIF of the journals they are published in.	H3b was not supported. There was no significant relationship between the rates of unsupported hypotheses and JIF within the SRs.

H3c: If RRs break the link between journal prominence and research outcomes, then there should be no relationship between rates of supported hypotheses, and JIF.	H3c was supported. There was no significant relationship between the rates of supported hypotheses and JIF within the RRs.
H3d: If RRs break the link between journal prominence and research outcomes, then there should be no relationship between rates of unsupported hypotheses, and JIF.	H3d was supported. There was no significant relationship between the rates of unsupported hypotheses and JIF within the RRs.

7.4. Discussion

7.4.1. Recap of Results

No significant difference was found in citation rates between the two article types. Overall, higher citation rates were associated with more positive findings and fewer negative findings within both the SRs and RRs. This therefore suggests that although there was support for hypothesis 2a and some partial support for hypothesis 2b regarding the SRs, hypotheses 2c and 2d (which suggested that RRs might break the association between citation rates and positive findings), are not supported. There was also no significant relationship between the journal impact factor and the proportion of hypothesis support, in either article type. This therefore supported hypothesis 3c and 3d that RRs should not show this association but does not support hypothesis 3a or 3b that SRs should show this association.

7.4.2. Comparison to Other Research

The lack of a significant difference between the article types is supported by Hummer et al.'s (2017) previous work which also showed similar citation rates for RRs and SRs, although they also suggested that there may be a slightly higher rate for the RRs than for the SRs. This slight citation advantage was not observed in the current sample. These slight differences in outcomes between the two studies could be due to the difference in the matching process for the control sample of SRs, as Hummer's approach considered a larger number of SRs and maintained a tighter timeframe for when these were published, although their matching process did not consider other article characteristics such as topic and study designs. It is possible that the differing choices of comparisons articles between the two studies might account for the differences in outcomes. Nevertheless, their findings broadly support the current study's finding that RRs do not appear to incur a citation cost, compared with SRs.

The finding that positive results are associated with higher citation rates, while null results are associated with lower citations, is broadly in keeping with previous research (Jannot et al., 2013; Misemer et al., 2016; Duyx et al., 2017). While these previous studies are focused on citations of medical research, the results appear to be consistent with the current study despite the different disciplines involved. Although it was hoped that RRs might break this pattern of greater citation of positive results, this change in citation practice was not found in the current study. As this association does not appear to have been previously studied in relation to RRs, this represents a novel finding and also shows that the general findings reported within the medical literature do also seem to apply to the psychological literature as well. This suggests that further efforts are needed to change norms and practices when citing research across disciplines, which may represent an important area for advocacy work.

The lack of any association between JIF and either positive or negative findings, contradicts previous research by Tang et al. (2014) which reported that positive trials tend to be published in higher impact journals. Despite a non-significant association between hypothesis support and JIF among RRs, there was also no association for this in SRs, suggesting a lack of evidence of any JIF-dependent publication bias generally.

7.4.3. Strengths and Limitations

This study provides a more detailed consideration of the citation counts of RRs and SRs than that conducted by Hummer et al. (2017). In addition to having a larger and more recent sample than the one that they used, their investigation relied only on Web of Science to gather citation counts. In contrast, the current study considers an average score derived from three separate sources. In particular, the current study found that a small number of the sources did not appear in Web of Science and so this may not be the most suitable source for gathering this data. In this way, the use of multiple sources and an averaged score may give a more complete view of the patterns involved for all of the included sources, particularly as citation counts for all of the included sources were obtained in Google Scholar, thereby ensuring that even sources that could not be located within the other two locations had at least one clear citation count to inform the calculation of the mean score. This experience of gathering the data is supported by Yang and Meho (2006) who showed that Google Scholar is a more comprehensive source for citation data than Web of Science or Scopus, although it may be overly inclusive and so is not the most stringent metric available. Hummer et al. (2017) also considered each article's Altmetric attention scores. This is a broader measure of the wider attention and impact regarding a particular article, including sources like social

media and news coverage. It is interesting that in Hummer et al.'s study, RRs appeared to achieve higher Almetric scores than the SRs, and although this may be due to the novelty of the RR format, it provides some encouraging evidence to show that RRs are not associated with less attention and consideration by others. This could have been an interesting addition to the current study and is a potential area for future research in order to verify Hummer et al.'s claims and to extent their investigation in a much larger sample.

A further possible advantage of Hummer et al.'s work compared to the current study is that they used many more comparison articles for each RR than used in the current study, which may help to ensure more robust comparisons. They also sought to match these within a much stricter timeframe than that used in the current study. However, their approach did incur some logistical challenges in gathering sufficient comparison articles within their selected timeframe of publication and meant that they were unable to fully follow their original plan for matching the control sample. Therefore, although Hummer et al.'s approach may be more thorough in some regards, it would not have been practical for the current study, particularly because the study in this chapter relied on the SRs that had already been matched previously for the work described in chapters 3 to 6. Furthermore, Hummer et al.'s approach did not account for the relevance of the article topic, study design, population or sample size, whereas the current study sought to do so wherever possible within the constraints of the available articles.

7.4.4. Implications and Future Directions

Overall, this chapter suggests that RRs are not associated with less biased publishing and citation practices than SRs are and so further work is needed in order to address this issue and attempt to shift researchers' focus to aspects of the studies that are not related to whether the outcomes are positive. However, as the greater levels of attention for supported than unsupported results seem to be so widespread and RRs have not yet solved this, this increased attention to supported findings could also be due to other reasons, such as researchers not feeling that a negative result is informative enough. Additionally, as methodological rigour has not been investigated in this study as a potential contributing factor to the citation rates, it may be that the studies' outcomes are not the only factor contributing to their citation impact. Therefore, future studies should include measures of the studies' methodological rigour and quality in order to determine whether this has any impact, particularly to understand whether there is any difference in this between RRs and SRs.

7.5 Conclusion

Overall, this chapter provides some discouraging, if unsurprising, evidence regarding the association between citation practices and positive findings in the psychology literature. There are however some encouraging signs such as the similar levels of citation counts for both RRs and SRs, as well as the finding that the JIF of the journals the articles are published in were not associated with the extent to which the hypotheses were supported, in either SRs or RRs. The citation counts themselves do, however, suggest that this association is present for both article types and so further work is needed to change the culture and focus of citation practices. Further work is also needed to understand the extent to which methodological rigour may contribute to citation rates and whether this differs between RRs and the regular literature.

7.6. References

Abbott, A., Cyranoski, D., Jones, N., Maher, B., Schiermeier, Q., & Van Noorden, R. (2010). Do metrics matter? Many researchers believe that quantitative metrics determine who gets hired and who gets promoted at their institutions. With an exclusive poll and interviews, Nature probes to what extent metrics are really used that way. *Nature, 465*(7300), 860. https://link.gale.com/apps/doc/A229527700/AONE?u=anon~4c2d7039&sid=googleScholar&xid=b9fb7401

Asknes, D. W., Langfeldt, L., & Wouters, P. (2019). Citations, citation indicators, and research quality: An overview of basic concepts and theories. *SAGE Open, 9*(1), 1-17. https://doi.org/10.1177%2F2158244019829575

Asknes, D. W. (2005). Citation rates and perceptions of scientific contribution. *Journal of the America Society for Information Science and Technology, 57*(2), 169-185. https://doi-org.abc.cardiff.ac.uk/10.1002/asi.20262

Atkinson, D. R., Furlong, M. J., & Wampold, B. E. (1982). Statistical significance, reviewer evaluations, and the scientific process: Is there a (statistically) significant relationship? *Journal of Counseling Psychology, 29*(2), 189–194. https://doi.org/10.1037/0022-0167.29.2.189

Bendels, M. H. K., Müller, R., Brueggmann, D., & Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PLoS ONE 13*(1), e0189136. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0189136

Brembs, B. (2018). "Prestigious science journals struggle to reach even average reliability": *Frontiers in Human Neuroscience, 12,* Article 376. https://doi.org/10.3389/fnhum.2018.00376

Brembs, B., Button, K., & Munafò, M. (2013). Deep impact: Unintended consequences of journal rank. *Frontiers in Human Neuroscience, 7,* Article 291. https://doi.org/10.3389/fnhum.2013.00291

Burghardt, J., & Bodansky, A. N. (2021). Why psychology needs to stop striving for novelty and how to move towards theory-driven research. *Frontiers in Psychology, 12,* Article 609802. https://doi.org/10.3389/fpsyg.2021.609802

Cabezas-Clavijo, A., Robinson-Garcia, N., Escabias, M., & Jimenez-Contreras, E. (2013). Reviewers' ratings and bibliometric indicators: Hand in hand when assessing over research proposals? *PLoS ONE, 8*(6), Article e68258. https://doi.org/10.1371/journal.pone.0068258

Carlsson, H. (2009). Allocation of research funds using bibliometric indicators—Asset and challenge to Swedish higher education sector. *InfoTrend, 64(*4), 82-88. https://www.sfis.nu/ojs/index.php/infotrend/article/download/119/112

Carpenter, C. R., Cone, D. C., & Sarli, C. C. (2014). Using publication metrics to highlight academic productivity and research impact. *Academic Emergency Medicine, 21*(10), 1160-1172

Christensen G, Dafoe A, Miguel E, Moore DA, & Rose AK (2019) A study of the impact of data sharing on article citations using journal policies as a natural experiment. *PLoS ONE, 14*(12), e0225883. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0225883

Colavizza G, Hrynaszkiewicz I, Staden I, Whitaker K, & McGillivray B (2020) The citation advantage of linking publications to research data. *PLoS ONE, 15*(4), e0230416. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0230416

de Vries, Y. A., Roest, A. M., de Jonge, P., Cuijpers, P., Munafò, M. R., & Bastiaansen, J. A. (2018). The cumulative effect of reporting and citation biases on the apparent efficacy

of treatments: the case of depression. *Psychological Medicine, 48,* 2453–2455. https://doi.org/10.1017/S0033291718001873

Dion, M., Sumner, J., & Mitchell, S. (2018). Gendered citation patterns across political science and social science methodology fields. *Political Analysis, 26*(3), 312-327. doi:10.1017/pan.2018.12

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology, 88,* 92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Fraley RC, Vazire S (2014) The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS ONE, 9*(10), e109019. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0109019

Hanel, P. H. P., & Haase, J. (2017). Predictors of citation rate in psychology: Inconclusive influence of effect and sample size. *Frontiers in Psychology, 8,* Article 1160. https://doi.org/10.3389/fpsyg.2017.01160

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology, 13*(3), Article e1002106. https://doi.org/10.1371/journal.pbio.1002106

Hegarty, P., & Walton, Z. (2012). The consequences of predicting scientific impact in psychology using journal impact factors. *Perspectives on Psychological Science, 7,* 72–78. https://doi.org/10.1177/1745691611429356

Hernández-Alvarez, M., Gomez Soriano, J., & Martínez-Barco, P. (2017). Citation function, polarity and influence classification. *Natural Language Engineering, 23*(4), 561-588. doi:10.1017/S1351324916000346

Hicks, D., Wouters, P., Waltman, L., de Rijcke, S., & Rafols, I. (2015). Bibliometrics: The Leiden manifesto for research metrics. *Nature, 520*(7548), 429-431. https://doi.org/10.1038/520429a

Holden, G., Rosenberg, G., & Barker, K. (2005) Bibliometrics. *Social Work in Health Care, 41*(3-4), 67-92, DOI: 10.1300/J010v41n03_03

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences of the USA, 117*(9), 4609-4616. 10.1073/pnas.1914221117.

Hummer, L., Thorn, F. S., Nosek, B. A., & Errington, T. (2017). *Evaluating Registered Reports: A naturalistic comparative study of article impact.* OSF Preprints. https://doi.org/10.31219/osf.io/5y8w7

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., & David, S. P. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. *Trends in Cognitive Sciences*, *18*(5), 235-241. https://doi.org/10.1016/j.tics.2014.02.010

Ioannidis, J. P. A. (2015). A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. *Journal of Psychosomatic Research, 78*(1), 7-11. https://doi.org/10.1016/j.jpsychores.2014.11.008.

Jannot, A. S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology, 66(*3), 296-301. https://doi.org/10.1016/j.jclinepi.2012.09.015

Jha, R., Jbara, A.-A., Qazvinian, V., & Radev, D. R. (2017). NLP-driven citation analysis for scientometrics. *Natural Language Engineering, 23*(1), 93-130.

Kiai, A. (2019). To protect credibility in science, banish "publish or perish". *Nature Human Behaviour, 3*(10), 1017-1018. https://doi.org/10.1038/s41562-019-0741-0

Kulkarni, A. V., Busse, J. W., & Shams, I. (2007). Characteristics association with citation rate of the medical literature. *PLoS One, 2*(5), Article e403. https://dx.doi.org/10.1371%2Fjournal.pone.0000403

Langham-Putrow A, Bakker C, Riegelman A (2021) Is the open access citation advantage real? A systematic review of the citation of open access and subscription-based articles. *PLoS ONE 16*(6), e0253129. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0253129

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., Spies, J. R., Thaney, K., Updegrove ,A., Woo, K., H., & Yarkoni, T. (2016) Point of View: How open science helps researchers succeed. *eLife, 5*, e16800. https://doi.org/10.7554/eLife.16800

McKiernan, E. C., Schimanski, L. A., Nieves, C. M., Matthias, L., Niles, M. T., & Alperin, J. P. (2019) Meta-Research: Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife, 8*, e47338. https://doi.org/10.7554/eLife.47338

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy and Research, 1*(2), 161–175. https://doi.org/10.1007/BF01173636

Misemer, B. S., Platts-Mills, T. F., & Jones, C. W. (2016). Citation bias favoring positive clinical trials of thrombolytics for acute ischemic stroke: A cross-sectional analysis. *Trials, 17,* Article 473. https://doi.org/10.1186/s13063-016-1595-7

Mlinarić, A., Horvat, M., & Šupak Smolčić, V. (2017). Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia Medica (Zagreb), 27*(3), Article 030201. https://doi.org/10.11613/BM.2017.030201

Ottaviani J (2016) The post-embargo open access citation advantage: It exists (probably), it's modest (usually), and the rich get richer (of course). *PLoS ONE 11*(8), e0159614. https://doi-org.abc.cardiff.ac.uk/10.1371/journal.pone.0159614

Rostami, F., Mohammadpoorasl, A. & Hajizadeh, M. The effect of characteristics of title on citation rates of articles. *Scientometrics, 98*, 2007–2010. https://doi-org.abc.cardiff.ac.uk/10.1007/s11192-013-1118-1

Safer, M. A., & Tang, R. (2009). The psychology of referencing in psychology journal articles. *Perspectives on Psychological Science, 4*(1), 51–53. https://doi-org.abc.cardiff.ac.uk/10.1111/j.1745-6924.2009.01104.x.

Sandström, U. (2009). Research quality and diversity of funding: A model for relating research money to output of research. *Scientometrics, 79,* 341–349. https://doi.org/10.1007/s11192-009-0422-2

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366. https://doi.org/10.1177%2F0956797611417632

Sotudeh, H. (2020). Does open access citation advantage depend on paper topics? *Journal of Information Science, 46*(5), 696–709. https://doi-org.abc.cardiff.ac.uk/10.1177/0165551519865489

Subotic, S., & Mukherjee, B. (2014). Short and amusing: The relationship between title characteristics, downloads, and citations in psychology articles. *Journal of Information Science, 40*(1), 115–124. https://doi-org.abc.cardiff.ac.uk/10.1177/0165551513511393

Szomszor, M., Pendlebury, D.A. & Adams, J. How much is too much? The difference between research influence and self-citation excess. *Scientometrics, 123,* 1119–1147 (2020). https://doi-org.abc.cardiff.ac.uk/10.1007/s11192-020-03417-5

Tang, P.A., Pond, G. R., Welch, S., & Chen, E. X. (2014). Factors associated with publication of randomized phase III cancer trials in journals with a high impact factor. *Current Oncology, 21*(4), 564-572. https://doi.org/10.3747/co.21.1937

Tressoldi, P. E., Giofré, D., Sella, F., & Cumming, G. (2013). High impact = high statistical standards? Not necessarily so. *PLoS One, 8*(2), Article e56180. https://doi.org/10.1371/journal.pone.0056180

van Aert, R. C. M., Wicherts, J. M., & van Assen, M. A. L.M. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PLoS One, 14*(4), Article e0215052. https://doi.org/10.1371/journal.pone.0215052

van Leeuwen, T. N., van der Wurff, L. J., van Raan, A. F. J. (2001). The use of combined bibliometric methods in research funding policy, *Research Evaluation, 10*(3), 195–201. https://doi.org/10.3152/147154401781777015

Yang, K. & Meho, L. I. (2006). Citation Analysis: A Comparison of Google Scholar, Scopus, and Web of Science. *Proceedings of the American Society for Information Science and Technology, 43*(1), 1-15.

Chapter 8: Comparative Analysis of HARKING Rates

8.1 Introduction

8.1.1. Forms of Bias in Research

There are many forms of bias which can affect researchers' approaches, their interpretation of data, and how they conceptualise and present their findings. For example, as outlined in chapter 3, confirmation bias is a widespread and concerning practice that can have negative implications for the conduct and assessment of research. Confirmation bias also manifests as, and intersects with, other forms of bias. For example, hindsight bias refers to the way in which one's knowledge about a situation's outcome can lead to exaggerating the likelihood of predicting that outcome beforehand (Roese & Vohs, 2012). This can be influenced by distortions of memory, by the person's beliefs about how likely certain outcomes are, and also by a person's own belief about their ability to predict the outcomes of a situation. This pattern of thinking can lead to selective recall of information, whereby only information that is consistent with the actual outcome is recalled or prioritised. Like confirmation bias, hindsight bias can make people feel overly confident in their ability to predict outcomes and may encourage them to discard or not fully consider other possible explanations for what has occurred (Roese & Vohs, 2012).

A related issue known as outcome bias refers to the way in which knowledge of the outcomes of a situation are prioritised without due consideration for the quality or appropriateness of the decisions that were made in order to reach that particular outcome (Henricksen & Kaplan, 2003). Whereas hindsight bias involves distorted memories that favour the individual's beliefs, outcome bias does not involve memory distortion in the same way. These various forms of bias are concerning because they can lead to questionable research practices and publication bias which can in turn undermine the replicability and reproducibility of the research findings (Banks et al., 2016; Chambers, 2014, 2017; Chambers & Tzavella, 2022; Field et al., 2020; Franco et al., 2014; Kelly et al., 2014; Leung., 2011; Nosek et al., 2015; Rubin, 2017, 2019; Scheel et al., 2021).

8.1.2. Hypothesising After Results are Known (HARKing)

As previously outlined in chapter 3, Hypothesising After Results are Known (HARKing) is a common practice in which hypotheses that are created after the study has been done are presented as being specified *a priori*, while hypotheses that were actually pre-specified may

be suppressed if they have not been supported by the results found (Rubin, 2017). This can also include retrieving hypotheses from other literature after the results are known in order to suggest greater concordance with the existing evidence. HARKing is thought to occur due to the incentives that exist for positive findings and neat storytelling in research reports, particularly due to publication bias from journals favouring such characteristics in the manuscripts they accept for publication. Self-deception may also be an important consideration here as confirmation and hindsight bias can lead researchers to believe that they did in fact predict the outcome all along.

Overall, HARKing is a concerning practice as, in addition to the dissemination of biased findings, Kerr (1998) argues that this may lead to such falsely positive results becoming widely accepted and difficult to eradicate, particularly if theory is then constructed to account for an illusory effect. Unfortunately, HARKing appears to be relatively widespread, although estimates vary considerably between studies. Kerr (1998) found that 40% of psychologists they surveyed had witnessed other authors engage in HARKing, while Bosco et al., (2016) show that approximately one-third of respondents admitted to changing at least one hypothesis between the completion of data collection, and publication. This is broadly in line with rates reported by John et al. (2012) which showed a self-admission rate of 27%. However, even anonymised, it is possible the researchers might fail to accurately disclose HARKing, either due to dishonesty, self-delusion, or unreliable memory. Therefore, John et al. considered not just the self-admission rate but also estimated prevalence rates and estimated prevalences rates derived from the admission rates, giving much higher scores than those that had been self-admitted. For example, although the overall self-admission rate for researchers reporting an unexpected finding as if it had been predicted was between 30 and 40%, the estimated prevalence rate they derived from this was almost 90%. Results from a study of Italian researchers indicate broadly similar findings although with higher rates of some practices, including changing hypotheses between data collection and publication which was self-reported by 37.4% of the sample (Agnoli et al., 2017). Meanwhile, their self-admission rates for claiming to have predicted an unexpected finding was around 40% and their derived prevalence estimates for this were 100%. Overall, estimates of the rates of these practices are clearly high, although the precise figures vary considerably depending on how they are measured and the likelihood of under-reporting by those surveyed.

8.1.3. RRs as a Proposed Solution for HARKing

Open science practices such as preregistration have been proposed as a potential solution for HARKing as this involves pre-specifying study details before the study is conducted and should therefore provide a record of which decisions were made *a priori*. However, preregistration has not proved as reliable for this as might be hoped. Outcome switching appears to still occur frequently even when a study has been publicly preregistered, and researchers do not typically appear to be held accountable for such deviations or for reporting these deviations honestly (Holst et al., 2023). Therefore, RRs may provide a useful alternative which holds researchers more accountable for honestly reporting their pre-specified hypotheses in their final manuscript, even while this format encourages authors to engage in *post hoc* exploration of the data, as long as it is reported transparently. Therefore, the confirmatory (deductive) and exploratory (inductive) research outcomes can be more clearly differentiated. This clear differentiation reflects one of the seven virtues of high-quality RRs, as reported by Chambers and Tzavella (2022), i.e., certifying that the conclusions stated in the final manuscript are based on the evidence presented *a priori* and are appropriately weighted in favour of the confirmatory outcomes.

The study reported in this chapter was conducted in collaboration with an MSc student for the purposes of their thesis project (White, 2022). While the study was designed in collaboration, the student in question undertook the data collection and coding processes independently, while I checked their work in detail, including the coding of the data.

8.1.4. Research Questions

This study aimed to look for discrepancies between preregistered protocols and their final published manuscripts to determine whether HARKing could be detected in a small sample of RRs and matched SRs, and to understand how rates of HARKing compare between these two article types. Because of the safeguards in place in the RR format, the rates of HARKing should be at or close to zero, while the rate in SRs was expected to be higher than this. In determining whether HARKing was present, the severity of the discrepancies between the protocols and the manuscripts were considered: HARKing was defined as at least one major change to at least one existing hypothesis, or the addition or removal of at least one hypothesis, while more minor changes to the wording of hypotheses that did not alter the meaning of the statement, were not considered to be HARKing.

- Question 1: What is the rate of HARKing in RRs?

    - Hypothesis 1 (H1): The rate of HARKing in RRs should be zero.
- Question 2: Is the rate of HARKing in RRs less than in SRs?
    - Hypothesis 2 (H2): The rate of HARKing is expected to be lower in RRs than in SRs.
- Question 3: Where *post hoc* changes in hypotheses occur, is the severity of these lower in RRs than in SRs?
    - Hypothesis 3 (H3): The severity of any *post hoc* changes in hypotheses is expected to be less in RRs than in SRs.

8.2. Methods

The plan for this study was pre-registered on 27th December 2021 on the Open Science Framework (OSF; https://osf.io/c8pha/). A full list of deviations from that protocol can be found in the online Supplementary Appendix 5 (https://osf.io/5pu4g/?view_only=96aec98ca4dd4d2eb9751cd916183133).

8.2.1. Sample Creation

Due to the time-consuming nature of the investigation process and the limited time and resources available, it was necessary to focus on a small sample of papers for this study. Based on the interests of the MSc student working on this project, the sample was restricted to studies in cognitive psychology and neuroscience. A sample of 12 RRs in this general topic area were taken from the larger database, with one of the main factors determining their inclusion being the availability of a stage 1 protocol for each of these papers. The other factors were that the RRs had been coded and checked by that point in time (beginning of February 2022) and that they also had existing SRs that had been matched, coded and checked.

These existing matched SRs for these RRs (n = 24 SRs) were checked to determine whether they had an associated preregistered protocol available. However, none had a protocol available and so the HARKing-related analysis could not be conducted on this sample of SRs. Efforts were also made to check if any of these paired SRs had associated preprints, as this could have given the opportunity to check for any HARKing influenced by the publication process, e.g., based on reviewers' feedback. Unfortunately, only two of the twenty-four SRs in this sample had an associated preprint. Therefore, this sample was not deemed suitable for inclusion in this particular study and a different SR sample was gathered instead. This new SR sample consisted of 12 SRs selected from the OSF registry which were matched to the

RRs on keywords or topic area, had a pre-registered protocol available, and whose results had already been published in a peer-reviewed journal.

8.2.2. Coding of HARKing

8.2.2.1. Initial Coding Process and Preparation

A range of details were extracted for each of the article types and their associated documents. For RRs, the stage 1 protocol and the published stage 2 manuscript (along with any supplementary information that had been published along with this), were both used. For the newly matched SR sample, the preregistered protocols and their final journal articles were used, along with any supplementary information provided with the articles. For the RRs, much of the necessary information for the initial stage of the coding process had already been gathered previously (e.g., quotes showing the articulation of the hypotheses in the stage 2 manuscript) and so this information was extracted from the existing dataset and was not independently recoded. For the new SR sample, and the details from the stage 1 RR protocols, the coding needed to be conducted in full.

The first round of coding for this study included capturing basic identifying details like the authors, year, and URLs for the documents. Other characteristics gathered included the exact text of the hypotheses that had been stated and the level at which these occurred in the document (i.e., article-level, study-level, hypothesis-level), and whether the statement of the hypothesis occurred in the main text or in supplementary information. As some additional research questions had initially been preregistered for this study, some additional details were also sought initially, although these were not analysed. Of these excluded outcomes, one was the level of support for each hypothesis, with four possible response options consisting of fully confirmed, partially confirmed, disconfirmed, and unknown/unclear. Whether this outcome was reported in the main text or in supplementary information was also recorded. Additional characteristics recorded included whether there was exploratory analysis used at any level of the paper, which was coded as either yes or no. A subjective rating of the coder's certainty in the accuracy of the basic coding was also documented to aid in the process of checking the coding.

8.2.2.2. Second Round of Coding

Based on the information gathered during this initial coding process, a series of characteristics were then coded to represent whether there were any differences between the protocols and the final manuscripts. For both the RRs and the unpaired SRs, categorical

responses were coded for whether there was a change of any kind in the wording of the hypothesis between the final manuscript and the protocol, whether any hypotheses were added to the final manuscript compared to the protocol, whether any hypotheses were removed from the final manuscript compared to the protocol and if so, free text entries were used to document these changes/additions/deletions.

Where there was any change of wording or addition of hypotheses in the final manuscript, the coding also indicated whether these were described in the main manuscript itself or in supplementary material accompanying this. Where any hypotheses had been added into the final manuscript that had not been in the protocol, a categorical variable measured whether that added hypothesis was fully confirmed, partially confirmed, disconfirmed, or unclear/unknown. However, these variables were not examined further in this study.

Where changes had occurred in the wording of the hypothesis between the protocol and the final manuscript, another categorical variable documented the extent (severity) of such changes, i.e., whether they were minor (consisting of wording changes only) or major (consisting of both wording changes and also changes to the meaning or precision). Furthermore, a categorical variable was created to indicate whether there was considered to be evidence of HARKing, which was defined as at least one major change to at least one existing hypothesis, or the addition or removal of at least one hypothesis. A categorical variable indicated whether there was any HARKing, while proportion scores were calculated for the proportion of major HARKing and the proportion of minor HARKing (i.e., of all the HARKing detected, what proportion was major and what proportion was minor, respectively). Finally, a subjective rating of the coder's certainty in the accuracy of the coding was recorded. This new coding was conducted initially by a trained MSc student, before being checked in detail by myself to ensure accuracy and consistency in the coding approach. From the variables created for the evidence of HARKing, *Yes* and *No* were re-coded as *1* or *0* respectively. Proportion scores were calculated to reflect what proportion of the changes were major or minor.

8.2.3. Analysis

8.2.3.1. Overview of Analysis Process

To test each of the three hypotheses, it was first checked whether there were any changes between each preregistered protocol and their final manuscript. If there were no changes, that paper was considered to not contain any HARKing. For this variable, *Yes* and *No* were re-

coded as *1* or *0* respectively. Where changes were found, these were coded for whether these were minor (e.g., rephrasing but with the same meaning) or major (e.g., changing the prediction, or adding/removing hypotheses in the manuscript). Only these major changes were considered to be HARKing for the purposes of this study, and analysis of this in both article types allowed us to test both hypothesis 1 and hypothesis 2. Proportion scores were also calculated to reflect what proportion of the changes were either major or minor. The proportion of the changes that were major was then compared between the article types in order to test hypothesis 3, with exploratory analysis comparing the proportion of minor changes between them. Figure 17 below, adapted from White (2022) shows the process of analysis of hypotheses between the protocol (or Stage 1 manuscript) and the final manuscript.

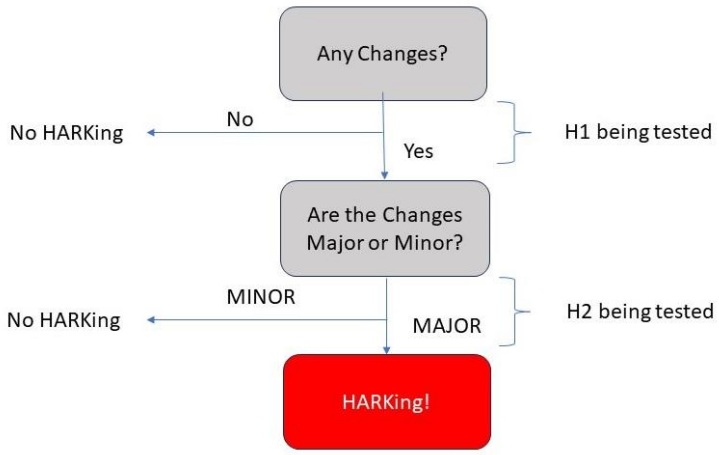

Figure 17: Workflow diagram showing the process for analysing H1 and H2

8.2.3.2. Analysis

In order to investigate hypothesis 1 (whether the rate of HARKing is zero in RRs), only descriptive analysis was needed as it was clear from the final dataset that there was no evidence of HARKing in the RRs and so no further analysis was considered necessary for this hypothesis.

Analysis for hypothesis 2 sought to compare the rates of HARKing between the article types to determine whether this rate was, as expected, higher in SRs than RRs. A Wald test was initially used to test for any association between the independent variable (article type) and the rate of HARKing. However, a chi-square test was later considered to be a more appropriate choice for this nominal variable and so this test was also run. The results of both

tests are reported in the results section, though the chi-square results are considered the more appropriate choice.

Analysis for hypothesis 3 sought to compare the proportion of major changes (between the final manuscripts and the protocols) between RRs and SRs. An independent samples *t*-test was initially used to test whether there was a difference in this between the RRs and SRs. Likewise, the proportions of minor changes were initially compared between RRs and SRs using an independent samples *t*-test. However, due to the small sample size and distribution of the data, Mann-Whitney U tests were later run instead. As above, both the results of the original test and the later test are reported in the results section, though the non-parametric test was more appropriate. Due to JASP's inability to test the difference in the proportions of minor changes for this small sample, the Mann-Whitney U tests for the differences in the proportions of minor and major changes, were conducted in jamovi instead (The Jamovi Project, 2023).

8.3 Results

All data for the study in this chapter can be found on the OSF Registry in the Data folder; Project 2 (https://osf.io/w27fm/)

8.3.1. Confirmatory Results

Descriptively, HARKing did not occur in any of the RRs, while it was detected in 4 of the 12 SRs (33%). This supports hypothesis 1 that the rate of HARKing should be zero in the RRs. As also reported in White (2022), a Wald test initially confirmed that the proportion of articles showing evidence of HARKing was significantly lower for RRs (M = 0, SE=0) than for SRs (M =0.33, SE = 0.136; z = -2.5, p = 0.014), supporting hypothesis 2. Later, chi-square analysis was instead thought to be a more appropriate test for this nominal data, and this showed that the overall findings were broadly consistent with those of the Wald test, i.e., that there was a significant difference in this characteristic between the two article types: $x^2(1, 24) = 4.8$, p = 0.028.

As would then be expected from these results, an independent samples *t*-test (as reported by White, 2022) also showed that the average proportion of hypotheses (per article) that underwent *major* changes relative to the total number of changes (major and minor) was significantly lower for RRs (M = 0, SE = 0) than for SRs [M = 0.33, SE = 0.136; $t(22) = -2.25$, p = 0.035]. The average proportion of hypotheses that underwent minor changes was also significantly lower for RRs [M = 0.33, SE = 0.14) than for SRs (M = 0.88, SE = 0.07;

$t(22) = -3.43$, $p = 0.002$]. This therefore supports hypothesis 3, that the severity of changes will be less in RRs than in SRs. However, due to the small sample size, non-parametric testing was also conducted at a later date, to verify these conclusions. As expected, the Mann-Whitney U tests confirmed a significant difference between the article types in the proportion of major changes ($U = 48.0$, $p = 0.04$) and the proportion of minor changes ($U = 30.0$, $p = 0.01$). Table 10 shows the proportion of major and minor changes within RRs and SRs.

Table 10

Proportion of Major and Minor Changes within Registered Reports and Standard Reports

	Article Type	N	Mean	Std. Deviation	Std. Error Mean
Proportion Major	RR	12	.000	.000	.000
	SR	12	.167	.257	.074
Proportion Minor	RR	12	.333	.492	.142
	SR	12	.880	.237	.069

Note. RR = Registered Reports, SR = Standard Reports, N = Sample Size

8.3.2. Exploratory Analysis: Granularity of HARKing

As the levels at which hypotheses were stated in each paper was considered in the initial coding process, an exploratory analysis used this information to assess the levels at which HARKing may occur within a paper. Out of the four occurrences of HARKing in the SR sample, three of those occurred at the hypothesis level, whereas the other occurrence of HARKing was found at the article level. No HARKing appeared to have occurred at the study level of coding. This overall pattern of granularity is unsurprising given that the vast majority of hypotheses were coded at hypothesis level, with both article and study level being much less common and study level being particularly rare.

8.4 Discussion

8.4.1. Recap of Results

All three hypotheses for this study were supported: the rate of HARKing was found to be zero among the RRs, while this was not the case for SRs, and finally, the proportion of major changes and the proportion of minor changes, relative to the number of total changes, were both significantly less in RRs than in SRs, showing that the severity of any post hoc changes was greater in SRs than in RRs. Exploratory analysis showed that the HARKing detected in the SRs followed the same pattern of granularity as the levels at which hypotheses were generally coded.

8.4.2. Discussion of Findings

The finding that there was no HARKing detected in this sample of RRs provides some initial evidence that rates of HARKing may be low when using this publication format. The comparatively higher rates of this among the SRs further supports the idea that RRs might be an effective solution to reduce HARKing and selective reporting generally. However, as an observational study, causal relationships cannot be determined. Furthermore, this was a very small sample and was restricted to RRs in the area of cognitive psychology and neuroscience. Therefore, any encouraging findings obtained from this limited sample may not apply to RRs in other areas or across the full range of RRs as a whole. Furthermore, these RRs were selected because they had publicly available protocols and so it is possible that those without available protocols might be less effective as a safeguard against HARKing. However, further research would be needed in order to verify this.

Nevertheless, the reported higher rate of HARKing in SRs is supportive of the idea that without some type of preregistration (and in particular, accountability for adhering to this), questionable research practices (QRPs) are evident in the general literature and that these occur at higher rates in standard reports than in RRs. This therefore broadly supports the intended aims of RRs (Chambers, 2014, 2017; Scheel et al., 2021; Hardwick & Ioannidis, 2018). The findings of this study also provide evidence that when changes to the hypotheses do occur within SRs, they can occur as either major changes or minor changes, whereas when assessing changes within RRs, there is only evidence of minor changes in the phrasing but not the meaning of the predictions. This also supports the notion that RRs are working as intended in preventing QRPs and bias, even though some minor changes in wording do occasionally still occur in RRs. Overall, this appears to be a novel area of investigation in relation to RRs and so there is limited previous research that can be compared with this study.

The findings from the SRs in this sample do however, support concerns expressed by others regarding the potential rates of undisclosed discrepancies between preregistered protocols and their published manuscripts (Goldacre et al. 2019; Mathieu et al., 2009; Mathieu et al., 2013; Claesen et al., 2021). While some discrepancies may be for justifiable reasons such as the preregistered study design being poorly designed or inappropriate, clear disclosure of such discrepancies and the reasons for them should be declared transparently. However, Claesen et al. (2021) found that approximately 93% of the preregistered studies they examined from the journal Psychological Science, contained deviations from the preregistered plan, with only one study disclosing all such changes. While they examined a range of different aspects of the studies, not just the hypotheses, these findings highlight the lack of transparency in the literature even when the studies have been preregistered. Claesen et al. also report finding it challenging to distinguish all of the pre-planned and *post hoc* elements of studies even when comparing the manuscripts against the preregistered protocols; this is in line with our experience in the current study and also with the lack of clarity and specificity of protocols reported by TARG Meta-Research Group & Collaborators (2022). Potential solutions include requiring more detailed preregistrations and encouraging the use of templates (Claesen et al., 2021), as well as introducing 'discrepancy review' into the peer review process when a manuscript is submitted that has been preregistered (TARG Meta-Research Group & Collaborators, 2022). This latter suggestion in particular may help to increase the level of accountability for researchers publishing preregistered SR manuscripts.

8.4.3. Limitations

Due to a lack of time and resources available, some additional research questions and hypotheses that were mentioned in the preregistered protocol could not be investigated. These were in relation to how HARKing might be introduced by the peer review process, as well as comparing the unregistered exploratory analyses and preregistered confirmatory analyses in RRs in order to investigate to what extent they seem susceptible to interpretive bias, and thus consider whether RRs are more immune to this interpretive bias than SRs are. These points are also outlined in the document in the online repository regarding the deviations from the protocol. The reason for not including these elements of the study was primarily due to the timeframe available for the MSc. student working on this project, as these additional aspects were more supplementary and would also have been much more time-consuming to code. Although these additional aspects would have provided interesting data to support and extend

this investigation, their inclusion was not considered feasible. They do, however, present potential directions for future research.

As the coding of the hypotheses in this chapter was done by only one person, the issue of subjectivity within the coded data could be considered a limitation of the study's approach. HARKing is not a straightforward phenomenon to detect and the approaches to HARKing can differ in how they manifest as well as in their severity. Furthermore, due to the primary coder's lack of experience in the area of neuroscience and cognitive psychology, the topic area was a challenging one for them to work on, particularly in terms of ensuring a full and nuanced understanding of the hypotheses involved. As the statements of the hypotheses could be somewhat difficult for a novice in this area to fully understand, misinterpretation of hypotheses may be an issue which could undermine the quality of the data. While the coding was not independently double-coded, my own checking of the MSc student's judgements about this was rigorous and so this should help to ensure accuracy of the coding despite the potential subjectivity involved in these decisions.

Another important limitation of this study might be the fact that the SRs were preregistered which may in itself be associated with a lower rate of HARKing and other questionable practices than in the more general literature, because authors who preregister may be more diligent than those who don't, and may also be more aware of the visibility of any changes compared to those who do not have a public protocol and so can commit HARKing without a record of their previous hypotheses being publicly available. Therefore, this study's findings can only be taken as evidence of lower rates of HARKing in RRs than in preregistered SRs, without the ability to generalize the conclusions to more general psychology literature that has not been preregistered (and for which HARKing is undetectable). It is reasonable to speculate that such literature may be even more likely to contain HARKing than the current SR sample, which would be concerning given the extent of this in the current sample; however, without a public record of the pre-specified hypotheses this was not possible to investigate for the more general (non-preregistered) literature. It may be possible however to use grey literature to gain some understanding of HARKing in non-preregistered SRs, for example, if student theses or conference presentations subsequently lead to peer reviewed publications, these may provide an earlier record of the hypotheses that could be compared against those stated in the final manuscript.

8.4.4. Implications and Future Directions

As previously mentioned, a number of research questions proposed in the preregistration but not yet investigated, could provide interesting further insights in the future. For example, as suggested in the original protocol for this study, by creating a new SR sample for which both preregistered protocols and preprints are available, we could compare these documents against final published manuscripts to investigate the extent to which HARKing might be introduced by the peer review process. Additionally, as HARKing can often co-occur with other questionable research practices such as *p*-hacking (Scheel et al., 2021; Rubin, 2019), it could be beneficial to investigate the co-occurrence of *p*-hacking and HARKing and how this compares between RRs and SRs, to provide a deeper understanding of how different QRPs might influence each other, as has been previously suggested by others (Scheel et al., 2021; Rubin, 2019).

8.5 Conclusion

This study aimed to investigate whether RRs appear to be working as intended by being associated with very low rates of apparent HARKing as compared with SRs. The findings reveal no evidence of HARKing in a sample of RRs, and higher rates in SRs. This preliminary study does, however, provide only initial evidence within one specific topic area. Therefore, further research in a larger sample would be beneficial to verify these findings in a larger and more diverse sample of the literature.

8.6. References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS One, 12*(3), e0172792.

Banks, G. C., Rogelberg, S. G., Woznyj, H. M., Landis, R. S., & Rupp, D. E. (2016). Editorial: Evidence on questionable research practices: The good, the bad, and the ugly. *Journal of Business and Psychology*, *31*(3), 323-338. https://doi.org/10.1007/s10869-016-9456-7

Bosco, F. A., Aguinis, H., Field, J. G., Pierce, C. A., & Dalton, D. R. (2016). HARKing's threat to organizational research: Evidence from primary and meta-analytic sources. *Personnel Psychology, 69*(3), 709-750.

Chambers, C. (2017). *The seven deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton University Press.

Chambers, C. D. (2014). Ten reasons why journals must review manuscripts before results are known. *Addiction, 110*(1), 10-11. https://web.s.ebscohost.com/ehost/pdfviewer/pdfviewer?vid=0&sid=39eae983-338e-47af-b829-5e6cd89c5139%40redis

Chambers, C. D., & Tzavella, L. (2020). The past, present, and future of registered reports. https://doi.org/10.31222/osf.io/43298

Claesen, A., Gomes, S., Tuerlinckx, F., & Vanpaemel, W. (2021). Comparing dream to reality: an assessment of adherence of the first generation of preregistered studies. *Royal Society Open Science, 8*(10), 211037.https://doi.org/10.1098/rsos.211037

Field, S. M., Wagenmakers, E., Kiers, H. A., Hoekstra, R., Ernst, A. F., & Van Ravenzwaaij, D. (2020). The effect of preregistration on trust in empirical research findings: Results of a registered report. *Royal Society Open Science, 7*(4), 181351. https://doi.org/10.1098/rsos.181351

Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science, 345*(6203), 1502-1505. https://doi.org/10.1126/science.1255484

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., Marston, C., Powell-Smith, A., Heneghan, C., & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials, 20*(1), 1–16. https://doi.org/10.1186/s13063-019-3173-2

Hardwick, T. E. & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2*, 793–796.

Henriksen, K. & Kaplan, H. (2003). Hindsight bias, outcome knowledge and adaptive learning. *BMJ Quality & Safety, 12*, ii46-ii50.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and Tharking. *Journal of Management, 43*(1), 5-18. https://doi.org/10.1177/0149206316679487

Holst, M. R., Halsberger, M., Yerunkar, S., Strech, D., Hemkens, L. G., & Carlisle, B. G. (2023, 21 February). *Hidden changes to prespecified primary outcomes of clinical trials completed between 2009 and 2017 in German University Medical Centres: A meta-research study*. MedRxiv. https://www.medrxiv.org/content/10.1101/2023.02.20.23286182v1

Kelly, J., Sadeghieh, T., & Adeli, K. (2014). Peer review in scientific publications: Benefits, critiques, & a survival guide. *National Library of Medicine, 25*(3), 227–243. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4975196/

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196-217. https://doi.org/10.1207/s15327957pspr0203_4

Leung, K. (2011). Presenting post hoc hypotheses as a priori: Ethical and theoretical issues. *Management and Organization Review*, *7*(3), 471-479. https://doi.org/10.1111/j.1740-8784.2011.00222.x

Mathieu, S., Boutron, I., Moher, D., Altman, D. G., & Ravaud, P. (2009). Comparison of registered and published primary outcomes in randomized controlled trials. *JAMA, 302*(9), 977–984.

Mathieu, S., Chan, A-W, & Ravaud, P. (2013) Use of trial register information during the peer review process. *PLoS ONE*, *8*(4), e59910. doi:10.1371/journal.pone.0059910

Noor, I. (2020, June 10). *Confirmation bias*. Study Guides for Psychology Students - Simply Psychology. https://www.simplypsychology.org/confirmation-bias.html#:~:text=%20Confirmation%20Bias%20%201%20Confirmation%20bias%20is,interpret%20it%20in%20a%20biased%20way.%20More%20

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D. P., Hesse, B., Humphreys, M., … Yarkoni, T. (2015). Promoting an open research culture. *Science*, *348*(6242), 1422-1425. https://doi.org/10.1126/science.aab2374

O'Mahony, A., Morey, C., & Chambers, C. (2021). *Comparative database of registered reports and baseline articles*. [Unpublished database]. School of Psychology, Cardiff University

Patil, P., Peng, R. D., & Leek, J. T. (2016). *A statistical definition for reproducibility and replicability*. BioRxiv https://doi.org/10.1101/066803

Roese, N. J., & Vohs, K. D. (2012). Hindsight bias. *Perspectives on Psychological Science, 7*(5), 411–426. https://doi-org.abc.cardiff.ac.uk/10.1177/1745691612454303

Rubin, M. (2017). When does harking hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, *21*(4), 308-320. https://doi.org/10.1037/gpr0000128

Rubin, M. (2019). The costs of HARKing. *The University of Chicago Press*. https://doi.org/10.1093/bjps/axz050

Scheel, A. M., Schijen, M. R., & Lakens, D. (2021). An excess of positive results: Comparing the standard psychology literature with registered reports. *Advances in*

Methods and Practices in Psychological Science, *4*(2), 251524592110074. https://doi.org/10.1177/25152459211007467

TARG Meta-Research Group and Collaborators (2022). Discrepancy review: a feasibility study of a novel peer review intervention to reduce undisclosed discrepancies between registrations and publications. *Royal Society Open Science, 9*(7), 220142. https://doi.org/10.1098/rsos.220142

The jamovi project (2023). *jamovi* (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org

White, M. (2022). *Are Registered Reports Immune to HARKing? A Comparative Analysis of Hypotheses in Registered Reports and Standard Reports in Cognitive Psychology and Neuroscience*. [Unpublished MSc thesis]. Cardiff University.

Chapter 9: Descriptive Analysis of Total RR Sample

9.1. Introduction

As outlined in the previous chapters, RRs are thought to be associated with greater rigour, quality, and transparency than standard research reports. The previous chapters have provided some evidence to support this, including lower rates of supported hypotheses, higher rates of data, code and materials sharing, higher rates of replication studies and the inclusion of exploratory analysis as clearly distinguished from confirmatory analysis. Overall, much of the evidence thus far supports the view that RRs are associated with greater rigour and transparency. However, the sample of RRs in these comparative analyses was restricted to only 170 articles due to feasibility constraints. This chapter aimed to use a much larger sample of RRs to examine some descriptive statistics within the RRs alone, in order to gain a better understanding of what these reports actually consist of. The chapter also briefly considers whether the descriptive statistics for the key characteristics of interest for both samples suggest that the 170 articles used as the comparative RR sample, appears to be representative of the larger sample of RRs gathered overall (n = 359).

9.1.1. Research Questions

Questions examined in this chapter are as follows. As the focus was on descriptive findings, these questions were exploratory and did not have specific predictions about the outcomes.

- Question 1: What are the rates of supported and non-supported hypotheses within this sample of RRs?
- Question 2: What is the rate of sampling plans in this sample of RRs?
- Question 3: What is the rate of replication studies included within this sample of RRs?
- Question 4: What is the rate of exploratory analysis included in this sample of RRs?
- Question 5: What is the rate of manipulation checks included in this sample of RRs?
- Question 6: What is the rate of non-registered studies included within this sample of RRs?
- Question 7: What are the rates of data, code, materials, and protocol availability in this sample of RRs?

9.2. Methods

9.2.1. Overview of Coding Process and Variable Creation

As outlined in chapters 3 to 6, characteristics of the articles were initially coded according to a detailed protocol which can be found in Appendix 2 and also on the online repository located here: https://osf.io/5pu4g/. This coding had been completed for a sample of 359 RRs. However, due to the time taken for the matching and coding processes, only a subset of 170 of those RRs had their SRs matched, coded, and checked towards the end of this project's original timeline. The decision was therefore taken to restrict the comparative analysis to that set of 170 RRs and their 340 SRs only. As the larger total sample of 359 RRs had already been coded, however, the coding of these was also checked and the data converted into the variables for analysis so that these could still be examined for some key characteristics of interest, which will be described in this chapter.

9.2.1.1. Hypothesis Variables

9.2.1.1.1. Hypothesis Support

As outlined in chapter 3, the presence of clearly and partially identifiable hypotheses had been documented, followed by whether these were supported. Whether hypotheses had been supported was initially coded at each of the three levels of the paper (article, study, and hypothesis level), as applicable. Response options were Yes, No, Partially, N/A or Unclear and further details about these response options can be seen in chapter 3 or in the coding protocol. Where support was not stated explicitly, the research findings had to be interpreted to determine whether they appeared to have supported the hypothesis. Based on the information gathered during the initial coding process, five variables were created at each of the three levels of the paper and at an overall level. However, only the overall-level variables were examined in this chapter. Values for each variable were given as proportion scores. These represented the total proportion of hypotheses across each paper that were either fully supported, partially supported, fully or partially supported (combined), or not supported. The median proportions were examined for the rates of fully supported hypotheses, of combined fully and partially supported hypotheses (referred to simply as 'supported'), and of non-supported hypotheses.

9.2.1.1.2 Granularity of Hypotheses

The granularity of the hypotheses was not explicitly coded for during the initial coding process. However, as outlined in chapter 3, the structure of the hypothesis coding across the three levels of the paper was clear from the output of the initial coding process and so this

was used to create specific variables reflecting the granularity of the hypothesis structure within each paper. Three variables were therefore created to reflect this structure: the proportion of hypotheses in the paper that were stated at article level, the proportion of hypotheses in the paper that were stated at study level, and the proportion of hypotheses in the paper that were stated at hypothesis level. Median proportions were examined for each of these three variables.

9.2.1.2. Study Characteristics

9.2.1.2.1. Sampling Plan

Initial data was gathered for what kind of sampling plan was used, as outlined in chapter 5. Response options used at that stage were as follows: Frequentist power, Bayesian, Other, Unclear, and N/A. 'Other' approaches to the sampling plan would include being based on the sample size from a previous study (when this has not been used to inform a power calculation), or if recruitment plans were limited by resource constraints, as outlined in chapter 5. The use of a sampling plan was initially coded at each of the three coding levels as applicable, depending on how this information had been presented in the paper, but variables for the analysis were ultimately created at a broader overarching level encompassing each paper in its entirety rather than sub-divided into levels.

Based on the information gathered during the initial coding process, two categorial variables were created, one of which indicated whether any kind of sampling plan had been used, while the other indicated whether there was a statistical sampling plan used. Additional categorical variables documented whether there was a power calculation, a Bayesian approach, any 'other' approach, or if the approach was unclear. Frequencies were examined for each of these categorical variables in order to see how common each type of sampling plan was among the RRs.

9.2.1.2.2. Inclusion of Replication Studies

Studies within each paper were coded for whether the study was (or appeared to be) original or a replication attempt. Replications could be direct or indirect and could be either a replication of a different study (an external replication), or an internal replication of their own earlier study in the same paper. Response options used were as follows: original, direct replication, indirect replication, direct internal replication, and indirect internal replication. Further details on these response options are available in chapter 5 (section 5.2.7.). As

outlined in that section, it was often challenging to determine which category was most appropriate.

Five categorical variables were created based on the data gathered during the initial coding process. One of these was used to indicate whether any kind of replication was included in each paper. Four categorial variables were then created to indicate the presence of each kind of replication specifically: direct (external) replications, indirect (external) replications, direct internal replications, and indirect internal replications. Frequencies were examined for each of these categorical variables.

9.2.1.2.3. Inclusion of Exploratory Analysis

At each of the three levels, as applicable, information was gathered for whether there was, or appeared to be exploratory analysis. This was accounted for regardless of whether the rest of that paper or study also contained any hypotheses i.e., whether there was also a confirmatory aspect of the paper. Response options used for this were 'yes', 'no', 'unclear', or N/A. Unless there was a clear distinction between confirmatory and exploratory/additional parts, it could sometimes be difficult to tell whether and where there was a difference in this and so this sometimes needed to be inferred. This characteristic was initially coded at each of the three sub-levels, but the variable ultimately created for this was at the overall-level across the papers. At this overall level, one categorical variable was created to reflect whether exploratory analysis had been included in the paper.

Where exploratory analysis was considered to have been included, the nature of this was coded as either general exploration, follow up on hypotheses, both, or unclear. Determining or inferring the nature of the exploratory analysis was sometimes challenging as the distinction between the two could be highly subjective. The nature of the exploratory analysis was initially coded at each of the three sub-levels as applicable. Subsequently though, variables were only created for the analysis at the overall level, although this did reflect the data that had previously been coded at each of the three sub-levels. Two categorical variables were ultimately created for the nature of the follow up analysis; one of these indicated whether the exploratory analysis was follow-up, and another indicated whether this was general exploration. Frequencies were examined for whether exploratory analysis had been included in the paper and also for whether this exploratory analysis was general exploration or follow up analysis.

9.2.1.2.4. Inclusion of Non-Registered Studies

As outlined in chapter 5, data was initially coded for whether each study had been pre-registered or not. Response options for this were RR, 'preregistered non-RR', 'non-preregistered non-RR', or unclear. Only the rate of non-preregistered non-RR studies was considered for this study; to investigate this, a categorial variable was created to indicate whether there was any non-registered study included in each paper and frequencies were examined for this.

9.2.1.2.5. Inclusion of Manipulation Checks

As described in chapter 5, the use of manipulation checks was initially coded at each of the three sub-levels, as applicable, with response options being either 'yes', 'no', 'unclear' or N/A. As mentioned in chapter 5, determining whether manipulation checks had been included was often considered challenging and/or subjective. An overall-level categorical variable was created based on the data gathered during the initial coding process, and this represented whether any manipulation checks had been included in the paper. Frequencies were examined for this variable.

9.2.1.3. Author Demographics

9.2.1.3.1. Author Seniority

As described in chapter 6, content analysis was used to gather data on seniority of the authors for each of the included articles. Categorical variables were created for whether the author had a PhD, and whether they could be considered an early career researcher (ECR). For further information see chapter 6 (section 6.2.1.) This chapter will focus only on the first author for each of these variables.

9.2.1.3.2. Non-Western Authors

As mentioned in chapter 6, the countries the authors were based in were initially coded using the information provided about each authors' institutional affiliation. All the countries represented in each paper were recorded, although each was only recorded once per paper so if multiple authors were based in the same country, this country was only documented once. A categorical variable was created for whether there was a non-Western country represented in the paper, and this was coded as either 0 (no) or 1 (yes). The creation of these variables did not take into account the number of authors from each country; whether there was a single author or multiple authors from a particular country, this country was only represented once per paper. Frequencies were examined for this variable.

9.2.1.4. Availability of Data and Study Materials

As outlined in chapter 4, data was initially gathered regarding the availability, and the use of availability statements, for each articles' data, analysis code, materials, and protocols. For protocol availability, this was coded as yes or no. For data availability, two pieces of information were included: whether there was a statement about the availability of the data, and how available the data was. The latter characteristic initially had a range of possible options: publicly available, partially available, already available, gated, unclear, not available with justification, or not available without justification. For each of the other three characteristics (availability of analysis code, digital materials, and original non-digital materials), three pieces of information were coded: whether this type of item was used in the study, whether there was a statement about its availability, and how available it was, using the same range of response options as listed above for the data availability characteristic.

Categorical variables were created for each of these characteristics. Specifically, a categorical variable was created for protocol availability whereby 0 indicated 'no' and 1 indicated 'yes. The presence of an availability statement (for the data, code, and digital and non-digital materials, respectively) was coded as 1 while the absence of such a statement was coded as 0. For the actual availability of these items, multiple categories had been used during the initial coding process. These were condensed so that those initially coded as publicly available, gated, partially available, or already available were grouped together as being 'available' and coded as 1, while those initially coded as available on request, not available with justification, not available without justification, or unclear, were all grouped together as 'not available', and coded as 0. The frequencies were examined for each of these categorical variables.

9.3. Results

9.3.1. Hypothesis Statement and Support

In terms of the granularity of the hypotheses, most were stated at hypothesis level (M = 0.91, SD = 0.21, median 1), while only a small proportion were stated at either article (M = 0.07, SD = 0.20, median 0) or study level (M = 0.02, SD = 0.07, median 0). The total proportions of hypotheses at each level are shown in Figure 18. There was a higher average proportion of clearly identifiable hypotheses (median 1, M = 0.63, SD = 0.45) than of partially identifiable hypotheses (median 0, M = 0.37, SD = 0.45). Only 7.17% of the articles included any competing hypotheses.

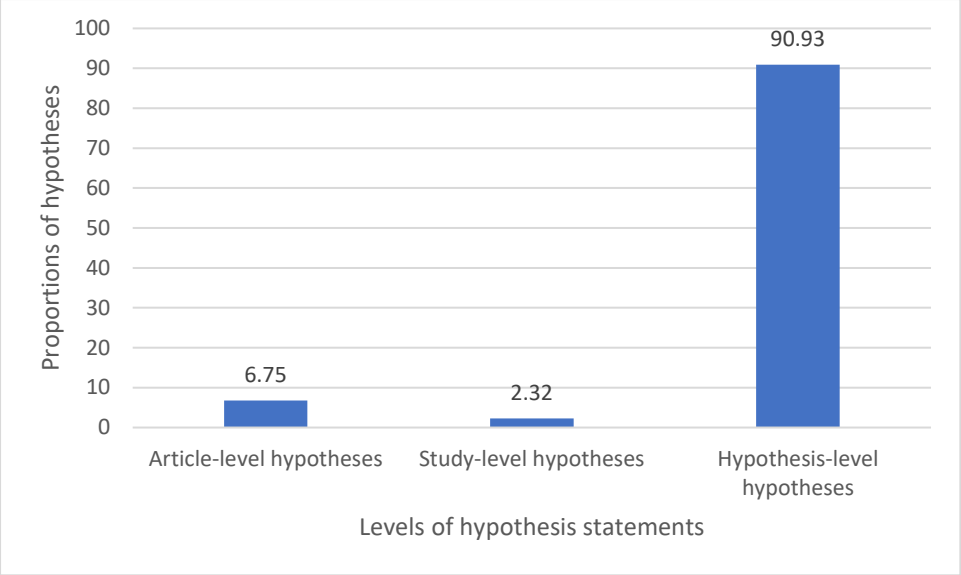

Figure 18: Total proportions of hypotheses at each level of the full RR sample. As with the samples in the comparative analysis from chapter 3, the vast majority of the hypotheses were stated at hypothesis level.

Less than half of the hypotheses were found to be supported; this is demonstrated by the median score of 0.44 (*M* = 0.45, *SD* = 0.37) when the fully and partially supported hypotheses were combined in one score. When the analysis was restricted to only the hypotheses that had been coded as being fully supported, a median score of only 0.25 (*M* = 0.33, *SD* = 0.36) was found. A median score of 0.5 (*M* = 0.50, *SD* = 0.38) was found for the hypotheses that were not supported, while the remaining hypotheses whose support was coded as unclear made up only a small proportion of the overall sample (median 0, *M* = 0.06, *SD* = 0.17). The total proportions for each type of hypothesis support are outlined in Figure 19 below.

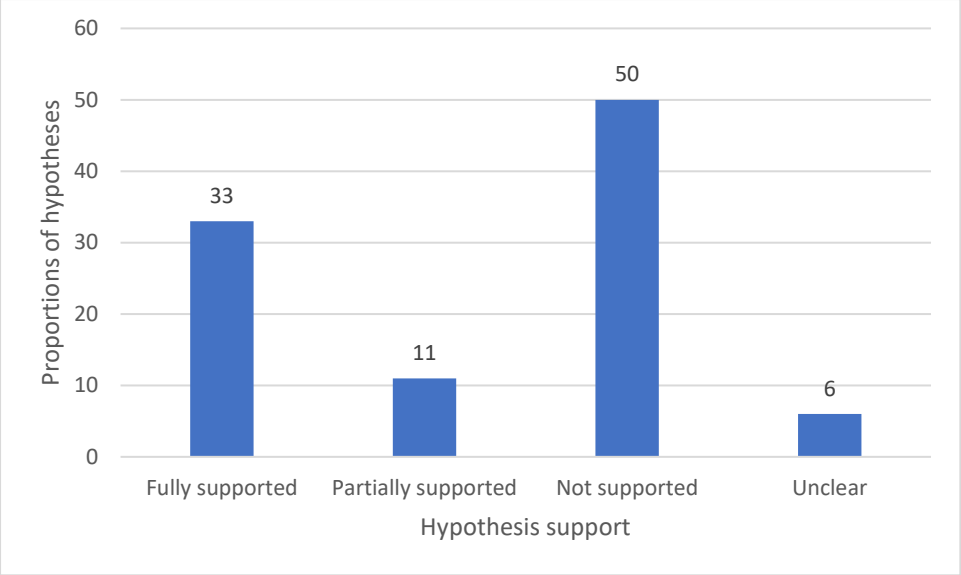

Figure 19: Total proportions (%) of hypothesis support in the full RR sample.

9.3.2. Study Characteristics

9.3.2.1. Sampling Plan

The majority of the RRs had some kind of sampling plan (72.42%), although just under half of the total sample had a statistical sampling plan (46.80%). When the specific types of sampling plans were examined, 41.23% used a power calculation, 6.13% used Bayesian sampling, and 29.25% used some other kind of sampling plan such as resource constraints or other practical considerations. These last three categories were not necessarily mutually exclusive as papers containing multiple studies could potentially use different sampling approaches in different studies.

9.3.2.2. Replication Studies

Just under half of the RRs included at least one replication study of some kind (48.47%), with 29.25% being direct external replications. A further 18.11% were indirect external replications, while few included internal replications (4.74% direct internal replications, and 5% indirect internal replications). These categories were not necessarily mutually exclusive since the same article could potentially contain more than one kind of replication study.

9.3.2.3. Exploratory Analysis

Most RRs included some exploratory analysis (71.87%) and the nature of this exploratory was evenly split between follow up analysis and general exploration with both having a median of 0.5.

9.3.2.4. Other Study Characteristics

Only 35.66% of the RRs appeared to have included manipulation checks. Relatively few of the RRs included any non-registered studies (8.65%).

9.3.3 Author Demographics

RRs had a mean of 7.39 authors (*SD* = 17.73, median = 4), and only 9.19% of RRs had authors based in a non-Western country. Of the articles for which this information could be obtained, the majority of the first authors appeared to have had a PhD (67.59%), and just over half (52.94%) appeared to be early career researchers.

9.3.4. Availability of Data, Code, Materials, and Protocols

The majority of the RRs (67.69%) had a statement regarding the availability of their data, and most also had their data available (65.18%). Slightly more than half of the total RR sample (n=193, i.e., 53.76%) reported using analysis code or could be inferred to have used it. Of the articles that had used analysis code, most of these contained a statement about the availability of this code (76.17%). Furthermore, a total of 88.60% of the RRs that used analysis code had made this available.

Of the 268 RRs that used digital materials, just over half of these (51.87%) had a statement about the availability of the materials. Furthermore, just slightly over half of the RRs that used digital materials had made them available (52.61%). Only 38 of the RRs used original non-digital materials. Of these, 55.26% had a statement about the availability of these materials and slightly more than half of these RRs (55.26%) had made their non-digital materials available.

Finally, the majority of the RRs (70.47%) had available protocols.

9.3.5. Exploratory Analysis: Comparisons between RR Samples.

In order to determine whether the smaller sample of RRs used in the comparative analyses of the earlier chapters (typically n = 170) seem representative of the larger population of RRs, the key descriptive statistics were compared between the two RR samples. The results are outlined in Table 11 and show that the rates of these characteristics are very similar for both RR samples. Any differences that were present between the two samples tended to be minimal. However, specific characteristics that were somewhat higher in the comparative sample of RRs included the overall use of any kind of sampling plan (80.59% vs. 72.42%), and the rates of open practices, i.e. sharing of data, code and materials and including specific statements about the availability of these items.

Table 11

Comparisons of characteristics between RR samples

Characteristic	n for total RRs in chapter/ analysis	Result in chapter	n for total RRs in chapter 9	Result in chapter 9
Hypothesis variables				
Granularity of hypotheses	n = 140	Article level: M = 0.08, SD = 0.23. Study level: M = 0.03, SD = 0.08. Hypothesis level: M = 0.89, SD = 0.24.	n = 307	Article level: M = 0.07, SD = 0.20. Study level: M = 0.02, SD = 0.07. Hypothesis level: M = 0.91, SD = 0.21.
Proportion of clearly vs. partially identifiable hypotheses	N = 140	Clearly identifiable: M = 0.69, SD = 0.43 Partially identifiable: M = 0.32, SD = 0.43	n = 307	Clearly identifiable: M = 0.63, SD = 0.45 Partially identifiable: M = 0.37, SD = 0.45
Inclusion of competing hypotheses	N = 140	7.86%	N = 307	7.17%
Hypothesis support (fully supported)	N = 140	37.31%	N = 307	33%
Hypothesis support (fully and partially combined)	N = 140	47.59%	N = 307	44%
Unsupported hypotheses	N = 140	47.42%	N = 307	50%
Hypothesis support unclear	N = 140	4.98%	N = 307	6%
Sampling characteristics				
Sampling plans (any)	N = 170	80.59%	N = 359	72.42%
Statistical sampling plans	N = 170	50.59%	N = 359	46.8%
Power calculation	N = 170	42.94%	N = 359	41.23%
Bayesian sampling	N = 170	8.24%	N = 359	6.13%
Other sampling plan (e.g. resource constraints, etc.)	N = 170	32.94%	N = 359	29.25%

Replication characteristics				
Replication included (any)	N = 170	50%	N = 359	48.47%
Direct external replication	N = 170	28.24%	N = 359	29.25%
Indirect external replication	N = 170	21.77%	N = 359	18.11%
Direct internal replication	N = 170	5.29%	N = 359	4.74%
Indirect internal replication	N = 170	3.53%	N = 359	5%
Other methodological characteristics				
Exploratory analysis included	N = 170	75.29%	N = 359	71.87%
Manipulation checks included	N = 170	33.53%	N = 359	35.66%
Non-registered studies included in paper	N = 170	9.41%	N = 359	8.65%
Author demographics				
Number of authors	N = 170	M = 8.02, SD = 22.0, median = 4	N = 359	M = 7.39, SD = 17.73, median = 4
Non-Western country represented	N = 170	8.82%	N = 359	9.19%
First author had a PhD	N = 144	64.58%	N = 359	67.59%
First author was an ECR	N = 143	55.94%	N = 359	52.94%
Open Research Practices				
Data availability statement included	N = 170	74.12%	N = 359	67.69%
Data available	N = 170	70%	N = 359	65.18%
Analysis code used	N = 170	50.59%	N = 359	53.76%
Analysis code availability statement included	N = 86	82.56%	N = 193	76.17%
Analysis code availability	N = 86	89.54%	N = 193	88.6%
Digital materials used	N = 170	83.53%	N = 359	74.65%
Digital materials availability statement included	N = 142	57.75%	N = 268	51.87%
Digital materials available	N = 142	56.34%	N = 268	52.61%
Non-digital materials included	N = 170	14.12%	N = 359	10.58%

Non-digital materials availability statement included	N= 24	70.83%	N = 38	55.26%
Non-digital materials available	N = 24	70.83%	N = 38	55.26%
Protocol available	N = 170	75.88%	N = 359	70.47%

9.4. Discussion

9.4.1. Recap of Results

Overall, the patterns of descriptive results for the larger RR sample indicated high rates of characteristics associated with greater rigour, with RRs having higher rates of unsupported than supported hypotheses, and RRs being more likely to have some kind of sampling plan than to not have one, although there is some room for improvement in this, particularly in the rates of statistical sampling plans which were found in a little under half of the RRs. A considerable proportion of the articles included some form of 'other' or non-statistical sample plan. The majority of RRs had made their data, analysis code, and protocols available, while only just over half had made their digital materials available.

Early career researchers are fairly well represented among the first authors of RRs, although those who did not yet (appear to) hold a PhD made up a smaller proportion of the sample. Just under 10% of the articles had authors that appeared to be based in a non-Western country.

Finally, most of the characteristics examined show similar rates in the smaller comparative sample of RRs, as in the larger sample of RRs reported in this chapter. Where differences were noted, these were primarily in relation to open research practices, and the use of sampling plans, each of which was somewhat higher in the comparative sample of RRs than in the larger sample of RRs as a whole. However, these differences in the descriptive statistics for these two groups were still small and were not tested statistically. The slightly higher rates of these practices in the comparative sample may be a reflection of the disciplines included in the two samples. Specifically, the majority of the RRs included in the comparative sample were from psychology, with the rest from closely-related areas, whereas the broader sample used in this current chapter included RRs from more diverse areas, such

as cancer biology, politics, and finance. Use of the aforementioned practices may be less common in such disciplines, which could account for the slightly lower rates overall for this larger and more diverse sample as a whole.

9.4.2. Comparison to Previous Literature

The pattern of results for the levels of support for the hypotheses were in line with expectations, with only 25% of these fully supported, and 50% not supported, although when fully and partially supported hypotheses were combined this overall rate for support did increase. Even so, this indicates lower rates of supported hypotheses than non-supported hypotheses, particularly when using the more stringent score of fully supported hypotheses only. These findings are in line with the rates of hypothesis support reported in other studies. For example, Scheel et al. (2021) reported very similar rates of supported hypotheses at 44%. Rates of non-supported findings reported by Allen and Mehler (2019) were slightly higher than those found in the current study (60.5% in their work, compared with 50% in the current study) but these findings are still broadly in line with each other. These similarities are encouraging, particularly the similarity with Scheel's results. The current sample represents a much larger number of RRs than that included in any previous study, as Scheel et al.'s main analysis was based on only 71 RRs while Allen and Mehler's study had 127 papers, providing 153 hypotheses. The current study's sample of 359 RRs is therefore a considerable increase and so should provide a more comprehensive view of the available reports.

Furthermore, the majority of the RRs had some kind of sampling plan although this is slightly lower than the rates found by Soderberg et al. (2020) who reported 90% of their RR sample having some kind of sampling plan. However, their sample of RRs was much smaller than in the current study and so the larger sample size may have allowed for greater variation. Furthermore, Soderberg et al. only examined this characteristic in the last study of each paper, while the current study considers this across the paper as a whole, which may help to explain the differing findings. Just under half of the total sample in the current chapter had a statistical sampling plan, which is similar to that found in the comparative sample studied in chapter 5. This is slightly higher than that reported by Norsa (2022) who found that 36.95% of the RRs they studied had what they referred to as a well-planned power analysis based on the exiting literature and/or pilot studies. However, Norsa's study only considered 46 RRs in total. Like Norsa (2022) claims, many of the RRs did not specify how sample sizes had been planned, and in cases where they did, if a calculation was performed, many did not specify

details like what software had been used for their power analysis, an issue that may hinder efforts to reproduce those power analyses if needed.

Just over one-third of the RRs appeared to have included manipulation checks, which is similar to the rates reported by Hauser et al. (2018) in relation to the general literature, although less than rates reported by Ejelov and Luke (2020). Due to the speculated greater degree of attention to methodological rigour in RRs than in standard research reports, it might be expected that the rates would be higher in RRs. However, this may be due to differences in the coding approaches taken across different studies, or due to the differing journals, sub-disciplines of psychology, and study designs represented in these different studies.

Most RRs included some exploratory analysis and the nature of this was evenly split between follow up analysis and general exploration. However, as stated in chapter 5, the coding of the nature of this type of analysis was highly subjective and so this should only be taken as a rough estimate of this characteristic. Additionally, no previous studies appear to have considered the extent to which RRs have included non-registered studies and so the rates of these in the current study represent a novel finding. Just under half of the RRs included at least one replication study of some kind with 29% being direct external replications. These rates are similar to that found by Scheel et al. (2021).

Claims that RRs are not accessible to ECRs appear to be contradicted by the rates of first authors who were ECRs, although the criteria used for this in the coding process could have been stricter and if so, this could have given lower estimates. One of the most concerning findings is the low rate of RRs with authors based in non-Western countries. Although this follows trends across scientific publishing more generally, it highlights an important area for further development to investigate and improve the accessibility and use of the format internationally. It is worth noting that using the institution at which the author was based at the time is not necessarily an accurate metric of their own nationality or ethnicity, but given the challenges of trying to determine this using publicly available information, this compromise was considered necessary and a deeper investigation into the diversity and geographic spread of the authors using this report format would be an important future direction.

Rates of data sharing in RRs were similar to findings reports by Obels et al. (2020), but higher than other sources (Vanpaemel et al., 2015) while the rates of code sharing in RRs in the current study were much higher than that found by Obels et al. (2020) or Laurinavichyute

et al. (2022). The current study offers promising evidence of high rates of sharing of data and analysis code, but lower rates of sharing of materials. Due to the limited attention in the literature regarding the sharing of study materials other than data and code, there was limited other work to compare the current study to and so this is an area where more research would be beneficial in order to understand materials sharing practices, barriers, and areas for improvement. The RRs did show much higher rates of sharing of original non-digital materials than of their digital materials, but this was a very small proportion of the overall sample, while studies using digital materials were much more common.

Finally, the rates of protocol availability in the current study were broadly similar to the rates found by Hardwicke and Ioannidis, indicating that most RRs do appear to be associated with a publicly available stage 1 protocol although there is room for improvement in this.

9.4.3. Limitations

Limitations of the methods used to obtain the data in this chapter are the same as those outlined in the chapters 3 to 6, as the same coding approaches were used.

9.4.4. Implications and Future Directions

The findings of this chapter show that while RRs are generally associated with particular characteristics indicating greater rigour and transparency, some improvements are still needed. For example, there is a need for greater standards or requirements for sampling plans, whether those are statistical plans or not, in order to justify their approach and try to avoid underpowered studies. While the use of non-statistical sampling plans and justifications (as found in approximately 29% of the RRs in this sample) can also be important, particularly when working with hard-to-reach populations or within resource constraints, a statistical plan would still be preferable where possible in order to allow readers to know what sample size would be needed for what level of power in order for them to be able to evaluate outcomes more thoroughly. Alternatively, a statistical power calculation could potentially be used alongside non-statistical aspects of sample planning to at least give an indication of the power of any sample that has mainly been restricted by more practical logistical issues. Increasing the rates of statistical sampling plans whenever possible is therefore an important area for improvement and future work and where this is not possible, some form of sample planning should be considered beforehand, and should be acknowledged in the final paper. Likewise, although rates of availability of the data, code, and materials are quite high, consideration should also be given to how this might be increased. Furthermore, the rates of protocol

availability for the RRs is less than optimal, as these should be made available for all published RRs to ensure transparency. Efforts to increase these practices should therefore be encouraged.

The geographic diversity of the authors was very limited, highlighting an important area for future work. Not only is this important for future research but it should also be taken on board as an important consideration for journal editors, and open science advocates in general.

9.5 Conclusion

Overall, the findings in this chapter shed some light on how particular characteristics are represented within RRs, with these findings coming from a larger sample than has been included in any previous study of RRs. They indicate that there is further work needed to improve rates of certain key practices, even when many such practices are quite common among RRs.

9.6. References

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

Ejelöv, Emma & Luke, Timothy. (2019). "Rarely safe to assume": Evaluating the use and interpretation of manipulation checks in experimental social psychology. *Journal of Experimental Social Psychology, 87*. 10.1016/j.jesp.2019.103937.

Hardwicke, T.E., Ioannidis, J.P.A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2,* 793–796

Laurinavichyute, A., Yadav, H., & Vasishth, S. (2022). Share the code, not just the data: A case study of the reproducibility of articles published in the Journal of Memory and Language under the open data policy. *Journal of Memory and Language, 125*, 104332. https://doi.org/10.1016/j.jml.2022.104332

Norsa, R. (2022). *Do registered reports improve sample size planning in psychology? An exploratory study*. Thesis. Padua Thesis and Dissertation Archive. https://thesis.unipd.it/handle/20.500.12608/32381

Obels P, Lakens D, Coles NA, Gottfried J, Green SA. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in*

Methods and Practices in Psychological Science, 3(2), 229-237. doi:10.1177/2515245920918872

Scheel, A. M., Schijen, M. R. M. J., & Lakens, D. (2021). An excess of positive results: comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Soderberg, C.K., Errington, T.M., Schiavone, S.R., Bottesini, J., Thorn, F. S., Vazire, S., Esterling, K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour, 5*, 990–997. https://doi-org.abc.cardiff.ac.uk/10.1038/s41562-021-01142-4

Vanpaemel, W., Vermorgen, M., Deriemaecker, L., &Storms, G. (2015). Are we wasting a good crisis? The availability of psychological research data after the storm. *Collabra: Psychology, 1*(1), 3. https://doi.org/10.1525/collabra.13

Chapter 10: General Discussion

10.1 Recap and Contextualisation of Key Findings

The current study aimed to determine the feasibility of creating a large comparative database of stage 2 RRs and a yoked sample of standard reports (SRs), and to investigate how these two article types differ on indicators of greater rigour and transparency. Overall, the matching and coding processes were found to be feasible to implement, although not without occasional challenges. The majority of the articles were considered 'somewhat easy' to code, and the majority of the SRs were considered 'somewhat relevant' or 'very relevant' to their RRs.

Overall, RRs were found to be more strongly associated with a range of positive characteristics compared to SRs. Specifically, RRs showed lower rates of supported hypotheses and higher rates of unsupported hypotheses compared with SRs, which in turn implies a reduction in publication bias and/or reporting bias. These findings are consistent with previous research despite key differences in the approaches used to investigate this (Allen & Mehler, 2019, Scheel et al., 2021). Use of open practices (i.e., the sharing of data, analysis code, materials, and protocols) was higher among RRs than SRs. Rates of data sharing in RRs were consistent with previous research (Obels et al., 2020), while code sharing rates appear to be less consistent with rates that have been previously reported (Obels et al., 2020), albeit still higher among RRs than SRs. Explicit statements about whether data and materials were available were also more common among RRs than SRs, although both article types would benefit from more consistent inclusion of such declarations.

RRs were generally more strongly associated with methodological practices indicative of greater rigour and transparency compared with SRs. This supports previous conclusions drawn by Soderberg et al., (2021) albeit using different methods of evaluation. Comparative analysis of author demographics revealed few differences between the article types, and RRs showed lower rates of ECR first authors than have been previously reported (Chambers, 2019; Chambers and Tzavella, 2022). This analysis also suggested that increasing the geographic diversity of RR authorship may be an important goal for future efforts. Broadly in line with previous research (Hummer et al., 2017), article citation rates showed no significant difference between the two article types. Higher citation rates were at least partially associated with more positive findings and with fewer negative findings within both the SRs

and RRs, but there was no statistically significant relationship between the journal impact factor and whether hypotheses were supported, for either article type. These latter points represent a novel finding in relation to RRs, although the association between citation rates and positive findings has been previously reported for SRs in medical research (Jannot et al., 2013; Misemer et al., 2016; Duyx et al., 2017).

A preliminary analysis of 12 RRs and 12 SRs revealed that HARKing appears to be non-existent in these RRs, while some evidence of HARKing was observed in the SRs. These findings suggest that RRs may be performing as hoped in reducing this questionable research practice. However, as a small pilot study, this preliminary observation is subject to replication.

Finally, to provide a broader overview of how common various practices are within RRs, some brief descriptive analysis was conducted on a range of characteristics in the larger sample of RRs only (n = 359). For example, analysis of protocol availability within the RR-only sample supported previous concerns that stage 1 protocols are not consistently available for all stage 2 RRs, although they are available for most (Hardwicke and Ioannidis, 2018). Increasing protocol availability, and perhaps improved standardisation of the available protocols, would be a useful future goal, as also indicated by Hardwicke and Ioannidis (2018) and Chambers and Mellor (2018).

Furthermore, the consistency between the two RR samples in the rates of most characteristics suggests that the smaller comparative sample used in most of the chapters was generally representative of the larger sample of RRs reported on in chapter 9 and so some of the conclusions from the earlier chapters might generalise more broadly, though further investigation of this would be warranted. Where some small differences were observed between the two RR samples in certain characteristics (e.g., use of open practices and of any kind of sampling plan), the lower rates of these in the larger RR sample might be explained by the broader range of study designs and disciplines represented in this sample, compared with the more psychology-focused comparative sample. As most of the current research on RRs has focused primarily on psychology, further exploration of RRs across other disciplines could provide useful insights into how the format has been implemented more widely and how their characteristics may differ across disciplines.

10.2 Strengths and Limitations

One strength of the current work is the precision with which SRs were matched to RRs, which was more consistent and detailed than in previous studies (e.g., Scheel et al., 2021; Soderberg et al., 2021). The use of a specific metric for the perceived relevance of each SR to their specific RR also enabled us to monitor how closely matched each SR was considered to be. While this metric was very broad and subjective to apply, it does provide an indication of the article's overall relevance, and for each of the matching criteria.

The study also has a larger sample than that used in previous studies (Allen & Mehler, 2019; Scheel et al., 2021; Soderberg et al., 2021; Norsa, 2022), thereby lending greater statistical sensitivity to the analyses and representing a broader section of the relevant literature. While some of the analyses were conducted on restricted sub-samples where needed, the overall sample size for most of the analyses reflected an improvement on previous studies. However, extending this work in the future with a larger sample would be beneficial to further verify the findings. The sample could also be considered more diverse than that of previous studies, which primarily focus on psychology (Scheel et al., 2021; Norsa, 2022; Soderberg et al., 2021), although Allen and Mehlers's (2019) work did also consider some biomedical papers. As with previous work, the current study did focus primarily on psychology and neuroscience but also included healthcare research published in the *Journal of Medical and Internet Research*, as well as articles from disciplines related to psychology, such as education. This therefore provides a slightly more diverse overview of how the format has been used across different, although still closely-related, areas. Further examination of these characteristics in an even larger and more diverse sample would help to build on this further, although obtaining expert review of the coding would be beneficial for some of the disciplines involved. The descriptive overview of the larger sample of RRs reported in chapter 9, does include a much broader range of disciplines, including some reports from areas like cancer biology, finance, and politics.

Additionally, some previous studies considered only parts of the articles they included. For example, Soderberg et al.'s (2021) study considered only the last study reported in each article as that was most likely to be the study that received the in-principle acceptance (IPA) decision, whereas earlier studies included in the articles are often pilot studies or other preliminary work that was not subject to the in-principle acceptance decision. Furthermore, Scheel et al. (2021) restricted the hypotheses that they examined, considering only the first hypothesis stated in the article, not all of those included. Therefore, the broader focus of the

current project offers a more comprehensive overview not only of the RR format but on the features that constitute each report as a whole. This means that the characteristics examined in this work reflect all studies within each article type, not just the unique RR reviewed studies. Arguably, it may be more appropriate to only consider the specific studies within an RR paper that have actually been pre-reviewed and were subject to in-principle acceptance. However, as readers would likely read and evaluate most papers as a whole, determining the overall qualities of the paper across all the different studies included may be a more rounded and useful approach. At the same time, the use of this approach may make comparison of the findings with other studies of RRs more challenging.

While most previous studies considered a very limited number of characteristics, the current study takes a more holistic approach, providing a broader overview of various aspects of the articles. Soderberg et al.'s (2021) study does constitute something of an exception to this as they assess a broad range of different criteria, but their study generally required assessors to make more subjective value judgements about the articles (e.g., the study's overall rigour, creativity, etc.), rather than more objective recording of a particular characteristic.

While any differences found between the article types may indicate a reduced level of bias in the conduct and reporting of the studies, this work does rely on observational data which is a major limitation since causal impacts cannot be inferred from this dataset. Nevertheless, it is hoped that this study will give a promising indication of the differences between these article types and provide a useful overview of the characteristics of this publishing format. In the future, a randomised controlled trial would be beneficial to establish causal impacts of the RR format (see section 10.3.3.4 below).

Additionally, the differences observed between the article types could be due to other factors. For example, the RR format may appeal to researchers who are more conscientious about rigorous practices, thereby contributing to higher standards in the RRs than in the SRs. As the SRs used in this work were not matched on the basis of authors, this possibility cannot be ruled out and future replication using this author-focused approach may be informative. For example, this could involve not just having RRs and SRs matched by author, but also considering whether the SRs that have been previously published by the authors of RRs demonstrate greater rigour and openness than SRs published by other authors. However, in this case it was considered more important to match the SRs from within the same journals and timeframe as the RRs were published in, to account for potentially differing journal

policies and practices which may otherwise have led to excessive variation between the articles.

As mentioned in chapter 3, the coders' awareness of the article types being coded means there is potential for bias in our interpretations of the article characteristics. The potential for awareness of the article type to influence judgements of the articles can also be seen in Soderberg's study where the blinding of the reviewers does not appear to have been very successful, with a majority of reviewers guessing correctly whether the article they had been evaluating was an RR or not. Given the overwhelmingly positive assessments of RRs in their study, compared to the assessments of the SRs, it is possible that this lack of appropriate blinding may had influenced the reviewers' judgements. Likewise, such bias could have crept into the judgements made by the coders in the current study, particularly myself as I conducted most of the coding and checking. Since the coding processes for most of the characteristics were mostly based on factual recording of data rather than value judgements like those used by Soderberg et al., this may be less of a concern in the current study, although some subjectivity was required in interpreting whether some characteristics were present, whether hypotheses were clearly stated and supported, and so forth. In any case, it was not considered feasible within the current study for the coders to be blinded to the article type, particularly as, in order to be truly blinded to this, certain information would need to be redacted from the articles, especially from the RRs. This was also highlighted by Scheel et al. (2021) as a limitation of their study. Blinded experimental or quasi-experimental approaches to the coding of articles could be very beneficial in the future to help overcome this and build on current evidence, although it would be challenging to implement.

Additionally, although care was taken to try to develop a protocol that could be applied to any type of design, there were some study approaches that did not fit quite as well within parts this coding framework, such as systematic reviews or qualitative studies. While these do not constitute a large part of the study sample, it does highlight the diversity of approaches that use this format and that some characteristics examined may be less relevant for certain study designs.

Finally, a major challenge in conducting this study was accurately identifying which articles were actually RRs and which were not. A considerable number of standard articles were incorrectly labelled as RRs in various journals, particularly the *Journal of Research in Personality* and the *International Journal of Psychophysiology*. Furthermore, the term

‘Registered Report’ is used much more loosely by the *Journal of Medical Internet Research* (JMIR), so that it is used to refer to any study with a protocol published in any journal, not just one published by the journal JMIR Research Protocols and therefore subject to in-principle acceptance at another JMIR journal. This overall lack of clarity in labelling and metadata led to many articles being excluded during the initial sampling process after they had initially been thought to be RRs (e.g. 17 articles from the *International Journal of Psychophysiology* were excluded at this early stage as they had been labelled as RRs but were found not to be). Furthermore, a large number of articles were included and coded but then had to be excluded much later in the process as it was only revealed much later that they had not actually been RRs, despite being labelled as such. This was a notable issue at the *Journal of Research in Personalit*y, and at JMIR journals, with approximately 40 inappropriate RRs from JMIR journals excluded by the time the overall coding process had been completed. While there were occasionally issues with articles that were in fact RRs not being clearly labelled as such, this was much less common, and it was generally possible to determine this by reading the article and/or finding their IPA document. The lack of clarity about whether articles were in fact RRs has also been noted as a challenge by other authors (Hardwicke & Ioannidis, 2018). Overall, the pervasiveness of this issue suggests the need for much clearer and more reliable labelling and metadata for RRs within journals or publishing platforms. This is essential for accurately representing the article types to readers and for ensuring accurate sampling in future meta-scientific research.

10.3. Implications and Future Directions

10.3.1. Implications and Directions for Journals and Publishers

This study provides encouraging evidence that RRs appear to be performing as expected by being associated with greater rigour and transparency compared to SRs. This evidence could be helpful in informing journal policies and decision-making regarding adoption and implementation of the RR format. This may be particularly useful in informing such decisions within fields where RRs are not yet widely known, although considerable advocacy efforts may be needed to reach journal editors within such areas.

The findings also help to illuminate some areas in which the RR model would benefit from further improvements or greater consistency in how it is implemented at different journals. Specifically, greater public availability of stage 1 protocols would ensure increased transparency for the format. This may require changes to journal policies or processes, including the guidance that they provide to authors. More consistent inclusion of explicit

statements regarding the availability of study materials such as data and code, would also be beneficial (for instance, as required by Level 2 or higher of the Transparency and Openness Promotion guidelines; https://www.cos.io/initiatives/top-guidelines). Even if these materials themselves are not actually available, ensuring clear declarations about (lack of) availability (and accompanying reasons) should be an easy addition to any manuscript. While these practices were observed more frequently in RRs than in SRs, they remained absent from many RRs, indicating room for improvement in increasing transparency. Likewise, the explicit inclusion of some kind of sampling plan or justification would be important and a reasonable expectation, particularly if justifications like those coded as 'other' in the current study are permitted; these included resource limitations, and working with hard-to-reach populations, as well as basing intended sample sizes on those used in previous studies without employing a power calculation or alternative statistical sampling plan.

Clear guidelines for how to prepare a stage 1 (and possibly the subsequent stage 2) RR submission would be helpful at each journal to make the process more streamlined and easier for authors to follow. Meghreblian's proposed interactive study design template (2021) may be a useful tool to support this objective. However, flexibility would also be needed due to the diversity of designs and study topics that may be used in RRs, as too rigid an approach may pose challenges for disciplines and designs that are currently less well represented within RRs. For example, the use of the format for qualitative studies and systematic reviews (though so far primarily within the JMIR journals), suggests the need for researchers, editors and reviewers to be flexible in conceptualising what an RR is and consists of, and the need to be wary of 'one-size-fits-all' mandates for this format. While the core concepts of stage 1 peer review and in-principle acceptance need to be upheld, expecting more diverse study approaches to meet methodological requirements more appropriate for quantitative and experimental designs should be avoided or at least considered with due caution and respect for differing standards across disciplines and methodological approaches. For example, concerns have been raised about the need for journals to expand the use of the RR format for approaches such as secondary data analysis and clinical studies, and the need for specific clear guidance (Kirtley & Lafit, 2021; Tzavella, Meghreblian & O'Mahony, 2021). Likewise, there has been interest in the use of the RR approach for qualitative studies (Chambers, 2016; Karhulahti, 2022), and Karhulahti et al.'s (2023) recent primer provides a useful starting point for designing and evaluating qualitative RRs. Consideration should also be given by journal editors to any potential adaptations or flexibility needed to ensure that the format is

accessible to a broader range of research approaches and disciplines. More formal guidance from journals, and/or training and resources to support researchers when writing their stage 1 submissions, could help to increase the uptake of the format for more diverse approaches. Meanwhile, training and guidance for reviewers and editors may also be helpful to ensure that studies using such approaches are evaluated fairly.

As mentioned in the section above, clear and accurate metadata is required to confirm that particular articles are actually RRs as opposed to just preregistered studies, or even more general unregistered studies. Where possible, efforts to increase the machine-readability of RRs could be helpful in achieving this.

Finally, RRs have also been highlighted as being an important initiative for improving medical research but there has been limited uptake of the format by journals in this area. For example, having the RR format as an option within all medical journals is cited as a core recommendation of the Declaration to Improve Biomedical and Health Research (Graham & Bradley, 2020; Bradley, 2021). Furthermore, since preregistration is already required for clinical trials, making RRs mandatory for clinical trials could be a useful next step to consider, particularly given how prone these studies currently are to bias and QRPs, as outlined in previous chapters. This move has been previously suggested by Chambers and Tzavella (2022), while also highlighting the potential for increased use of RR funding partnership models (Drax et al., 2021; Clark et al., 2021) to increase efficiency. In order to achieve such changes, extensive advocacy and awareness-raising about the format is needed within healthcare disciplines, and existing guidance on preregistration of clinical research could potentially be adapted to help provide clearer instructions for researchers considering the RR format for their clinical trials. Furthermore, concern has been raised about the clinical studies that have been published using the RR format (Anthony et al., 2023), almost all of which have come from the JMIR journal group. Concerns centred around the observation that the dates that in-principle acceptance of stage 1 protocols were awarded were not clear, that evidence of primary outcome switching was common, and that this outcome switching was generally not acknowledged. Therefore, although RRs may have good potential for improving clinical research, editors must ensure that the basic features and standards for this format are upheld consistently.

10.3.2. Database Sharing as a Future Direction

In future, the database will be made publicly available so that it can be used by the wider community. There is potential to make this a more open, interactive, and living public resource for the wider research and open science community, and for this to be supported and maintained by the community in the future. This could then offer a reliable and comprehensive source of published stage 2 RRs, providing an open dataset that may be of interest to others, and providing a source of information and examples for researchers who are considering writing a RR. For example, where links to stage 1 protocols have been made available, these are included in the database and so researchers seeking examples of how to write or structure a stage 1 RR for a particular journal or specific study design could find this more easily. Even after the database has been shared, it could be used as a basis for future research as additional other variables could be added in.

Making the database more interactive and user-friendly would be essential. Tagging of article characteristics in particular should be improved in order to make the database easier to navigate. For example, characteristics such as subject area and study design have been deliberately kept as broad as possible for the current study. However, other researchers using the database as a source of information and examples would likely need more granular details in order to identify papers of most interest to them. Article keywords can be added to facilitate this, along with any other relevant tags. This highlights how the approach that has been necessary in order to code information into broader categories that are helpful for the analysis, may conflict with the needs of users of the database for other purposes in the future. Deeper consideration must therefore be given to how to make the database more user friendly and suitable for the practical needs of the community in the future.

10.3.3 Potential Future Directions for Research

10.3.3.1. Additions to Database

As previously mentioned, the database sample could be used as the basis for future meta-scientific studies, as additional characteristics of interest could be coded and added into the existing resource. This could be done based on the interests of the research community and there is potential for future work on this to be crowdsourced.

10.3.3.2. Replication of HARKing Study

As previously mentioned, the investigation of indicators of HARKing within RRs and SRs which was reported in chapter 8, was a small pilot study and thus should be replicated on a much larger scale to verify the findings.

10.3.3.3. Statcheck Comparison of Statistical Errors in RRs and SRs

Statcheck is a resource available online and as an R package, which detects statistical inconsistencies where results are reported in APA style (Nuijten, Hartgerink, van Assen, Epskamp & Wicherts, 2016). This resource can therefore be used to investigate differences in the rates of statistical errors between RRs and SRs. This was considered as a potential additional project for this thesis but ultimately was not pursued. This has, however, already been examined by de Munck (2019) in a sample of 85 RRs and 85 SRs and they report finding no convincing evidence for RRs having fewer gross statistical inconsistencies. They do however recommend that this should be investigated further in larger samples and thus, further investigation of this in the larger sample used in the current study could be interesting in replicating and expanding on this.

10.3.3.3. Peer Review Studies of RRs and SRs

An additional suggestion for future research, (which had been initially considered as part of this PhD, is the use of blinded peer review by experts to assess the perceived methodological quality of the articles, with the expectation that RRs would be more highly rated than SRs. This is an important consideration due to the importance of peer review in scientific research and so could help to illuminate how RRs are perceived to compare to standard research articles without the knowledge of the study being an RR. A feasibility study was previously conducted by an undergraduate student within the lab group, to determine the logistics of carrying out such a study (Mohamed, 2021). This was considered successful due to the number of reviewers recruited who successfully completed the peer review of their assigned paper. Although there were few significant differences found due to the small sample size, trends were noted in some of the expected directions. Furthermore, although a similar study of RRs was conducted by Soderberg et al. (2021), this feasibility study improved on some important aspects of this such as stricter blinding of participants regarding the article type, and a manipulation check to assess whether participants guessed the article type they were reviewing, which was found to be successful as none guessed this correctly. Larger scale studies could be conducted (possibly as an RR) using this stronger methodology and using

papers from a broader range of disciplines as this feasibility study was limited to a small number of experimental psychology and experimental social psychology articles only.

10.3.3.4. Investigation of Causal Impacts

As noted, experimental designs are needed to establish any causal impacts of the RR format and whether these support the conclusions drawn from this observational work. Randomised controlled designs would be challenging to implement, however, as it would not be realistic to control authors' research or publishing plans. Therefore, an alternative solution that has been suggested is to introduce this at the peer review stage of standard reports, in cases where additional experiments or studies are requested by reviewers (i.e., the 'Registered Revisions' approach: Haber, 2023; Center for Open Science, n.d.). This is a key stage where questionable research practices can be introduced into the process due to the pressure to obtain desirable results in the reviewers' suggested experiments. Conducting the trial at this stage would allow randomisation of the submissions to either receive, or not receive, in-principle acceptance of the subsequent studies regardless of outcome. However, there is potential for sampling bias to be introduced into such a design if authors are expected to opt into the trial since those most open to RRs may be more likely to participate. Having the option for authors to consent to being part of *any* trials that the journal might conduct on their submission process, without providing specific details about the studies being conducted or receiving confirmation of whether they have been included, may help to overcome this bias, but could be considered ethically questionable since consent may not be considered fully informed. Additionally, if the same editors are involved in both arms of the study, their handling of the submissions could be biased by their knowledge of the allocation process. Therefore, a cluster RCT in which some journals provide IPA for the reviewer-requested studies and some journals do not, could help to mitigate this issue.

10.4. Conclusion

This project aimed to establish the feasibility of building a comparative database of the characteristics of a large sample of stage 2 RRs and a yoked sample of standard research articles, to enable comparison between these different article types. Overall, the evidence presented in this work demonstrates that, while there are still areas for improvement, RRs do appear to be working as intended in being associated with improved research practices, although further work is needed to determine causal impacts. There are ample opportunities for future work to support and expand the use of the RR format, such as the sharing of the database as a public resource, future research directions, increased provision of training and

resources for researchers, editors and reviewers, particularly regarding more diverse study designs, and potential areas for refining journal policies and practices.

10.5. References

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

Anthony, N., Tisseaux, A., & Naudet, F. (2023). Published registered reports are rare, limited to one journal group, and inadequate for randomized controlled trials in the clinical field. *Journal of Clinical Epidemiology, 160,* 61-70.

Bradley, S. (2021). The Declaration to Improve Health Research two years on. *HealthSense UK Newsletter,* 116, 5-7. https://www.healthsense-uk.org/publications/newsletter/newsletter-116/236-116-bradley.html

Center for Open Science. (n.d.). *Impact of Registered Revisions Within the Journal Peer Review Process.* https://www.cos.io/r3ct/registered-revisions

Chambers, C. D. (2016, September 19). Registered Reports for qualitative research: a call for feedback from humanities and social science researchers. *NeuroChambers*. http://neurochambers.blogspot.com/2016/08/registered-reports-for-qualitative.html

Chambers, C. (2019). What's next for Registered Reports? *Nature, 573*, 187-189. https://doi-org.abc.cardiff.ac.uk/10.1038/d41586-019-02674-6

Chambers, C. D., & Mellor, D. T. (2018). Protocol transparency is vital for registered reports. *Nature Human Behaviour, 2*(11), 791-792.

Chambers, C.D., & Tzavella, L. (2022). The past, present and future of Registered Reports. *Nature Human Behaviour, 6*, 29-42.

Clark, R., Drax, K., Chambers, C. D., Munafo, M., & Thompson, J. (2021). Evaluating Registered Reports Funding Partnerships: a feasibility study. *Wellcome Open Research, 6,* 231. https://doi.org/10.12688/wellcomeopenres.17028.1

De Munck, S. (2019). *Registered Reports and statistical reporting inconsistencies in psychology*. [Master thesis, Tilburg University]. Tilburg University Library. https://arno.uvt.nl/show.cgi?fid=148954

Drax, K., Clark, R, Chambers, C. D., Munafo, M., & Thompson, J. (2021). A qualitative analysis of stakeholder experiences with Registered Reports Funding Partnerships. *Wellcome Open Research, 6,* 230. https://doi.org/10.12688/wellcomeopenres.17029.1

Duyx, B., Urlings, M. J. E., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. *Journal of Clinical Epidemiology, 88,* 92-101. https://doi.org/10.1016/j.jclinepi.2017.06.002

Graham, C. J. & Bradley, S. H. (2020). Declaration to improve biomedical and health research. *The Journal of the Royal College of Physicians of Edinburgh, 50*(3), 343-350.

Haber, N. [@NoahHaber]. (2023, June 6). *Announcing: Impact of Registered Revisions Within the Journal Peer Review Process at COS (@OSFramework)! Interested in: * Methods commitment devices?* [Tweet]. Twitter. https://twitter.com/NoahHaber/status/1666099562606624772

Hardwicke, T.E., & Ioannidis, J.P.A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour*, 2, 793-796.

Hummer, L., Thorn, F. S., Nosek, B. A., & Errington, T. (2017). *Evaluating Registered Reports: A naturalistic comparative study of article impact*. OSF Preprints. https://doi.org/10.31219/osf.io/5y8w7

Jannot, A. S., Agoritsas, T., Gayet-Ageron, A., & Perneger, T. V. (2013). Citation bias favoring statistically significant studies was present in medical research. *Journal of Clinical Epidemiology, 66(*3), 296-301. https://doi.org/10.1016/j.jclinepi.2012.09.015

Karhulahti, V.-M. (2022). Registered reports for qualitative research. *Nature Human Behaviour, 6,* 4-5.

Karhulahti V.-M., Branney, P., Siutila, M., & Syed, M. (2023, January 31). A primer for choosing, designing and evaluating registered reports for qualitative methods. *Open Research Europe, 3*, 22. https://doi.org/10.12688/openreseurope.15532.1

Kirtley, O. & Lafit, G. (2021). *Accelerating the adoption of Registered Reports in clinical psychology and psychiatry.* Unconference session at Society for the Improvement of Psychological Science conference 2021, Virtual., 23rd June 2021.

Meghreblian, B. (2021). *Encouraging Registered Reports – Metascience & Tool Development*. Lightning talk at Metascience 2021 conference, Virtual, 16th September 2021.

Misemer, B. S., Platts-Mills, T. F., & Jones, C. W. (2016). Citation bias favoring positive clinical trials of thrombolytics for acute ischemic stroke: A cross-sectional analysis. *Trials, 17,* Article 473. https://doi.org/10.1186/s13063-016-1595-7

Mohamed, I. (2021). *A Peer Review Investigation: Registered Reports Compared to the Traditional Publishing Model*. [Undergraduate thesis, Cardiff University]. Unpublished.

Norsa, R. (2022) *Do Registered Reports Improve Sample Size Planning in Psychology? An exploratory study*. [Thesis, Università degli studi di Padova]. Thesis and Dissertation Padua Archive. https://thesis.unipd.it/handle/20.500.12608/32381

Nuijten, M. B., Hartgerink, C. H. J., van Assen, M. A. L. M., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods, 48*(4), 1205-1226.

Obels, P., Lakens, D., Coles, N., Gottfried, J. & Green, S. (2020). Analysis of open data and computational reproducibility in registered reports in psychology. *Advances in Methods and Practices in Psychological Science, 3*(2), 229-237.

Scheel, A.M., Schijen, M.R.M.J., & Lakens, D. (2021). An excess of positive results: comparing the standard psychology literature with registered reports. *Advances in Methods and Practices in Psychological Science, 4*(2), 1-12.

Soderberg, C. K., Errington, T. M., Schiavone, S. R., Bottesini, J. G., Singleton Thorn, F., Vazire, S., Esterling,m K. M., & Nosek, B. A. (2021). Initial evidence of research quality of registered reports compared with the standard publishing model. *Nature Human Behaviour, 5,* 990-997.

Tzavella, L., Meghreblian, B. & O'Mahony, A. (2021). *How can we improve Registered Reports for authors, reviewers, and editors?* Unconference session at Society for the Improvement of Psychological Science conference 2021, Virtual., 23rd June 2021.

Appendix 1: Standard Report (SR) Selection Protocol

Protocol last updated:

02/10/2020: Details added about completing the SR justification table.

01/07/2021: General proofreading and occasional clarification of phrasing. No content changes.

In order to compare the characteristics of Registered Reports (RRs) with standard research articles these standard articles must be chosen carefully, and matched on relevant characteristics where possible, using a structured and consistent approach. Two of these standard reports (SRs) are selected for each RR. The decisions are documented in a structured way in the SR justification table. Also, in the last columns of the SR coding sheet, you will need to code for the relevance of the articles to show how closely they match the RR, both for each important characteristic, and as an overall rating – see the final section of the coding protocol document for details of how to do this (Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant).

General Matching Process

Foremost, SRs should come from the same journal and be published around the same time (same issue or one issue either side, if possible) as the RR. Within these limits, attempt to find the most relevant standard research papers in terms of the topic, research design, population, and sample size. Often, you may need to balance these considerations and choose between these characteristics because an SR with the same study design as the RR may not be the most relevant paper available in that issue in terms of the topic or the population.

Although there are six characteristics that we match the SRs on (journal, timeframe, topic, design, population, and sample size), there is a somewhat hierarchical approach to these characteristics. Choosing from within the same journal is essential (except for Comprehensive Results in Social Psychology). Then, it is important to match as closely as possible within the same timeframe. Aim first to match SRs within the same issue of the journal, and if the options are not satisfactorily relevant, widen your search to within one issue either side.

Within the appropriate timeframe/range of journal issues, matching on topic is then typically most important, and similarity of study design should also be considered as much as possible. Considering the similarity of the population is often a little less important, depending on the type of papers being matched -population similarity can be more important in certain topic areas e.g., people with similar conditions for clinical psychology or health research, whereas within experimental psychology they may often be student samples or MTurk participants, in which case populations may be broadly similar anyway. Sample size is generally the least important characteristic to consider when matching SRs; although it is not unimportant, it should not have too strong an impact on your judgement of a paper that otherwise matches well on other characteristics. Despite the slightly different degrees of importance of these

characteristics during the actual process of matching them, when coding the relevance of the characteristics in the SR database we give equal weight to each of these characteristics.

It is important to choose from the same article type wherever possible. For example, some journals publish 'short research articles' or 'brief reports' as well as standard length research articles. As these will be shorter than a standard-length article, they will generally have less detail and the methods section may have a lower word count. Therefore, try to avoid selecting these unless there is really no other option available and they are not excessively short (e.g. 2 or 3 pages would be too short to be a fair comparison), and they are highly relevant on the relevance criteria mentioned below. If this seems necessary, try instead to search one more issue either side for possible matches even if these papers are not very relevant in terms of the other characteristics, because we really want to avoid selecting these shorter article types if at all possible.

Additionally, do not choose systematic/literature review papers as SRs for articles containing primary research studies, as these are too different to be comparable on most characteristics.

When downloading and saving SRs you have selected, ensure the file name reflects which RR the SR is matched to – see the names given to the pdfs of the practice SRs in the practice coding folder.

When selecting SRs it is important to use the SR justification table to document these choices as well as the characteristics that were used to match them on.

SR justification table

It is important to document the process/justification of choosing SRs for each RR in order to transparently document how rigorously this has been done. The SR justification table has several columns: the first states the matching characteristics that need to be documented, then there is a column to complete this for the RR, and then two columns to complete this for the two SRs selected. There is then a further column to document these characteristics if another article was closely considered or 'shortlisted' for being chosen as a SR but was ultimately not chosen (if applicable). There will not always be a third potential option for this but when there is it should be documented because this helps to show the decision-making process in a transparent way.

Each SR article should only be chosen once (i.e., a SR chosen for one RR can't also be an SR for a different RR. However, if a third article was 'shortlisted' as a potential choice (and included in the SR justification table in the 'Other column) but was not ultimately chosen, this paper can still be considered or chosen as a SR for a different RR.

When listing/describing each characteristic of the SR papers, please try to colour-code the potential relevance of these as follows: Highlight in green if you are likely to code that characteristic/aspect of that characteristic as 'Very relevant'. Without using any highlighting, use dark green text colour if you are likely to code it as 'Somewhat relevant'. Without using any highlighting, use red text colour if you are likely to code it as 'Somewhat Irrelevant'.

Highlight in red if you are likely to code it as 'Very Irrelevant'. See the example SR justification table for a guideline of how this should look.

Sometimes there might be a mixture of these likely rankings for a particular characteristic of a particular paper, e.g. one aspect of the paper's topic might be very relevant but another aspect of its topic might be somewhat irrelevant – you will need to balance these considerations in making a final judgement about how to code that characteristic – Very relevant might still be appropriate, or you may need to compromise and code it as 'Somewhat relevant', depending on how important that irrelevant aspect of the topic is.

Journal

To begin selecting SRs, check what journal, and what volume and issue of that journal the RR was published in. The most important characteristic to match on is the journal itself. For all journals except Comprehensive Results in Social Psychology, SRs should be selected from the same journal as the RR. As Comprehensive Results in Social Psychology only publishes RRs, Journal of Social Psychology has been chosen as a baseline journal for these RRs instead. This is because Journal of Social Psychology is similar in terms of scope and subject area.

As long as the SR comes from the same journal as the RR, it can be coded as 'Very relevant' for this characteristic (column CO of the SR coding sheet).

For articles taken from Journal of Social Psychology as SRs for the RRs from Comprehensive Results in Social Psychology, these can be coded as 'Somewhat relevant' for the journal relevance characteristic (column CO).

Timeframe

It is important that SRs are published around the same time as the RR, as different journal policies could be introduced over time that influence the quality/characteristics of the articles published there. If possible, try to select SRs from the same issue of the journal as the RR was published in (although this will depend somewhat on whether any of the articles are relevant to the topic of the RR – see next characteristic). If the SRs are selected from the same issue as the RR, this characteristic (column CP) can be coded as 'Very relevant'. If there are not reasonably topic-relevant articles in the same issue, try searching within one issue either side of the issue the RR was published in – an article chosen from one issue either side can be coded as 'Somewhat relevant' for the timeframe (column CP). The volume and issue number should be included in column CN of the coding sheet to help clarify this.

When RRs have been published as a special issue of the journal, there are typically only RRs within that issue and no standard research articles. In that case, choose SRs from within the issue either side of that special issue, and continue to code the timeframe as 'Somewhat relevant'.

Try to always choose SRs from within these time limits if possible, even if you have to choose papers that are not very relevant in terms of topic, design, etc, However, if there are not enough standard articles available within this range of issues, you can broaden out your

search to one more issue either side, and code the timeframe characteristic of these SRs as 'Somewhat Irrelevant', and so on. This situation may be particularly likely when matching SRs for RRs that have come from special issues, because if no SR papers are available from the same issue, there may not be enough articles in the issue either side to have two different SRs for each RR. Likewise, this could also occur if journals don't publish many articles per issue.

Topic

While an exact topic match may not be very likely, if it is the same general topic area, code as 'Very relevant', e.g. if both papers are about app-based interventions for depression or low mood, or both about online therapy for young people's mental health, or both about prejudice and in-group/out-group biases. If the SR's topic is not close enough to be coded as 'Very relevant' but is still not too dissimilar, code as 'Somewhat relevant', e.g. if they are both about some form of cognitive processing, or if there are some minor aspects of the topic in common.

Within many journals (within the timeframe limit of same issue or one issue either side), it is likely that there often will not be a clear match in terms of the topic but you should still try to choose the most relevant (or least irrelevant) paper possible within these limits. Try to balance this with the relevance of the study design (next characteristic).

When documenting this characteristic in the SR justification table, usually you would need to use the main important terms from the title and any keywords given (except those that relate to the methodology). You can also add any other terms that you feel are important for showing the relevance of these papers (e.g. if the title, keywords etc. all relate to a specific concept or theory but the main similarity between the papers is that they are all about different types of cognitive processing, you can add 'cognitive processing' as a term in this section. Likewise, all of the papers might relate to different forms of social judgements, or different types of decision-making processes.

Design

Where possible, try to match SRs on similar research designs. If the designs are very similar code as 'Very relevant', e.g. if they are both an RCT, or both a systematic review. If an SR contains an online experiment while the RR has an in-person experiment, this should be coded as 'Somewhat relevant'.

You may need to balance choosing the most relevant design, with choosing papers focusing on the most relevant topic.

When describing the study designs in the SR justification table, use broad categories: Observational / Systematic review / Experiment / Intervention (although you can specify if it is a pilot intervention study). You can add more detail to this part if it helps to show the relevance of the papers to each other (e.g. if both the RR and SR design used online surveys, you can also include this as an extra detail). Make sure to specify whether studies (such as

Experiments) were done online or in-person as this is an important distinction that influences how you should code the relevance of the paper for this characteristic.

Often keywords (i.e. those given in the article keywords) that relate to the methodology are too specific to the particular design to be very useful when comparing the relevance so you do not have to include these unless they seem relevant and/or broad enough to be helpful.

Population

While an exact population match may not be very likely, if it is the same general population group, code as 'Very relevant', e.g. older adults with cancer, or teenagers with depression, or neurotypical undergraduate students. If the population is not too dissimilar but not close enough to be considered very similar, code as 'Somewhat relevant', e.g. if they are an adult population from very different cultures (e.g. China vs. USA), or adult caregivers of relatives with dementia vs. parents/caregivers of children with disabilities, or adult employees in a certain organisation vs. university students (which are still technically an adult population but this distinction should be made clear since university students are a specific population group).

When describing the population in the SR justification table, you do not need to go into much detail. Just a couple of keywords is generally sufficient (e.g. Adolescents with depression / University students / Adults caregivers of people with dementia).

Sample Size

Where possible, consider the general relevance of the sample size, although the other characteristics are more important to match on where possible.

There is unlikely to be a close match on this characteristic and the other characteristics should be prioritised instead when selecting SRs, but we can at least code for whether the sample size is completely dissimilar or not, e.g. if the RR has a sample of 2,000 people and the SR has a sample of 15 people, this would be 'Very Irrelevant', whereas if the SR had a sample of 1,750, this might reasonably be coded as 'Very relevant'. There will be a certain degree of subjectivity in the coding of this characteristic.

When describing this information in the SR justification table, if there are multiple studies in the paper, give the sample size per study (e.g. 36 + 45 + 67).

Overall SR relevance

Based on the coding of the relevance of the SR characteristics, you will need to make an overall judgement on how relevant the SR is. This process is described below and is also available in the coding protocol, which should be used in conjunction with this current document.

Based on the coding of the relevance of the SR characteristics in the previous 6 columns, select an overall ranking for how relevant the SR is for the RR it is matched with. As a guide, use the following ranking system:

For each of the 6 relevance characteristics, they are coded on a 4-point scale (Very Irrelevant to Very relevant). If we assign a number to each of these, we can mark each characteristic out of 3:

Very Irrelevant = 0

Somewhat Irrelevant = 1

Somewhat relevant = 2

Very relevant = 3

If we mark each of these 6 characteristics out of 3 like this and them add them all together, we get a total number out of a maximum possible total of 18. In order to determine an overall relevance ranking for the BA, use the following (approximate) scale:

0-4 = Very Irrelevant

5-9 = Somewhat Irrelevant

10-14 = Somewhat relevant

15-18 = Very relevant

While this scale is not totally even, it gives an approximate system for assigning an overall category for the BA's relevance, based on the relevance of the characteristics it was matched on.

Notes

There is a specific notes column in the SR coding sheet which allows you to explain any major issues with the matching process. There is also a column for notes/comments in the SR justification table.

Appendix 2: Registered Reports & Standard Reports Initial Coding Protocol

LAST UPDATED

01/06/2021: Addition of specific competing hypothesis response options, and explicit addition of the updated approach to coding pilot studies. General proofreading and clarifying some descriptions.

18/06/2021: General proofreading and clarifying some descriptions.

Introduction

The coding sheet is used to store and analyse information we are interested in, about the characteristics of Registered Reports (RRs) and standard reports (SRs).

There are two separate sheets, one for RRs and one for SRs. While the characteristics being coded are almost identical between the RRs and the SRs, there are a couple of important differences and some additional characteristics to code for the SRs.

Specifically, we look at the following areas:

1) General article information (columns A to Y)
    - ID numbers (RR-ID and RR-SR-ID)
    - Authors
    - Year
    - Article title
    - Journal name
    - Journal subject area
    - Country of authors
    - Term used
    - Protocol availability
    - Link to protocol
    - No. of authors
    - Author name and seniority
    - Data availability statement
    - Level of availability of data

- Analysis code used
- Analysis code availability statement
- Level of availability of analysis code
- Digital materials used
- Digital materials availability statement
- Level of availability of digital materials
- Other (non-digital) original materials used
- Other original materials availability statement
- Level of availability of other original materials

2) Article level characteristics (columns Z to AR)
   - Hypothesis stated at ARTICLE level
   - Articulation of hypothesis at ARTICLE level
   - Support for hypothesis at ARTICLE level
   - Articulation of support for hypothesis at ARTICLE level
   - Intended/planned sample size at ARTICLE level
   - Pre-exclusion sample size at ARTICLE level
   - Post-exclusion sample size at ARTICLE level
   - Sampling plan at ARTICLE level
   - Sampling unit at ARTICLE level
   - Word count of methods section at ARTICLE level
   - Exploratory/additional analysis included at ARTICLE level
   - Clear distinction between confirmatory and exploratory analysis at ARTICLE level
   - Nature of exploratory analysis at ARTICLE level
   - Manipulation check used at ARTICLE level (x3)
   - Passed manipulation check at ARTICLE level (x3)
3) Study level characteristics (columns AS to BP)
   - No. of studies
   - Type of study
   - Study design
   - Registered
   - Hypothesis stated at STUDY level
   - Articulation of hypothesis at STUDY level
   - Support for hypothesis at STUDY level

- Articulation of support for hypothesis at STUDY level
- Intended/planned sample size at STUDY level
- Pre-exclusion sample size at STUDY level
- Post-exclusion sample size at STUDY level
- Sampling plan at STUDY level
- Sampling unit at STUDY level
- Replication
- Word count of methods section at STUDY level
- Exploratory/additional analysis included at STUDY level
- Clear distinction between confirmatory and exploratory analysis at STUDY level
- Nature of exploratory analysis at STUDY level
- Manipulation check used at STUDY level (x3)
- Passed manipulation check at STUDY level (x3)

4) Hypothesis level characteristics (columns BQ to CJ)
    - Hypothesis stated at HYPOTHESIS level
    - Articulation of hypothesis at HYPOTHESIS level
    - Support for hypothesis at HYPOTHESIS level
    - Articulation of support for hypothesis at HYPOTHESIS level
    - Intended/planned sample size at HYPOTHESIS level
    - Pre-exclusion sample size at HYPOTHESIS level
    - Post-exclusion sample size at HYPOTHESIS level
    - Sampling plan at HYPOTHESIS level
    - Sampling unit at HYPOTHESIS level
    - Exploratory/additional analysis included at HYPOTHESIS level
    - Clear distinction between confirmatory and exploratory analysis at HYPOTHESIS level
    - Nature of exploratory analysis at HYPOTHESIS level
    - Manipulation check used at HYPOTHESIS level (x3)
    - Passed manipulation check at HYPOTHESIS level (x3)
5) Notes (column CK and CL)
    - Coding Notes
    - Level of difficulty to code
    - Reason for coding difficulty

6) Relevance of SRs to the RRs (SR coding sheet only, columns CN to CV)
    - Journal volume and issue number
    - Journal relevance
    - Timeframe relevance
    - Topic relevance
    - Design relevance
    - Population relevance
    - Sample size relevance
    - Overall relevance
    - Notes on SR relevance coding
7) Additional columns (columns CN to CR for RRs, columns CW to DC for SRs)
    - Non-final copy coded
    - Coded by
    - Coding double-checked by
    - Coding triple-checked by
    - SR selected by (SR coding sheet only)
    - SR selection checked by (SR coding sheet only)

This step-by-step protocol is designed to enable a consistent approach to coding these characteristics. However, many articles will have unique issues or subjectivity that need to be addressed. Try to code the information as best you can, which may require making somewhat subjective judgements on what response option is appropriate. If you run into problems with this while learning to use this protocol, please highlight the relevant piece of information in the coding sheet, describe any issue in the 'Notes' column (marking the information with an asterisk if doing this - see next paragraph). If there is a relatively minor issue that you need to check, change the text colour to red. If there is a major issue, you are very uncertain of your coding judgement, or you cannot code that piece of information at all, highlight that space in yellow to ensure it is noticeable and describe the problem in the 'Notes' column. This will be helpful when comparing or checking the work of different coders later on in order to investigate any difficulties or differences in the coding, as any explanation of major difficulties encountered in coding that particular variable may help in the process of reaching a consensus.

If you wish to make notes in the 'Notes' columns to clarify something about the coding of a particular variable, ensure that you use an asterisk both in the characteristic column, and then in the Notes section where you add any extra details about it. As this may be necessary for multiple issues within the same article, increase the number of asterisks for each issue within that article e.g. the first time use one (*), the second time use two (**), etc. (See example coding in Excel sheet)

The following sections describe the general coding protocol as it applies to the RRs and SRs. This approach is mostly the same for each coding sheet, but there are a few key differences between the RRs and SRs to look out for, such as additional columns at the end of the SR coding sheet to code for the relevance of the SR to the RR it was matched with.

Some characteristics have specific response options to choose from while others have more open responses. When coding from a list of response options for a particular characteristic, ensure that you format the response options as given in the protocol and the column heading, e.g. in relation to use of uppercase and lowercase letters, and placement of hyphens or underscores.

1. General article information

The initial columns contain general identifying information about the article, most of which is more for administrative purposes.

Column A: "RR-ID", or "SR-ID"

This is the article's ID number within the coding sheet, i.e. if it is the first/second/third RR or SR coded, etc. In the RR coding sheet this heading will be "RR-ID", while in the SR coding sheet it will be "SR-ID".

Example: RR_1 , RR_79 , SR_6 , SR_35.

Column B: "RR-SR-ID", or "Filler column"

In the RR coding sheet, leave this column blank – it has been added in to make the coding sheets even in terms of the number of columns.

In the SR coding sheet, this column should contain a number linking each SR with the RR it is matched to.

For example, for RR 79, there are two SRs so the first SR for that RR would be 79_1. The second SR chosen for that RR would be 79_2.

Column C: "Authors"

Surname of authors. If there are more than two authors, first author followed by et al. is used. Examples: Huff & Kruszewska , or Lynott et al.

Column D: "Year"

Year of publication, e.g. 2016

Column E: "Article title"

Title of article, e.g. Effects of subtitles, complexity, and language proficiency on learning from online education videos

Column F: "Journal name"

Name of the journal the article was published in e.g. Journal of Media Psychology

Notes/Issues in coding:

Give the full name of the journal, not the abbreviated form of the name

Column G: "Subject area"

General subject area of the journal e.g. Neuroscience

Notes/Issues in coding:

This is the subject area of the journal, not of the article.

If this is not immediately clear from the journal name, search for the journal online to read what they describe as their scope/discipline. When the journal focuses on a specific subdiscipline, code it as this, e.g. as Health Psychology or Social Psychology rather than just Psychology.

Some journals may need to be coded as 'Interdisciplinary' if necessary.

Column H: "Country"

Country of authors, e.g. Germany

Notes/Issues in coding:

Get this information from the author by-line (institutional affiliation, etc.). In cases where there are multiple countries (e.g. an international collaboration), include all of the countries.

Country should be included for all authors' institutions not just the maximum of four authors included in columns M and N.

Column I: "Term used"

Term used to refer to the article format.

Examples: For registered reports this could be: 'Registered report', 'Pre-registered direct replication' (as used in Psych Science), or 'Replication' (as used in eLife, where it is the stage 1 report that is referred to as the RR while the stage 2 report is referred to as a replication).

For SRs, this might just be something like 'Original Article', 'Research Article' or 'Review'. If there is no term given, you can use N/A.

Notes:

RRs are generally labelled as Registered Report. RRs are called preregistered direct replications in Psychological Science, and stage 2 RRs are called replications in eLife as they refer to the stage 1s as RRs instead.

For JMIR papers, if they have an International Registered Report ID number (IRRID) this marks them as an RR. Even though JMIR refer to these papers as the article type, e.g. 'Original article' (or sometimes as a 'Review' if it's a systematic review or similar), if they have an International Registered Reported ID number (IRRID), code the term used as Registered Report instead. If the JMIR paper doesn't have an IRRID but was published before May 2018, check if there is a protocol in JMIR Research Protocols – this could still be an RR.

Column J: "Protocol availability"

Check if the paper contains a link to the protocol, or if the protocol is otherwise identifiable e.g. by an ID number linking the protocol with a stage 2 article as per JMIR RRs.

Response options: Yes / No

Notes:

Search the document for relevant terms, e.g. protocol, proposal, preregistration/pre-registration.

Also search for any OSF links, as they sometimes contain the protocol along with data etc, even if they don't state this directly in the paper.

For JMIR RRs, the protocol should be in JMIR Research Protocols. Search the JMIR website for article details, author names, etc., and filter by JMIR Research Protocols. The first author on JMIR stage 2s is not always the same first author as on the protocols, so you may need to search several of the authors to find it. Try searching the IRRID on the JMIR website if necessary. If the protocol can't be found on JMIR site this way, google article titles and

'protocol', to see if you can find the protocol in a different journal. If this is not successful, then Google the IRRID number.

If they state that the protocol is available but it is not actually accessible (e.g. the link provided doesn't work), try to find it in any other way and if unsuccessful, code as No for availability and then code as N/A for the link in the next column, but write a note in the Notes column explaining that the link provided doesn't work, etc.

Column K: "Link to protocol"

Include link if the protocol/preregistration document exists. Otherwise, use N/A as a response.

Column L: "No. of authors"

Number of authors listed, e.g. 4

Column M: "Author name"

Full name (e.g. Scott J Peters) is given for the first three authors, and the last author.

Notes:

Inclusion of this information here is really to give context to the information coded in the next column

Column N: "Seniority of authors"

Seniority of each author, at time of publication, not at time the research was done. This is currently only being coded for the first three authors and the last author.

Examples: Postdoctoral researcher / Research Assistant / Senior Lecturer

Notes:

Currently, just write the relevant job title in. This will be changed to categories at a later date for easier analysis.

Search online for the author, particularly in combination with the name of the institution given in the article's author information. Search particularly for profile pages on institution websites, LinkedIn profiles, and CVs. You may need to use the term 'CV' in the search.

Authors may have moved on from that institution or onto a different role since work on that project was done, especially for older papers. For coding purposes, only focus on their role at the time of publication, not their role when the work was done, and not their current role, if these roles are different.

If you can't find authors' information for that time period but can find info for their current role, this can be included in the database in RED font with the word 'currently', (e.g. Currently – Assistant Professor) and specify in the notes column that this is the person's current role but their role title in the year of publication is unclear. This is potentially useful for more recent papers, as authors' roles may not have changed, but this still needs to be verified in the future.

If no information can be found, code as 'unclear' and highlight in yellow. Contact can be attempted at a later date to clarify current roles if needed.

Column O: "Data availability statement"

Whether authors explicitly state whether the data is available (not whether the data is actually available)

Response options: Yes / No

Notes:

Check for any statement in the paper about data being available, or for links to repositories such as OSF, as they often mention data availability when giving the OSF link.

Occasionally a paper may state that data is available, particularly as part of Supplementary Materials or an online appendix, but when examining the files it's not raw data but additional/supplementary data or even just tables or figures from supplemental analyses rather than that used in the paper - if the statement refers to the data being available but doesn't specify that it is only supplementary and not the actual data that is central to the study, the statement can be coded as Yes (with an asterisk and an explanation in the Notes column), but the following column about the level of availability should not be coded as being available: Unclear would likely be the most suitable response option in such cases, and the asterisk should be used here as well to link it to the same explanation in the Notes column.

A similar approach should be taken if a paper states that the data are available at a particular link but are found to not actually be there when the contents of this link are examined (Code as Yes for statement and Unclear for availability of data).

For a systematic review/scoping review, etc. a completed data extraction table could be considered as the study's data so if this is available it can be coded as Yes

Column P: "Availability of data"

Level of availability of data

Response options: Publicly available / Gated / Available on request / Already available / Not available, without justification / Not available, with justification / Unclear

Notes:

- Publicly available: data fully available and accessible to the public without making accounts or other processes to obtain the data, e.g. available on OSF (check links provided in paper to ensure data is there). If a link is given to the data but that link doesn't work, and you can't find the data in any other way, code as Unclear and explain this in the notes column.
- Gated: technically the data have been made available but you would need to make an account to access it, or other similar process required to obtain it from a public source. It may be behind a paywall. If an article states that their data etc. are available as an appendix in the online version of the article, check that online version of the article to see if this data is accessible and if the article is behind a paywall for the public (i.e. when not accessed through your university library account but just from searching through Google, etc.), this could be considered gated access. If this occurs, please make a note in the Notes column explaining this.
- Available on request: any mention of data being available from the author, or about needing to contact the author to obtain it.
- Already available: In some instances, authors may have obtained the data from a source where it is already publicly available e.g. census data. In this case it is publicly available but not provided specifically by the authors. (E.g. some AERA Open RR papers).
- Not available WITHOUT justification: Only use this if the article explicitly states that the data are not available, without a reasonable justification for why not (if they have a reason, code as 'Not available, WITH justification' instead).
- Not available, WITH justification: Any other reason given for why data is not available e.g. not available due to confidentiality issues or if funders don't permit data sharing, etc.
- Unclear: If paper makes no mention of whether data is available, or the availability is otherwise unclear and doesn't fit any of the other codes. If a link is given to the data but that link doesn't work, and you can't find the data in any other way, code as Unclear and explain this in the Notes column.

Column Q: "Analysis code used"

Whether analysis code was used in the study – coding this is included in order to inform the variables in the subsequent columns (regarding availability of analysis code).

Response options: Yes / No / Unclear

Notes:

- Yes: Specific mention of code, syntax, scripts, or software that uses code, such as R or Python (search for these terms within the document and read the analysis section carefully for relevant terms). Also check any OSF links or other repositories in case they have provided the analysis code there – this is not always mentioned explicitly in the paper even if it is available in a repository.
- No: Explicit mentions of software that doesn't rely primarily on syntax such as SPSS, or other can be a useful indication of whether code was used for the analysis although this is not always definitive. For example, syntax from SPSS analyses can be extracted and shared, so do check for this, especially on OSF links, etc.
- Unclear: If authors do not mention code being used or available, or if they don't specify the analysis software used, etc.

Column R: "Analysis code availability statement"

Whether authors explicitly state whether analysis code is available (if code was used)

Response options: Yes / No/ N/A

Notes:

- Yes: Specific mention of analysis code/syntax/scripts being available.
- No: Any lack of statement about the availability of analysis code (even if code is actually available in a repository, if it is not explicitly stated in the paper, code as 'No')
- N/A: Non-applicable if code was not used for the analysis (i.e. If previous column was coded as 'no')

Column S: "Availability of analysis code"

Level of availability of analysis code (if used)

Response options: Publicly available / Gated / Available on request / Not available, without justification / Not available, with justification / Unclear / N/A

Notes:

- Publicly available: code fully available and accessible to the public without making accounts or other processes to obtain the data E.g. available on OSF (check links provided in paper to ensure code is there). If link is given to the code/repository but link doesn't work, highlight this and mention it in the notes section.
- Gated: technically available but may need to make an account to access it, or other similar process required to obtain it from a public source. May be behind a paywall.
- Available on request: any mention of the code being available from the author.
- Not available, WITHOUT justification: Only use this if it explicitly states that the code is not available, without a reasonable justification for why not (if they have a reason, code as 'Not available, WITH justification' instead).
- Not available, WITH justification: Any other reason given for why code is not available.
- Unclear: If paper makes no mention of whether code is available, or the availability is otherwise unclear and doesn't fit any of the other codes.
- N/A: Non-applicable, i.e. if column Q was coded as 'No'.

If the variable for whether code was used (Column Q) has been coded as unclear, the availability should also be coded as unclear, as the other response options would not be suitable. In this situation, the variable for whether there is a statement about the availability of code (Column R) would be coded as 'No'.

Column T: "Digital materials used"

Whether digital materials were used in the study (e.g. online surveys, stimuli, or code for running experiments), to inform other variables regarding availability of digital materials

Response options: Yes / No / Unclear

Notes:

Digital materials may include the code for running experimental tasks administered on a computer or online, stimuli or videos used in experimental tasks, online surveys, or any

online/digital intervention used in the studies such as online CBT interventions, apps, or e-learning materials. Recordings and subsequent transcripts of interviews could be considered digital. If coding a systematic review or similar paper, availability of a document with the search terms used, list of inclusion/exclusion criteria (in addition to what is described in paper about these criteria) or a list of the excluded full-text papers with reasons for exclusion, etc., could all be considered digital materials – if they only provide one of these documents, they can be coded as 'Yes' in this column, and coded as 'Partially available' in column V.

It was initially difficult to determine if some materials count as digital or not e.g. a list of questions from an online survey – survey was administered online so would be digital but a pdf with a list of questions that were used in that survey could be construed as non-digital. Following discussion about this early on during the development of the protocol, such materials should be coded as 'Yes' here, and then coded as 'Partially available' in column X regarding the availability of the digital materials, due to how they were originally administered making them digital materials. See information for column V for further info on how to recognise Partially available materials

- Yes: Clear mention of digital materials being used e.g. within 'materials' or 'methods' section of paper.
- No: Clearly evident that no digital materials were used e.g. if study only used in-person tasks and paper-based questionnaires.
- Unclear: If unsure whether materials were delivered digitally

Column U: "Digital materials availability statement"

Whether authors explicitly state whether digital materials are available (if used)

Response options: Yes / No / N/A

Notes:

- Yes: Specific mention of whether digital materials are available. Also if they just mention 'materials' generally being available but digital materials were used in the study, likely also coded as yes here – check the available materials if they are linked to ensure they are the relevant ones.
- No: Any lack of statement about the availability of digital materials, even if materials are actually available in a repository, etc.

- N/A: Non-applicable if digital materials were not used (i.e. if previous column T was coded as ‘No’).

Column V: “Availability of digital materials”

Level of availability of digital materials (if used – see info for column T for examples/clarification of what constitutes digital materials).

Response options: Publicly available / Partially available/ Gated / Available on request / Not available, without justification / Not available, with justification / Unclear / N/A

Notes:

- Publicly available: materials fully available and accessible to the public without making accounts or other processes to obtain the data E.g. available on OSF (check links provided in paper to ensure materials there). If link is given to the materials/repository but link doesn’t work, code as Unclear and explain this in the notes section (using asterisk as needed)
- Partially available: this applies if only some of the digital materials are publicly available, e.g. if it includes a list of the questions used in an online survey but these are only available as a pdf, or if they provide the code for running a task but not the other digital materials that were part of it such as images or video clips – these would all be coded as being partially available.
- Gated: technically available but would need to make an account to access it, or other similar process required to obtain it from a public source. May be behind a paywall.
- Available on request: any mention of contacting the author for the materials.
- Not available, WITHOUT justification: Only use this if it explicitly states that the materials are not available, without a reasonable justification for why not (if they have a reason, code as ‘Not available, with justification’ instead)
- Not available, WITH justification: Any other reason given for why materials are not available, e.g. confidentiality issues, copyright issues, etc.
- Unclear: If paper makes no mention of whether materials are available, or the availability is otherwise unclear and doesn’t fit any of the other codes
- N/A: Non-applicable if digital materials were not used (i.e. if column T was coded as ‘No’).

Column W: "Other Original Materials used"

Whether any other original materials (non-digital) were used in the study (e.g. paper-based questionnaire, paper-based focus group or interview question list), to inform subsequent variables regarding availability of these other materials

Response options: Yes / No / Unclear

Notes:

This is rarely applicable but was added in as a characteristic because these types of materials have been used in a small number of studies. To be relevant, these must be original non-digital materials, i.e. developed by the authors/for the study, not just using established questionnaires/materials developed by others.

- Yes: Clear mention of these materials being used e.g. within 'materials' or 'methods/ section of paper.
- No: Clearly evident that no non-digital original materials were used, e.g. if no non-digital materials were used, or if the non-digital materials used were not original (e.g. if they used a questionnaire that already exists).
- Unclear: If unsure whether materials were delivered digitally

Column X: "Other original materials availability statement"

Whether authors explicitly state whether other original materials are available (if used)

Response options: Yes / No / N/A

Notes:

- Yes: Specific mention of these materials being available. Also if they just generically mention 'materials' being available but original non-digital materials were used in the study, check any linked repository or similar to see whether these materials are included there and if so, code as Yes.
- No: Any lack of statement about the availability of these types of materials, even if materials are actually available in a repository, etc.
- N/A: Non-applicable if these types of materials were not used (i.e. if previous column was coded as 'No').

Column Y: "Availability of other original materials"

Level of availability of other original (non-digital) materials (if used).

Response options: Publicly available / Partially available/ Gated / Available on request / Not available, without justification / Not available, with justification / Unclear / N/A

Notes:

- Publicly available: materials fully available and accessible to the public without making accounts or other processes to obtain the data E.g. available on OSF (check links provided in paper to ensure materials there). If link is given to the materials/repository but link doesn't work, code as Unclear and explain this in the notes section (using asterisk as necessary).
- Partially available: this applies if only some of this type of materials used are publicly available.
- Gated: technically available but may need to make an account to access it, or other similar process required to obtain it from a public source. May be behind a paywall.
- Available on request: any mention of contacting the author for these materials.
- Not available, WITHOUT justification: Only use this if it explicitly states that the materials are not available, without a reasonable justification for why not (if they have a reason, code as 'Not available, WITH justification' instead)
- Not available, WITH justification: Any other reason given for why materials are not available, e.g. confidentiality issues, copyright issues, etc.
- Unclear: If paper makes no mention of whether materials are available, or the availability is otherwise unclear and doesn't fit any of the other codes.

2. Article level characteristics

As hypotheses and methodological details can be stated at more than one level of the paper, this initially led to difficulties in fitting these variables into the database format. As a result, the database was spread out into three levels of granularity (article, study, and hypothesis level) and these variables are coded at each level applicable, for maximum clarity

An article level only exists in the coding of a paper when the article contains multiple studies, so that there may be important information at the overall article level that is not captured within the individual studies, such as an overarching general hypothesis that the authors hope to answer using a series of studies, or some overall exploratory analysis (e.g. internal meta-analysis) conducted by pooling the data from all of the studies after they have been conducted.

Coding Patterns for Availability Variables

Only code the availability of code or materials) as N/A if you have coded whether it was used as No (i.e. No for whether it was used, then N/A for whether there is an availability statement, then N/A for the level of availability.

If you have coded whether code/materials were used as Unclear, don't use N/A when coding the statement or level of availability. Instead, code as Unclear for whether it was used, then No for whether there was an availability statement, and then Unclear again for the level of availability. These are the typical patterns of coding when the response for whether that thing was used is not Yes.

If you code whether the code or materials were used as Yes but there is no clear statement about its availability (which would be coded as No), then the level of availability should be coded as Unclear (not as N/A, because the level of availability and whether there is a statement of the availability are only truly non-applicable of they have definitely not been used in the first place i.e. if the column for whether they are used is coded as No)

If an article contains only one study, then there is unlikely to be any additional useful information at the article level other than the availability of data, code, materials, etc. These columns (from statement of the hypothesis at article level, to the use and passing of manipulation checks at article level) can therefore be coded as N/A in papers with only one study.

Column Z: "Hypothesis stated at ARTICLE level"

Whether a hypothesis has been clearly (or partially) stated. This is coded separately for each applicable level of the paper – here we focus on whether it was stated at ARTICLE level (i.e. at an overall/overarching level in a multi-study paper).

Response options: Yes / No / Partially

Response options if competing hypotheses: Yes – competing / Partially - competing

Notes:

A hypothesis might be stated at article level if there are multiple studies reported within the paper, which all investigate the same overarching hypothesis (which is only stated at this overall article level and which may or may not be broken down further or elaborated on at study or hypothesis level but those would then be coded at those levels as well if applicable). For examples, see the annotated example papers and the example coding sheet – the Lynott paper has an article level hypothesis.

Also see example for the next column of how hypotheses may be articulated.

Ideally, an article level hypothesis would be stated within the introduction section but sometimes it may not be stated explicitly until the overall discussion section.

<u>Coding levels</u>

If an article contains only one study, and the study contains only one hypothesis, this is technically stated at all three of these levels, but as we want to capture information at the lowest level of granularity possible, this info should then be coded at the lowest level possible (e.g. hypothesis level if applicable). Therefore, in an article with only one study, there is unlikely to be any additional useful information at the article level anyway - you can therefore code this characteristic as N/A, as there would not be a 'useful' article level in this instance; the same goes for the other characteristics from this column up to the columns coding for use and passing of manipulation checks at ARTICLE level – as this information will realistically not be provided at this level when there is only one study in the article, these can be coded as N/A rather than 'No'.

<u>Moving hypotheses</u>

If there are hypotheses stated at article level, consider whether these can be moved to a lower level of granularity, e.g. if all hypotheses stated at article level are clearly tested in each of the studies within a paper, and you can clearly see their findings related to the support for these hypotheses within each study, these could be moved to be coded at hypothesis level for each of these studies instead. In this case, code as Yes at hypothesis level but add an asterisk to explain in the notes column that it has been moved from article level, and code as No at article level but with an asterisk to link it to the explanation in the notes column (unless there is also support given for each of the article level hypotheses at an overall level of the paper, e.g. in an overall discussion section, in which case you can also code this as Yes at article level, and in subsequent columns give the articulation and support for these article level hypotheses as presented at this overall paper level).

Moving a hypothesis from article level to hypothesis level may be necessary even if there is only one hypothesis at article level in a multi-study paper, if there is no other articulation of the hypothesis at any other level, and this hypothesis is tested in each of the studies. Again in this case code as Yes (or Partially) at hypothesis level with an asterisk to explain in the notes column that it has been moved from article level, and code as No at article level with an asterisk, unless there is also support for the hypothesis presented separately/on an overall article level basis for that article level hypothesis in which case the article level hypothesis can be coded as Yes (or Partially).

Using asterisks in this manner to indicate hypotheses being moved is only necessary when moving from article level to hypothesis level as there is more of a distinction between these levels than there would be between study level and hypothesis level.

<u>Breaking down articulations into separate hypotheses</u>

Sometimes, articulations presented as a single sentence may actually contain multiple predictions and so these articulations may need to be broken down into separate predictions (particularly in order to clearly code the support for these). If breaking a sentence down into separate hypotheses from an articulation that was originally presented as one sentence, code as ‘Yes’ or ‘Partially’ depending on the level of clarity but use an asterisk and explain in the notes section that the articulation has been broken down. You may need to cut some of the words out of the quotes you use from the paper for the articulation of each hypotheses - add ellipses (…) within the quotes to indicate where you have done this.

Sometimes it can be difficult to decide whether it is necessary to break the articulations down; The best way to decide this is often based on how they've presented the results/support for the hypotheses – if the support is presented all together as one clear articulation this may fit better as the support for a multi-faceted prediction that is presented as a single sentence or a single prediction. However, if you find you have to work to piece together multiple different quotes about the results/support in order to answer the different aspects of the prediction, it may be the type of articulation that is worth trying to break down into separate predictions. Usually when you need to break a sentence down into multiple different predictions in a way that is different to how it has been presented in the paper, this should probably be coded as 'Partially' stated, although this depends somewhat on the degree of work needed in order to break down this articulation appropriately.

Partially stated hypotheses

A hypothesis is not always clearly stated but can often still be inferred from clear research questions or attempts to replicate previous studies – this results in the inclusion of 'partially' as a response option. See examples below of this how this is articulated. If in doubt about whether to code this characteristic positively when you need to infer the prediction, aim to be as inclusive as possible – if you can reasonably guess what the authors hoped to find, it can be coded as being partially stated. However, if it is really difficult to make a judgement on whether something can be considered a hypothesis it may be best to just code it as No – whether this is appropriate will depend on the individual paper.

Common articulations of Hypotheses

Simple/Straight-forward coding (usually code as Yes)

- We hypothesize that…
- We predict…
- We expect..
- Testing the hypothesis that X affects Y
- X should affect Y

Coding as Yes or Partially

- "We aimed to replicate …" [might need to refer elsewhere in the text to original study description] . This may be coded as Yes or as Partially stated, depending on the level

and clarity of information given in the paper. Usually these types of articulations are coded as Partially.

Possible 'Partially' Coding (depending on whether hypotheses can be inferred from these statements)

- Some research questions and/or aims
- Examined the effect of ____ on ____
- Some 'whether' statements (Often not substantial enough, or too exploratory)
- Some 'explore' statements… (Often not substantial enough, or too exploratory)
    - "Specifically, we explored whether attributing racially disparate outcomes of police-citizen encounters to implicit, rather than explicit, bias reduces perceptions of police accountability"

Coding as No

Not every aim or research question specified in a paper constitutes a partially stated hypothesis. The following is an example of an aim that would not be substantial enough to be considered a partially stated hypotheses (i.e. would be coded as No):

"The present study sought to investigate the extent to which empathy and nonattachment predicted unique variance in prosociality, above and beyond self-esteem".

In this quote, there is a clear statement of the aim but it is clearly exploratory. The article may then go on to specify particular hypotheses more clearly, which might be coded more positively.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g. multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

Column AA: "Articulation of hypothesis at ARTICLE level"

Articulation of hypothesis, as stated in the paper, at the appropriate level (in this case, ARTICLE level). Provide a direct quote from that level of the paper showing how the hypothesis has been articulated at that level. Include a page number for the quote.

Response options: Open response (quote from paper, with page number for quote), or N/A

Notes:

See examples in example coding sheets and annotated pdfs for other examples of how the hypothesis is stated at various levels. See also the examples of terms used, in the instructions for the previous column.

Choose N/A if hypothesis was never stated at article level and is thus, non-applicable (i.e. if answered 'no' to previous characteristic of whether hypothesis was stated at article level)

Example of hypothesis articulation coded as 'yes'

"Starting from Payne and colleagues' (2010) model, we hypothesized that an individual-level version of the mod would permit us to disentangle individual process estimates for misattribution (parameter M which reflects the extent of confusing prime and target evaluations), attitudes (parameter A which reflects the affective responses to the primes provided that misattribution occurred), and target evaluations (parameter P which reflects the affective responses to the targets provided that misattribution did not occur), respectively. Accordingly, we expected that individual process parameter A to reflect a less confounded estimate of attitudes (here, prejudice against minorities) that could be used for subsequent correlational analyses", p217

Example of partially stated hypothesis articulation:

"The main objective of the present research was to conduct a conceptual replication study (Earp & Trafimow, 2015) of Ott et al. (2018) in order to confirm the newly established homogeneous multimedia effect of multiple symbolic representations, and beyond that, to extend Ott et al.'s investigation about heterogeneous multimedia effects by using functionally equivalent graphics", p4

Column AB: "Hypothesis supported at ARTICLE level

Whether the hypothesis is stated to be supported, or this can be clearly inferred from the study findings. Variable is coded separately at each level, as applicable.

Response options: Yes / No / Partially / N/A / Unclear

Response options if competing hypotheses: Yes – competing / Partially – competing / No – competing / Unclear - competing

Notes:

Where support is not stated explicitly, try to interpret the research findings and determine whether they support the hypothesis. Only use 'unclear' if really impossible to tell – otherwise use your best estimate.

Choose N/A if hypothesis was never stated at article level and is thus, non-applicable.

If some aspects of the hypothesis were supported bit not other aspects of it, or if there was mixed evidence for the hypothesis, this can be coded as 'Partially'.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g. multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

Although in theory competing hypotheses would be expected to have mutually exclusive support (i.e. if one is supported the other must not be supported), this may not always be the case in reality, so be sure to code the support for these as accurately as possible and keep an open mind about these hypotheses both being supported or partially supported, or both being not supported.

Column AC: "Articulation of support for hypothesis at ARTICLE level"

Articulation of support for the hypothesis (or articulation of the relevant study findings), as stated in paper. Included separately at each of the three different database levels as applicable.

Response options: Open response (quote from paper, with page number for quote), or N/A

Notes:

Choose N/A if hypothesis was never stated at article level (i.e. if answered ‘no’ to whether hypothesis was stated at article level) and is thus, non-applicable.

Example of articulation of hypothesis when clearly supported (coded as ‘Yes’) at article level: "Across both studies, the results indicated that dominance congruence reduced relationship conflict when both individuals are low on dominance, but dominance congruence was harmful at high levels of dominance", p359

Column AD: “Intended/Planned sample size at ARTICLE level”

Planned sample size as per sampling plan (e.g. result of power calculation, or other intended number for recruitment). Only code this at article level (as opposed to study level) if the same overall sample was used across all of the studies in a multi-study article.

Response options: Open response (in digits, not words) / Unclear / N/A

Column AE: “Pre-exclusion sample size at ARTICLE level”

Sample size recruited (at ARTICLE level), prior to any exclusions from analysis. Only code this at article level (as opposed to study level) if the same overall sample was used across all of the studies in a multi-study article

Response options: Open response / Unclear / N/A

AF: “Post-exclusion sample size at ARTICLE level”

Sample size in final sample (at ARTICLE level), after any participants have been excluded from the analysis. Only code this at article level (as opposed to study level) if the same overall sample was used across all of the studies in a multi-study article. If intent-to-treat analysis has been used this can be considered as the post-exclusion sample size if necessary and appropriate.

Response options: Open response / Unclear / N/A

AG: “Sampling plan at ARTICLE level”

Type of sampling plan used (if sample exists at article level)

Response options: Frequentist power / Bayesian / Other / Unclear / N/A

Notes:

Frequentist power: The authors may use terms like power calculation or sample size calculation, and they may have used G*Power. Power calculations are often informed by the samples used in previous studies.

Bayesian: The authors may use terms like Bayes factor, or BF.

Other: Other approaches to the sampling plan would include being based on the sample size from a previous study (when this has not been used to inform a power calculation), or may I recruitment plans are limited by availability of resources, e.g. limited funding to pay participants, or working with a population that is limited in numbers, such as trying to obtain a convenience sample of patients with a particular condition attending a certain outpatient clinic at a certain hospital where the study is being conducted – there will be a limited number of such patients and if the researchers aim to recruit as many participants as possible within these limits, this could be considered as 'Other'.

AH: "Sampling unit, at ARTICLE level"

Unit of the sample, to inform other variables (if sample exists at article level)

Response options: Open response but try to keep words used consistent/ N/A

Notes:

Examples: People / Articles / Mice / Neurons / Schools / Dyads

AI: "Word count at ARTICLE level"

This is the word count of the methods section, if the method section of a multi-study paper has been given at an overall article level as opposed to any other level. This is the number of words of the presented overall methods section of the paper, excluding tables and footnotes. To get the word count, copy and paste the methods section into a word document. Make sure any tables, footnotes or article page headers/footers (e.g. name of the journal, or page numbers), are not included in this – therefore you need to copy and paste page by page to avoid this. Don't include the heading 'Methods' when copying/pasting, but subheadings within the section can be retained. Also ensure that the spacing between words copies correctly, as sometimes some of the spaces or hyphens from the original document can be missing, which would cause errors in the word count given by the Word document.

Notes:

Not all elements included the same across papers, e.g., some have details of analysis techniques included in results rather than methods section, or lack of detail in methods sections in later studies of a multi-study paper because methods/measures described in detail in the earlier studies. Just whatever authors have presented as the methods section for that study is included in the word count.

AJ: "Exploratory/additional analysis included at ARTICLE level?"

Whether there is or appears to be exploratory/additional analysis (at article level) in the paper either in addition to, or in the absence of, clearly/partially stated hypotheses. In a multi-study paper, this would be overall exploratory analysis at the ARTICLE level, not exploratory analysis conducted just within a specific study within that article.

Response options: Yes / No / Unclear / N/A

Notes:

Unless there is a clear distinction between confirmatory and exploratory/additional parts, it can be difficult to tell whether and where there is a difference.

Remember that this must be at an overall ARTICLE level to be coded at this level, not for a specific study within that article.

If the article being coded only contains one study, N/A can be used here rather than 'No'.

AK: "Clear distinction between confirmatory and exploratory analysis?" (at ARTICLE level)

Whether there is a clear distinction between confirmatory and exploratory elements of the analysis/findings.

Response options: Yes / No / N/A

Notes/Issues:

It can be difficult to be sure of the distinction at certain points if not clearly stated by authors.

N/A should be used if exploratory analysis has not been included at this level (i.e. if column AJ was coded as 'No').

AL: "Nature of exploratory analysis" (at ARTICLE level)

Whether the exploratory analysis included is general exploratory analysis, or follows up on stated hypotheses, etc. Sometimes the exploratory analysis given, especially at article or study level, can do include both of these – follow up and general exploration, and so it should then be coded as 'Both'. N/A should be used if exploratory analysis has not been included at this level (i.e. if column AJ was coded as 'No').

Response options: Follow up on hypotheses / General exploration / Both / Other / N/A

Notes:

If the exploratory analysis is general but is for clearly specified exploratory aims or research questions, use an asterisk and make a note in the Notes column stating 'Exploratory analysis is r/t aims/RQs'

AM: "Manipulation check 1 at ARTICLE level"

The presence of a manipulation check has been included three times, and at each of the three database levels, to account for the possibility of multiple manipulation checks at each of these levels.

Response options: Yes / No / Unclear / N/A

Notes:

Manipulation checks may be referred to by other terms e.g. validation check, or positive control. Terms such as attention check may also be relevant. General outlier removal or other basic processing of data would not be considered a manipulation check (although participants' data is often removed if they have failed the manipulation check).

Due to lack of familiarity with many of the specialised research topics of the different papers, it can be difficult to identify whether manipulation checks have been used unless it has been clearly stated by the authors, as they may describe them without referring to them explicitly as manipulation checks, assuming readers who are experts in the topic area will recognise these tasks as such.

The use of manipulation checks should only be coded as Yes or No if there are other details at this level because if there are no such details coded at this level (e.g. hypotheses, sample characteristics, etc.) then we can't reasonably expect there to be any manipulation checks conducted at this level anyway, and so N/A would be more appropriate than coding as No.

However, in a study with multiple hypothesis-level hypotheses, if a manipulation check exists but it can't be attributed to any specific hypothesis-level hypotheses (e.g. because it underlies several hypotheses or it ensures basic quality of the study's data overall), this check can be coded at study level instead, even if no hypothesis has been coded at study level.

Note on manipulation checks in systematic reviews:

Systematic reviews etc. don't generally have an article level, and as they usually don't have any hypotheses stated, they often don't have a hypothesis level either – in these cases manipulation checks would be coded as N/A at these levels and the characteristic would only be relevant to study level.

AN: "Passed manipulation check 1?"

Included three times at each level of the paper, to account for multiple manipulation checks

Response options: Yes / No / Unclear / N/A

Notes/Issues:

If the manipulation check was successful, this can be coded as Yes. Also, if those who failed the manipulation check were excluded from the analysis, this can be coded as 'Yes'. Code as N/A if no manipulation check was used (i.e. if the previous column was coded as 'No'). Try to only code as 'Unclear' if absolutely necessary.

AO: "Manipulation check 2 at ARTICLE level"

As for AM

AP: "Passed manipulation check 2?"

As for AN

AQ: "Manipulation check 3 at ARTICLE level?"

As for AM

AR: "Passed manipulation check 3?"

As for AN

3. Study level characteristics

Many of the variables being coded at study level are the same type of characteristics as those coded at the other levels (i.e. as already seen for article level). Remember to default to the lowest level of granularity possible when coding information presented within a study, e.g. if a study contains only one hypothesis, it could technically be considered as either study or hypothesis level, but it is more important to code it at hypothesis level, as this is the lower level of granularity, and information used to test that one hypothesis should also then be coded at the hypothesis level if possible (e.g. sample size, if this is the sample size used to test that one hypothesis). If the methodological details are coded at hypothesis level, the characteristics at study level can be coded as No (for hypothesis being stated) or N/A (for other hypothesis columns, sample details, and manipulation checks), assuming there are no relevant details at study level once the details coded at hypothesis level are accounted for.

However, if there are multiple hypotheses at hypothesis level, methodological details are likely to be coded at study level unless they are different per hypothesis-level hypothesis (e.g. different sample size per hypothesis being tested at hypothesis level).

Most characteristics being investigated at study level are the same as those being investigated at the other levels. However, there are some additional characteristics at this level that have not been included at other levels, e.g., study design, and pre-registration status.

AS: "No. of studies reported"

Number of studies reported in paper or inferred from description within paper, e.g. 3

Notes:

It may sometimes be necessary to infer that there were multiple studies. For example, some papers may describe different phases of the study (e.g. focus group to inform development of an intervention, followed by developing the intervention, then pilot testing the intervention with a different sample – these could potentially be considered as up to three different studies despite being referred to as one study with three separate phases in the paper). If different samples are used for each phase, these should generally be coded as separate studies even if not presented as such in the paper. Similarly, in some papers, studies may describe a pilot study in the same section as they describe a main study – if these are important and have used a different sample than that main study, these would be sufficiently detailed to be considered as separate studies.

AT: "Type of study"

General description of study type

Response options: Clinical / Nonclinical / Animal / Simulation

Notes/Issues:

No other response options have been necessary so far but if something new comes up, add it

Coding Tip: Coding Levels for Methodological Details

Methodological details (i.e., sample details and manipulation checks) will most often be coded at study level unless it is a study with only one hypothesis or when these details are different per hypothesis-level hypothesis (in which cases they should be coded at hypothesis level).

In the relatively rare situations where methods for the studies are described altogether at an overall article level, the methodological details can then be coded at article level if necessary.

Remember that we want to code information at the lowest level of granularity possible.

in red text/highlighted in yellow.

AU: "Design"

Description of design, as stated by authors where possible. Otherwise, design is inferred from the description in the paper.

Response options: Intervention / Case control / Experiment / Observational / Review / Meta-analysis / Secondary data analysis.

Response options for pilot studies: Pilot study – experiment / Pilot study – intervention / Pilot study – observational, etc.

Choose one of these options if possible. If none of these options fit at all, write in appropriate terms and highlight to flag as needed.

Notes/Issues:

When design is not clearly stated by authors, try to infer the design from the paper – this can be difficult but use your best judgement.

Pilot studies

When a pilot study has been used, code it as being both a pilot study, and the closest design you can attribute to it, as per examples above. This is to ensure that pilot studies are clearly identifiable but also indicates what type of approach is involved in the pilot study. However, you do not need to do this if the pilot study is the only study reported in the whole paper, e.g. as in many JMIR papers, because then it is the main (only) study so coding it in this way is unnecessary.

AV: "Registered"

Whether the study has been pre-registered or not and if so, whether it is an RR, because many RR papers also include unregistered pilot studies, or studies that have been preregistered (e.g., on OSF or as clinical trials) but were not part of the accepted stage 1 RR.

Response options: RR / Prereg non-RR / Non-prereg non-RR / Unclear

Notes:

RR status of individual studies can be unclear at times if not clearly stated. One paper can contain both preregistered and non-preregistered studies, so each study must be checked carefully for indications of this. If a protocol is available, check it to see which studies within the paper were preregistered and whether they were part of the accepted RR.

Be careful of any use of hyphens in the response options – this should be consistent to avoid missing some instances if searching for the term later.

AW: "Hypothesis stated at STUDY level"

Whether a hypothesis has been clearly (or partially) stated. This is coded separately for each applicable level of the paper – here we focus on whether it was stated at STUDY level (i.e., at an overall/overarching level within a study, over and above the hypothesis-level hypotheses).

Response options: Yes / No / Partially

Response options for competing hypotheses: Yes – competing / Partially - competing

Notes:

Coding levels

If the study contains only one hypothesis, then this is technically stated at both study and hypothesis level. As we want to capture information at the lowest level of granularity possible, this info should be coded at the hypothesis level rather than study level. So should any relevant study characteristics (e.g. sample size, because although it is a characteristic of the study and so could be considered study level, it is also the sample size used to test that one hypothesis, so is then a hypothesis level characteristic. These details will therefore be coded at hypothesis level rather than study level because we want to default to the lowest level of granularity applicable).

Sometimes only some of the studies within a paper have a hypothesis stated, while others don't.

A hypothesis might be stated at study level if it is a general encompassing prediction over and above the specific hypothesis-level hypotheses. Therefore, it is generally helpful to code the hypothesis-level hypotheses before the study-level hypotheses in order to see if there is a broader or more general prediction left over once the specific hypothesis-level predictions are coded. There is typically not more than one study-level hypothesis per study, and often there are none.

If there is no hypothesis stated in the paper at all at any level, then it is important to code as 'Yes' for exploratory analysis, 'No' for whether there is a clear distinction between confirmatory and exploratory analysis and 'General' for the nature of the exploratory analysis. This will typically only be necessary at study level in such papers, or potentially at article level, but not at hypothesis level. Depending on the paper, it may be good to add an asterisk to the word 'General' and then add a comment in the notes column like 'Exploratory analysis is related to aims/RQs' – this would be in cases where there are no hypotheses that can be inferred but they do still have very clearly stated exploratory aims or research questions. No need to add the quote for the aim/RQ in such cases.

Partially stated hypotheses

A hypothesis is not always clearly stated but can often still be inferred from clear research questions or attempts to replicate previous studies – this results in the inclusion of 'partially' as a response option. See examples below of this how this is articulated. If in doubt about whether to code this characteristic positively when you need to infer the prediction, aim to be as inclusive as possible – if you can reasonably guess what the authors hoped to find, it can be coded as being partially stated. However, if it is really difficult to make a judgement on

whether something can be considered a hypothesis it may be best to just code it as No – whether this is appropriate will depend on the individual paper.

Common articulations of Hypotheses

Simple/Straight-forward coding (usually code as Yes)

- We hypothesize that…
- We predict…
- We expect..
- Testing the hypothesis that X affects Y
- X should affect Y

Coding as Yes or Partially

- "We aimed to replicate …" [might need to refer elsewhere in the text to original study description] . This may be coded as Yes or as Partially stated, depending on the level and clarity of information given in the paper. Usually these types of articulations are coded as Partially.

Possible 'Partially' Coding (depending on whether hypotheses can be inferred from these statements)

- Some research questions and/or aims
- Examined the effect of ____ on ____
- Some 'whether' statements (Often not substantial enough, or too exploratory)
- Some 'explore' statements… (Often not substantial enough, or too exploratory)
    - "Specifically, we explored whether attributing racially disparate outcomes of police-citizen encounters to implicit, rather than explicit, bias reduces perceptions of police accountability"

Coding as No

Not every aim or research question specified in a paper constitutes a partially stated hypothesis. The following is an example of an aim that would not be substantial enough to be considered a partially stated hypothesis (i.e. would be coded as No):

"The present study sought to investigate the extent to which empathy and nonattachment predicted unique variance in prosociality, above and beyond self-esteem".

In this quote, there is a clear statement of the aim but it is clearly exploratory. The article may then go on to specify particular hypotheses more clearly, which might be coded more positively.

Breaking down articulations into separate hypotheses

Sometimes, articulations presented as a single sentence may actually contain multiple predictions and so these articulations may need to be broken down into separate predictions (particularly in order to clearly code the support for these). If breaking a sentence down into separate hypotheses from an articulation that was originally presented as one sentence, code as 'Yes' or 'Partially' depending on the level of clarity but use an asterisk and explain in the notes section that the articulation has been broken down. You may need to cut some of the words out of the quotes you use from the paper for the articulation of each hypotheses - add ellipses (…) within the quotes to indicate where you have done this.

Sometimes it can be difficult to decide whether it is necessary to break the articulations down; The best way to decide this is often based on how they've presented the results/support for the hypotheses – if the support is presented all together as one clear articulation this may fit better as the support for a multi-faceted prediction that is presented as a single sentence or a single prediction. However, if you find you have to work to piece together multiple different quotes about the results/support in order to answer the different aspects of the prediction, it may be the type of articulation that is worth trying to break down into separate predictions. Usually when you need to break a sentence down into multiple different predictions in a way that is different to how it has been presented in the paper, this should probably be coded as 'Partially' stated, although this depends somewhat on the degree of work needed in order to break down this articulation appropriately.

In practice, this doesn't really seem to have been necessary in relation to study level hypotheses so far and instead appears to be much more applicable to article-level and hypothesis-level hypotheses. However, it is still worth bearing in mind as a possibility when coding study level hypotheses.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g., multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

AX: "Articulation of Hypotheses at STUDY level"

Articulation of hypothesis, as stated in the paper, at the appropriate level (in this case, STUDY level). Provide a direct quote from that level of the paper showing how the hypothesis has been articulated at that level. Include a page number for the quote.

Response options: Open response (quote from paper, with page number for quote) / N/A

Notes:

Choose N/A if this is non-applicable (i.e., if answered 'No' to whether a hypothesis was stated at this level).

Where the study reported in a paper is part of a larger overall study e.g. JMIR papers may report on a part of a much larger RCT, such as presenting a qualitative study with a subset of the RCT participants to investigate the usability/acceptability of the intervention, or other perceptions of something. Some such papers report the aims of the larger RCT as well as the aims of the paper itself – these overall aims may not be relevant to the paper you are coding. Similarly, when a study/paper uses secondary data analysis, occasionally they may state the aims of the study whose data they use – these usually will not be relevant to the study/paper you are coding.

<u>Example of hypothesis articulation coded as 'yes'</u>

"Starting from Payne and colleagues' (2010) model, we hypothesized that an individual-level version of the mod would permit us to disentangle individual process estimates for misattribution (parameter M which reflects the extent of confusing prime and target evaluations), attitudes (parameter A which reflects the affective responses to the primes provided that misattribution occurred), and target evaluations (parameter P which reflects the affective responses to the targets provided that misattribution did not occur), respectively. Accordingly, we expected that individual process parameter A to reflect a less confounded estimate of attitudes (here, prejudice against minorities) that could be used for subsequent correlational analyses", p217

Example of partially stated hypothesis articulation:

"The main objective of the present research was to conduct a conceptual replication study (Earp & Trafimow, 2015) of Ott et al. (2018) in order to confirm the newly established homogeneous multimedia effect of multiple symbolic representations, and beyond that, to extend Ott et al.'s investigation about heterogeneous multimedia effects by using functionally equivalent graphics", p4

AY: "Hypotheses supported at STUDY level?"

Whether the hypothesis is stated to be supported, or this can be clearly inferred from the study findings. Variable is coded separately at each level, as applicable.

Response options: Yes / No / Partially / N/A / Unclear

Response options if competing hypotheses: Yes – competing / Partially – competing / No – competing / Unclear - competing

Notes:

Where support is not stated explicitly, try to interpret the research findings and determine whether they support the hypothesis. Only use 'Unclear' if really impossible to tell. Otherwise, try to give best estimate you can.

Choose N/A if hypothesis was never stated at article level and is thus, non-applicable.

If some aspects of the hypothesis were supported but not other aspects of it, or if there was mixed evidence for the hypothesis, this can be coded as 'Partially'.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g. multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

Although in theory competing hypotheses would be expected to have mutually exclusive support (i.e. if one is supported the other must not be supported), this may not always be the case in reality, so be sure to code the support for these as accurately as possible and keep an open mind about these hypotheses both being supported or partially supported, or both being not supported.

AZ: "Articulation of support for Hypothesis at STUDY level"

Articulation of support for the hypothesis (or articulation of the relevant study findings), as stated in paper. Included separately at each of the three different database levels as applicable.

Response options: Open response (quote from paper, with page number for quote), or N/A

Notes:

Choose N/A if hypothesis was never stated at article level (i.e. if answered 'no' to whether hypothesis was stated at article level) and is thus, non-applicable

BA: "Intended/planned sample size at STUDY level"

Planned sample size as per sampling plan (e.g. result of power calculation, or other intended number for recruitment). Code this at study level if this is the sample size used for this entire study (i.e. if it is not broken down separately per hypothesis within the study).

Response options: Open response (in digits, not words) / Unclear / N/A

Notes:

Note on coding planned sample size for systematic reviews.

When coding a systematic review paper, it's not reasonable to expect a planned or intended maximum sample size of articles to include, as this will depend entirely on the search results. As an optimal number of articles (i.e. the 'sample size' of the systematic review) can't be predetermined, this variable can be coded as N/A for systematic reviews. Based on reviews coded so far, these methodological details for review papers will probably only be coded at study level, as they tend not to have article levels (because they usually only report their systematic review which is a single study) and the usually don't have specific hypotheses so do not have a usable hypothesis level. This information has therefore only been given at study level for this variable so far, although this could vary depending on the particular review paper.

BB: "Pre-exclusion sample size at STUDY level"

Sample size recruited (at STUDY level), prior to any exclusions from analysis. Code this at study level if this is the sample size used for this entire study (i.e. if it is not broken down separately per hypothesis within the study).

Response options: Open response / Unclear / N/A

Notes:

Note on coding pre-exclusion sample size for systematic reviews

When coding a systematic review paper, the pre-exclusion sample size should be the number of records found with duplicates removed, before these are screened. Based on reviews coded so far, these methodological details for review papers will probably only be coded at study

level, as they tend not to have article levels (because they usually only report their systematic review which is a single study) and the usually don't have specific hypotheses so do not have a usable hypothesis level. This information has therefore only been given at study level for this variable so far, although this could vary depending on the particular review paper.

BC: "Post-exclusion sample size at STUDY level"

Final sample size (at STUDY level), after any participants had been excluded from analysis. Code this at study level if this is the sample size used for this entire study (i.e., if it is not broken down separately per hypothesis within the study). If intent-to-treat analysis has been used this can be considered as the post-exclusion sample size if necessary.

Response options: Open response / Unclear / N/A

BD: "Sampling plan at STUDY level"

Type of sampling plan used (if sample exists at study level).

Response options: Frequentist power / Bayesian / Other / Unclear / N/A

Notes:

Frequentist power: The authors may use terms like power calculation or sample size calculation, and they may have used G*Power. Power calculations are often informed by the samples used in previous studies.

Bayesian: The authors may use terms like Bayes factor, or BF.

Other: Other approaches to the sampling plan would include being based on the sample size from a previous study (when this has not been used to inform a power calculation), or may I recruitment plans are limited by availability of resources, e.g. limited funding to pay participants, or working with a population that is limited in numbers, such as trying to obtain a convenience sample of patients with a particular condition attending a certain outpatient clinic at a certain hospital where the study is being conducted – there will be a limited number of such patients and if the researchers aim to recruit as many participants as possible within these limits, this could be considered as 'Other'.

BE: "Sampling unit at STUDY level"

Unit of the sample, to inform other variables (if sample exists at study level).

Response options: Open response but try to keep words used consistent / N/A

Notes:

Examples: People / Articles / Mice / Neurons / Schools / Dyads

BF: "Replication study"

Whether the study is original or a replication attempt, either for another author's work, or for their own work reported in an earlier study in the same paper (i.e. an internal replication)

Response options: Original / Direct replication / Indirect replication / Direct internal replication / Indirect internal replication

Notes/ Issues:

Indirect replications may be referred to as 'conceptual replications' in the paper.

The difference between direct and indirect replications can be quite subjective at times. Typically, a direct replication (or 'close' replication) uses the same study methods (or a slightly altered version of the same methods) as the study they are replicating. An indirect or 'conceptual' replication may be more likely to have a different methodology, such as different measurement tasks, different population, etc. but aims to replicate the same finding/effect/relationship as a previous study.

It can be difficult to be certain if a study is original since most research is based on or informed by previous studies to some extent. Therefore, unless the paper explicitly states that they were attempting to replicate some aspect of a previous study or that their study was based on one conducted by a certain author, it can probably be coded as original instead. Some papers may mention in the discussion section that their findings replicate the findings of a previous study. However, unless the authors state or seem to very clearly imply from the beginning of the paper that they were deliberately trying to replicate this previous study's finding, it is probably not a replication and can instead be coded as 'original'.

BG: "Word count of method section at STUDY level"

This is the word count of the methods section, if the methods section is presented at study level in the paper. This is the number of words of the presented methods section of the study, excluding tables and footnotes. To get word count, copy and paste the methods section into a word document. Make sure any tables, footnotes or article page headers/footers (e.g. name of the journal, or page numbers), are not included in this – you will need to copy and paste page by page to avoid this. Don't include the heading 'Methods' when copying/pasting, but subheadings within the methods section can be included in it. Always check that the spacing has copied correctly as some spaces or hyphens can sometimes be missing between words when copied, which would affect the word count detected by the Microsoft Word document.

Notes:

Not all elements included the same across papers, e.g., some have details of analysis techniques included in results rather than methods section, or lack of detail in methods sections in later studies of a multi-study paper because methods/measures described in detail in the earlier studies. Just whatever authors have presented as the methods section for that study is included in the word count.

BH: "Exploratory/additional analysis included at STUDY level?"

Whether there is or appears to be exploratory or additional analysis in the paper (at study level) either in addition to, or in the absence of, clearly/partially stated hypotheses.

Response options: Yes / No / Unclear / N/A

Notes:

Unless there is a clear distinction between confirmatory and exploratory parts, it can be difficult to tell whether and where there is a difference.

BI: "Clear distinction between confirmatory and exploratory analysis?" (at STUDY level)

Whether there is a clear distinction between confirmatory and exploratory/additional elements of the analysis/findings.

Response options: Yes / No / N/A

Notes/Issues:

If no exploratory analysis has been included at this level (i.e. column BH coded as 'No'), then this column can be coded as N/A.

BJ: "Nature of exploratory analysis" (at STUDY level)

Whether the exploratory analysis included is general exploratory/additional analysis, or follows up on stated hypotheses, etc. Can sometimes be both. If no exploratory/additional analysis has been included at this level (i.e. column BH coded as 'No'), then this column can be coded as N/A.

Response options: Follow up on hypotheses / General exploration / Both / Other / N/A

Notes:

If the exploratory analysis is general but is for clearly specified exploratory aims or research questions, use an asterisk and make a note in the Notes column stating 'Exploratory analysis is r/t aims/RQs'

Sometimes the exploratory analysis given, especially at article or study level, can consist of both follow up and general exploration – in these cases, code as 'Both'.

The exact distinction between general and follow up can be difficult to determine – use your best estimate for which of these is appropriate and highlight this section of your coding to be checked if you are not sure.

BK: "Manipulation check 1 at STUDY level"

The presence of a manipulation check has been included three times, and at each of the three database levels, to account for the potential for multiple manipulations checks at each of these levels.

Response options: Yes / No / N/A

Notes/ Issues:

Manipulation checks may be referred to by other terms e.g., validation check, or positive control. Terms such as attention check may also be relevant. General outlier removal or other basic processing of data would not be considered a manipulation check (although participants' data is often removed if they have failed the manipulation check).

Due to lack of familiarity with many of the specialised research topics of the different papers, it can be difficult to identify whether manipulation checks have been used unless it has been clearly stated by the authors, as they may describe them without referring to them explicitly as manipulation checks, assuming readers who are experts in the topic area will recognise these tasks as such.

The use of manipulation checks should only be coded as Yes or No if there are other details at this level because if there are no such details coded at this level (e.g. hypotheses, sample characteristics, etc.) then we can't reasonably expect there to be any manipulation checks conducted at this level anyway, and so N/A would be more appropriate than coding as No.

However, in a study with multiple hypothesis-level hypotheses, if a manipulation check exists but it can't be attributed to any specific hypothesis-level hypotheses (e.g., because it underlies several hypotheses or it ensures basic quality of the study's data overall), this check can be coded at study level instead, even if no hypothesis has been coded at study level.

Note on manipulation checks in systematic reviews:

Systematic reviews etc. usually don't have an article level or a hypothesis level, so manipulation checks would be coded as N/A at these levels, and so manipulation checks would typically only be relevant to study level for these types of papers.

BL: "Passed manipulation check 1?" (at STUDY level)

Included three times to account for multiple manipulation checks, at each of the three database levels (as per previous column)

Response options: Yes / No / Unclear / N/A

Notes:

If the manipulation check was successful, this can be coded as Yes. Also, if those who failed the manipulation check were excluded from the analysis, this can be coded as 'Yes'. Code as N/A if no manipulation check was used (i.e. if the previous column was coded as 'No'). Try to only code as 'Unclear' if absolutely necessary.

BM: "Manipulation check 2 at STUDY level"

As for BK

BN: "Passed manipulation check 2?"

As for BL

BO: "Manipulation check 3 at STUDY level?"

As for BK

BP: "Passed manipulation check 3?"

As for BL

4. Hypothesis level characteristics

If a study only has one hypothesis, or if there are multiple sub-hypotheses stated within the study, these would be coded at HYPOTHESIS level. If there are details about the study that relate to testing that hypothesis specifically, they should be coded at this hypothesis level if possible (e.g. if the study has only one hypothesis and so all methodological details for that study are then coded at hypothesis level; if only a subset of the overall study's sample was used to test one particular hypothesis; or if exploratory analysis was included for only one of the hypotheses but none of the others). However, if there are multiple hypotheses within a study, often the methodological details will apply to the whole study not just to each hypothesis individually (e.g. the same sample will be used for the whole study, not just for one particular hypothesis). Wherever possible, default to the lowest level of granularity possible when coding these characteristics. However, most methodological details will only be coded at hypothesis level if these details are broken down per hypothesis, or if the study has only one hypothesis and so all details for that study are then coded at hypothesis level. Otherwise, the lowest level of granularity for these study details would be the study level.

BQ: "Hypothesis stated at HYPOTHESIS level"

Whether a hypothesis has been clearly (or partially) stated. This is coded separately for each applicable level of the paper – here we focus on whether it was stated at HYPOTHESIS level (i.e. a specific prediction within a study).

Response options: Yes / No / Partially

Response options if competing hypotheses: Yes – competing / Partially - competing

Notes:

Coding levels

If the study contains only one hypothesis, then this is technically stated at both study and hypothesis level. As we want to capture information at the lowest level of granularity possible, this info should then be coded at the hypothesis level. So should any relevant study characteristics (e.g. sample size), because although it is a characteristic of the study and so could be considered study level, it is also the sample size used to test that one hypothesis so is then also a hypothesis level characteristic. It will therefore be coded at hypothesis level rather than study level because we want to default to the lowest level of granularity applicable.

Usually if there are multiple hypotheses within a paper these will be coded at hypothesis level if possible, although occasionally there can be multiple hypotheses at article level too if these can't be moved to hypothesis level (see explanation/notes on moving article level hypotheses to hypothesis level in the earlier section of the protocol, about the statement of hypothesis at article level).

If a paper has no clearly stated hypothesis at any level (i.e. no stated or partially stated hypothesis), there is no hypothesis level so column BQ should be coded as 'No' and the remaining columns at hypothesis level can then be coded as N/A. Methodological details (such as sample size, etc.) would then be coded at study level as that is the lowest level of granularity available when there is no hypothesis level.

Ideally, hypothesis-level hypotheses would be stated within the introduction section of a study but sometimes they may not be stated explicitly until the overall discussion section or even within the methods, analysis or results sections.

Moving hypotheses

If there are hypotheses stated at article level, consider whether these can be moved to a lower level of granularity, e.g. if all hypotheses stated at article level are clearly tested in each of the studies within a paper, and you can clearly see their findings related to the support for these hypotheses within each study, these could be moved to be coded at hypothesis level for each of these studies instead. In this case, code as Yes at hypothesis level but add an asterisk to explain in the notes column that it has been moved from article level, and code as No at article level but with an asterisk to link it to the explanation in the notes column (unless there is also support given for each of the article level hypotheses at an overall level of the paper,

e.g. in an overall discussion section, in which case you can also code this as Yes at article level, and in subsequent columns give the articulation and support for these article level hypotheses as presented at this overall paper level).

Moving a hypothesis from article level to hypothesis level may be necessary even if there is only one hypothesis at article level in a multi-study paper, if there is no other articulation of the hypothesis at any other level, and this hypothesis is tested in each of the studies. Again in this case code as Yes (or Partially) at hypothesis level with an asterisk to explain in the notes column that it has been moved from article level, and code as No at article level with an asterisk, unless there is also support for the hypothesis presented separately/on an overall article level basis for that article level hypothesis in which case the article level hypothesis can be coded as Yes (or Partially).

It can be difficult to decide whether to move the hypotheses in this way; The best way to decide this is often based on how they've presented the results/support for the hypotheses – if this is really only presented at an overall article level, coding at article level may be best, whereas if they give clear information within each study about the support for each of the hypotheses which were originally stated at article level, you should try to move the coding of the hypotheses and their support to hypothesis level.

Using asterisks in this manner to indicate hypotheses being moved is only necessary when moving from article level to hypothesis level as there is more of a distinction between these levels than there would be between study level and hypothesis level.

<u>Breaking down articulations into separate hypotheses</u>

Sometimes, articulations presented as a single sentence may actually contain multiple predictions and so these articulations may need to be broken down into separate predictions (particularly in order to clearly code the support for these). If breaking a sentence down into separate hypotheses from an articulation that was originally presented as one sentence, code as 'Yes' or 'Partially' depending on the level of clarity but use an asterisk and explain in the notes section that the articulation has been broken down. You may need to cut some of the words out of the quotes you use from the paper for the articulation of each hypotheses - add ellipses (…) within the quotes to indicate where you have done this.

Sometimes it can be difficult to decide whether it is necessary to break the articulations down; The best way to decide this is often based on how they've presented the results/support

for the hypotheses – if the support is presented all together as one clear articulation this may fit better as the support for a multi-faceted prediction that is presented as a single sentence or a single prediction. However, if you find you have to work to piece together multiple different quotes about the results/support in order to answer the different aspects of the prediction, it may be the type of articulation that is worth trying to break down into separate predictions. Usually when you need to break a sentence down into multiple different predictions in a way that is different to how it has been presented in the paper, this should probably be coded as 'Partially' stated, although this depends somewhat on the degree of work needed in order to break down this articulation appropriately.

Partially stated hypotheses

A hypothesis is not always clearly stated but can often still be inferred from clear research questions or attempts to replicate previous studies – this results in the inclusion of 'partially' as a response option. See examples below of this how this is articulated. If in doubt about whether to code this characteristic positively when you need to infer the prediction, aim to be as inclusive as possible – if you can reasonably guess what the authors hoped to find, it can be coded as being partially stated. However, if it is really difficult to make a judgement on whether something can be considered a hypothesis it may be best to just code it as No – whether this is appropriate will depend on the individual paper.

Common articulations of Hypotheses

Simple/Straight-forward coding (usually code as Yes)

- We hypothesize that…
- We predict…
- We expect..
- Testing the hypothesis that X affects Y
- X should affect Y

Coding as Yes or Partially

- "We aimed to replicate …" [might need to refer elsewhere in the text to original study description] . This may be coded as Yes or as Partially stated, depending on the level and clarity of information given in the paper. Usually these types of articulations are coded as Partially.

Possible 'Partially' Coding (depending on whether hypotheses can be inferred from these statements)

- Some research questions and/or aims
- Examined the effect of ____ on ____
- Some 'whether' statements (Often not substantial enough, or too exploratory)
- Some 'explore' statements… (Often not substantial enough, or too exploratory)
    - "Specifically, we explored whether attributing racially disparate outcomes of police-citizen encounters to implicit, rather than explicit, bias reduces perceptions of police accountability"

Coding as No

Not every aim or research question specified in a paper constitutes a partially stated hypothesis. The following is an example of an aim that would not be substantial enough to be considered a partially stated hypotheses (i.e. would be coded as No):

"The present study sought to investigate the extent to which empathy and nonattachment predicted unique variance in prosociality, above and beyond self-esteem".

In this quote, there is a clear statement of the aim but it is clearly exploratory. The article may then go on to specify particular hypotheses more clearly, which might be coded more positively.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g. multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

BR: "Articulation of Hypotheses at HYPOTHESIS level"

Articulation of hypothesis, as stated in the paper, at the appropriate level (in this case, HYPOTHESIS level). Provide a direct quote from that level of the paper showing how the hypothesis has been articulated at that level. Include a page number for the quote.

Response options: Open response (quote from paper, with page number for quote), or N/A

Notes:

See examples in example coding sheets and annotated pdfs for other examples of how the hypothesis is stated at various levels. See also the examples of terms used, in the instructions for the previous column.

Choose N/A if a hypothesis was never stated at hypothesis level and is thus, non-applicable (i.e. if answered 'no' to previous characteristic of whether hypothesis was stated at this level)

Example of hypothesis articulation coded as 'yes'

"Starting from Payne and colleagues' (2010) model, we hypothesized that an individual-level version of the mod would permit us to disentangle individual process estimates for misattribution (parameter M which reflects the extent of confusing prime and target evaluations), attitudes (parameter A which reflects the affective responses to the primes provided that misattribution occurred), and target evaluations (parameter P which reflects the affective responses to the targets provided that misattribution did not occur), respectively. Accordingly, we expected that individual process parameter A to reflect a less confounded estimate of attitudes (here, prejudice against minorities) that could be used for subsequent correlational analyses", p217

Example of partially stated hypothesis articulation:

"The main objective of the present research was to conduct a conceptual replication study (Earp & Trafimow, 2015) of Ott et al. (2018) in order to confirm the newly established homogeneous multimedia effect of multiple symbolic representations, and beyond that, to extend Ott et al.'s investigation about heterogeneous multimedia effects by using functionally equivalent graphics", p4

BS: Articulation of Hypothesis identical at study and hypothesis level?

Also need to add at article and study level comparison

Response options: Yes/no/ unclear

Notes:

This WAS being coded but this has been suspended. It remains in the database for now, but can be ignored: can highlight this column in red to make clear that it is no longer relevant

BT: "Hypotheses supported at HYPOTHESIS level?"

Whether the hypothesis is stated to be supported, or this can be clearly inferred from the study findings. Variable is coded separately at each level, as applicable.

Response options: Yes / No / Partially / N/A / Unclear

Response options if competing hypotheses: Yes – competing / Partially – competing / No – competing / Unclear - competing

Notes:

Where support is not stated explicitly, try to interpret the research findings and determine whether they support the hypothesis. Only use 'Unclear' if really impossible to tell. Otherwise, try to give best estimate you can.

Choose N/A if hypothesis was never stated at article level and is thus, non-applicable.

If some aspects of the hypothesis were supported but not other aspects of it, or if there was mixed evidence for the hypothesis, this can be coded as 'Partially'.

Competing hypotheses

Sometimes you may encounter competing hypotheses within a paper, i.e. two or more hypotheses where the support for these would, theoretically, be mutually exclusive, i.e. only one could be supported which would mean the other(s) could not be. In this case, use the response options specified for the competing hypotheses. If there are multiple sets of competing hypotheses within a paper at the same level (e.g. multiple sets of competing hypotheses at the hypothesis level of the same study), please use asterisks to indicate in the Notes columns which hypotheses are competing with which.

Although in theory competing hypotheses would be expected to have mutually exclusive support (i.e. if one is supported the other must not be supported), this may not always be the case in reality, so be sure to code the support for these as accurately as possible and keep an open mind about these hypotheses both being supported or partially supported, or both being not supported.

BU: "Articulation of support for Hypothesis at HYPOTHESIS level"

Articulation of support for the hypothesis (or articulation of the relevant study findings), as stated in paper. Included separately at each of the three different database levels as applicable.

Response options: Open response (quote from paper, with page number for quote), or N/A

Notes:

Choose N/A if hypothesis was never stated at hypothesis level (i.e. if answered 'no' to whether hypothesis was stated at hypothesis level) and is thus, non-applicable

BV: “Articulation of support for Hypothesis identical at study and hypothesis level?”

Also needs to be added at article and study level comparison

Response options: Yes/no/ unclear

Notes:

This WAS being coded but has been suspended. It remains in the database for now, but can be ignored: can highlight this column in red to make clear that it is no longer relevant

BW: ”Intended/planned sample size at HYPOTHESIS level”

Planned sample size as per sampling plan (e.g. result of power calculation, or other intended number for recruitment). Code this at hypothesis level if this is the planned sample size used for one particular hypothesis (i.e. if there is a different planned sample per hypothesis within the study, or if there is only one hypothesis within the study so that the methodological details used to test that one hypothesis are all coded at hypothesis level rather than study level).

Response options: Open response (in digits, not words) / Unclear / N/A

BX: “Pre-exclusion sample size at HYPOTHESIS level”

Sample size recruited (at HYPOTHESIS level), prior to any exclusions from analysis. Code this at hypothesis level if this is the planned sample size used for one particular hypothesis (i.e. if there is a different planned sample per hypothesis within the study, or if there is only one hypothesis within the study so that the methodological details used to test that one hypothesis are all coded at hypothesis level rather than study level).

Response options: Open response / Unclear / N/A

BY: “Post-exclusion sample size at HYPOTHESIS level”

Final sample size (at STUDY level), after any participants had been excluded from analysis. Code this at hypothesis level if this is the planned sample size used for one particular hypothesis (i.e. if there is a different planned sample per hypothesis within the study, or if there is only one hypothesis within the study so that the methodological details used to test that one hypothesis are all coded at hypothesis level rather than study level). If intent-to-treat analysis has been used this can be considered as the post-exclusion sample size if necessary.

Response options: Open response / Unclear / N/A

BZ: “Sampling plan at HYPOTHESIS level”

Type of sampling plan used (if sample exists at hypothesis level).

Response options: Frequentist power / Bayesian / Other / Unclear / N/A

Notes:

Frequentist power: The authors may use terms like power calculation or sample size calculation, and they may have used G*Power. Power calculations are often informed by the samples used in previous studies.

Bayesian: The authors may use terms like Bayes factor, or BF.

Other: Other approaches to the sampling plan would include being based on the sample size from a previous study (when this has not been used to inform a power calculation), or may I recruitment plans are limited by availability of resources, e.g. limited funding to pay participants, or working with a population that is limited in numbers, such as trying to obtain a convenience sample of patients with a particular condition attending a certain outpatient clinic at a certain hospital where the study is being conducted – there will be a limited number of such patients and if the researchers aim to recruit as many participants as possible within these limits, this could be considered as 'Other'.

CA: Sampling unit at HYPOTHESIS level"

Unit of the sample, to inform other variables (if sample exists at study level).

Response options: Open response but try to keep words used consistent / N/A

Notes:

Examples: People / Articles / Mice / Neurons / Schools / Dyads

CB: "Exploratory analysis included at HYPOTHESIS level?"

Whether there is or appears to be exploratory or additional analysis in the paper (at hypothesis level) either in addition to the clearly/partially stated hypotheses.

Response options: Yes / No / Unclear / N/A

Notes:

Unless there is a clear distinction between confirmatory and exploratory parts, it can be difficult to tell whether and where there is a difference

In a paper with multiple hypotheses at hypothesis level, code as yes at hypothesis level if you can attribute the exploratory analyses to a particular hypothesis. If this can't be attributed to a specific hypothesis at hypothesis level or if the exploratory analysis is more of an overall, study-level exploration, it can be coded at study level.

CC: "Clear distinction confirmatory vs exploratory analysis at HYPOTHESIS level"

Whether there is a clear distinction between confirmatory and exploratory/additional elements of the analysis/findings at the hypothesis level.

Response options: Yes / No / N/A

Notes/Issues:

If no exploratory/additional analysis has been included at this level (i.e. column CB coded as 'No'), then this column can be coded as N/A.

Can be difficult to be sure of the distinction at certain points if not clearly stated by authors

CD: "Nature of exploratory analysis at HYPOTHESIS level"

Whether the exploratory analysis included is general exploratory analysis, or follows up on stated hypotheses, etc. Can sometimes be both.

Response options: Follow up on hypotheses / General exploration / Both / Other / N/A

Notes:

If the exploratory analysis is general but is for clearly specified exploratory aims or research questions, use an asterisk and make a note in the Notes column stating 'Exploratory analysis is r/t aims/RQs'

Sometimes the exploratory analysis given, especially at article or study level, can consist of both follow up and general exploration – in these cases, code as 'Both'.

The exact distinction between general and follow up can be difficult to determine – use your best estimate for which of these is appropriate and highlight this section of your coding to be checked if you are not sure.

If no exploratory analysis has been included at this level (i.e. column CB coded as 'No'), then this column can be coded as N/A.

CE: "Manipulation check 1 at HYPOTHESIS level"

The presence of a manipulation check has been included three times, and at each of the three database levels, to account for the potential for multiple manipulations checks at each of these levels.

Response options: Yes / No / N/A

Notes/ Issues:

Manipulation checks may be referred to by other terms e.g., validation check, or positive control. Terms such as attention check may also be relevant. General outlier removal or other basic processing of data would not be considered a manipulation check (although participants' data is often removed if they have failed the manipulation check).

Due to lack of familiarity with many of the specialised research topics of the different papers, it can be difficult to identify whether manipulation checks have been used unless it has been clearly stated by the authors, as they may describe them without referring to them explicitly

as manipulation checks, assuming readers who are experts in the topic area will recognise these tasks as such.

The use of manipulation checks should only be coded as Yes or No if there are other details at this level because if there are no such details coded at this level (e.g. hypotheses, sample characteristics, etc.) then we can't reasonably expect there to be any manipulation checks conducted at this level anyway, and so N/A would be more appropriate than coding as No.

However, in a study with multiple hypothesis-level hypotheses, if a manipulation check exists but it can't be attributed to any specific hypothesis-level hypotheses (e.g., because it underlies several hypotheses or it ensures basic quality of the study's data overall), this check can be coded at study level instead, even if no hypothesis has been coded at study level.

CF: "Passed manipulation check 1?"

Included three times to account for multiple manipulation checks, at each of the three database levels (as per previous column)

Response options: Yes / No / Unclear / N/A

Notes:

If the manipulation check was successful, this can be coded as Yes. Also, if those who failed the manipulation check were excluded from the analysis, this can be coded as 'Yes'. Code as N/A if no manipulation check was used (i.e. if the previous column was coded as 'No'). Try to only code as 'Unclear' if absolutely necessary.

CG: "Manipulation check 2 at HYPOTHESIS level"

As for CE

CH: "Passed manipulation check 2?"

As for CF

CI: "Manipulation check 3 at HYPOTHESIS level?"

As for CE

CJ: "Passed manipulation check 3?"

As for CF

5. Notes

CK: Coding Notes

General notes regarding any issues or lack of clarity encountered – please use an asterisk to denote any variable causing problems in the appropriate column, to link it with the explanation/description in this notes column. If there are difficulties with multiple different

variables, use one asterisk for the first issue encountered, then add an additional asterisk for each additional variable/issue.

CL: Level of difficulty of coding

Response options: Very easy, somewhat easy, somewhat difficult, very difficult.

Indicate how easy/difficult you found the process of coding this particular article, and if appropriate, explain in notes column what aspects were difficult (e.g. granular level of coding, inferring hypothesis, inferring support for hypothesis, detecting manipulation checks, lack of clarity regarding sample sizes, etc.)

CM: Reasons for coding difficulty

Where difficulties were encountered in coding (i.e. if level of difficulty coded as Very difficult or Somewhat difficult), specify reasons for this. If level of difficulty was coded as Somewhat easy, reasons can still be given here if they caused enough difficulty/uncertainty; otherwise, this column can be coded as N/A. If you have coded the level of difficulty as 'Very easy' this column can be coded as N/A as there are not likely to be any specific difficulties if the paper was easy to code.

Response options (Choose as many as apply): Inferring hypotheses; Inferring support for hypotheses; Determining level of coding; Determining level of coding – hypotheses; Determining level of coding – methodological details; Identifying manipulation checks; Inferring support for manipulation checks; Lack of clarity regarding sample sizes; Determining number of studies in paper; Open response

The above list of response options is not exhaustive.

If you encounter other issues, mention them in your own words. In many cases, certain characteristics won't be clear and can be easily coded as Unclear (or Other, No, N/A, etc.) but these do not necessarily have to be included in this list of reasons. This column is more for characteristics that have caused significant difficulties. For example, if you can't see any evidence of a sampling plan being used this can be coded as Unclear, and this characteristic does not need to be mentioned as a reason in this column, whereas if there is some evidence to suggest a different response is more appropriate for the coding of this characteristic but understanding the authors' explanations is very difficult and time consuming, so that it has caused significant delays, difficulty or uncertainty of your coding judgement, this could then be included as a reason for the coding difficulties.

6. SR relevance coding

These columns are unique to the SR coding sheet as they code the relevance of the SR chosen to match that particular RR, in terms of the same journal and timeframe, and how relatively similar the research design, topic, population and/or sample size are, as well as an overall rating derived from the relevance of these six characteristics.

See the example SR justification table and the SR selection protocol for further information about the matching process and how it should be documented.

Column CN: "Journal volume and issue number"

Open response, e.g. 12(4), or 6(3).

Column CO: "Journal relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

As long as the journal is the same, code as 'Very relevant'. For SRs from Journal of Social Psychology chosen as SRs for papers from Comprehensive Results in Social Psychology (which only publishes RRs), code as 'Somewhat relevant'.

Column CP: Timeframe relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

If the SR is from same issue as the RR, code as 'Very relevant'. If it is from one issue either side of the issue the RR was in, code as 'somewhat relevant'. If it is one more issue out, code as "somewhat irrelevant".

CQ: "Topic relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

While an exact topic match may not be very likely, if it is the same general topic area, code as 'Very relevant', e.g. if both papers deal with effects of app-based interventions for

depression or low mood. If the BA's topic is not close enough to be coded as 'Very relevant' but is still not too dissimilar, code as 'Somewhat relevant'.

CR: "Design relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

If the designs are very similar select as 'Very relevant', e.g. if they are both an RCT, or both a systematic review. If comparing the relevance of a study using an online experiment vs. one using an in-person experiment, code as "somewhat relevant".

CS: "Population relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

While an exact population match may not be very likely, if it is the same general population group, code as 'Very relevant', e.g. older adults with cancer, young people with depression, or neurotypical undergraduate students. If the population is not too dissimilar but not close enough to be coded as very similar, code as 'Somewhat relevant'.

CT: Sample size relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

There is unlikely to be a close match on this characteristic and it does not seem to be particularly helpful when selecting SRs but it is important to acknowledge any major differences in sample size between an RR and its matched SRs, particularly if the other matching characteristics have been considered very relevant. Coding for this will be fairly subjective but consider if the sample size is completely dissimilar or not, e.g., if the RR has a sample of 2,00 and the SR has a sample of 15, this would be 'Very Irrelevant', whereas if the SR had a sample of 1,500 or 1,750, this could reasonably be coded as 'Very relevant'. There will be a certain degree of subjectivity in the coding of this characteristic so use your best judgement.

CU: "Overall SR relevance"

Response options: Very relevant / Somewhat relevant / Somewhat Irrelevant / Very Irrelevant

Notes/ Issues:

Based on the coding of the relevance of the SR characteristics in the previous 6 columns, select an overall ranking for how relevant the SR is for the RR it is matched with. As a guide, use the following ranking system:

For each of the 6 relevance characteristics, they are coded on a 4-point scale (Very Irrelevant to Very relevant). If we assign a number to each of these, we can mark each characteristic out of 3:

Very Irrelevant = 0

Somewhat Irrelevant = 1

Somewhat relevant = 2

Very relevant = 3

If we mark each of these 6 characteristics out of 3 like this and them add them all together, we get a total number out of a maximum possible total of 18. In order to determine an overall relevance ranking for the BA, use the following (approximate) scale:

0-4 = Very Irrelevant

5-9 = Somewhat Irrelevant

10-14 = Somewhat relevant

15-18 = Very relevant

While this scale is not totally even, it gives an approximate system for assigning an overall category for the BA's relevance, based on the relevance of the characteristics it was matched on.

CV: "Notes on SR relevance coding"

Open response. Use asterisks to link notes with relevant columns re: relevance.

7. **Additional columns (Column depends on whether in RR or SR coding sheet)**

CN (RR sheet) / CW (SR sheet): Non-final copy coded?

Response options: Preprint / Pre-proof / Accepted copy or authors' copy / Other

If any other option, please specify if possible or make a note in the Notes column.

Sometimes the pdf provided is not a final version. If you encounter a pdf that is not a final version, check online to see if you can find a more recent version, and if so, use that version instead (but check with me before using it, just to be sure). Even when an article has been published for years, sometimes the pdf that is downloadable from the article's webpage on a journal's site is still marked as being the accepted manuscript, or author's copy. Or, if a paper has been published recently, occasionally it may be a pre-proof article (e.g. hasn't been copyedited for language clarity etc but the content should all be the same). While this shouldn't occur very often, if we encounter a paper where the only available copy is like this, it should be documented in this column using the above response options just in case this ever becomes an issue in the future.

CO (RR sheet) / CX (SR sheet): Coded by

Open response: First coder's initials

CP (RR sheet) / CY (SR sheet): Coding double- checked by

Open response: second coder / checker's initials

CQ (RR sheet) / CZ (SR sheet): Coding triple-checked by

Open response: third coder / checker's initials

DA (SR sheet only): SR selected by

Open response: First selector's initials

DB (SR sheet only) SR selection checked by

Open response: second checker's initials

CR (RR sheet) / DC (SR sheet): Additional tags (qualitative / mixed methods)

Response options: Qualitative / Mixed methods / N/A