Researcher Degrees of Freedom in the Psychology of Religion

Charles, SJ, Bartlett, JE, Messick, K, Coleman III, T & Uzdavines, A

DOI 10.1080/10508619.2019.1660573
ISSN 1050-8619

Publisher: Taylor and Francis

Researcher Degrees of Freedom in the Psychology of Religion

[a]Sarah J. Charles, [a, b]James E. Bartlett, [a]Kyle J. Messick, [a, c]Thomas J. Coleman III, and [d]Alex Uzdavines

[a]Coventry University, Brain, Belief, and Behaviour Research Lab; Centre for Advances in Behavioral Science, UK

[b]Arden University, School of Psychology, Law, and Social Science, UK

[c]Grand Valley State University, Department of Psychology, USA

[d]Case Western Reserve University, Department of Psychological Sciences, USA

Corresponding author: Sarah J. Charles, Charle42@uni.coventry.ac.uk

ORCID ID: Charles: 0000-0002-3559-1141; Bartlett: 0000-0002-4191-5245; Messick: 0000-0002-0452-0922; Coleman: 0000-0002-3003-5090; Uzdavines: 0000-0001-5829-9648

Abstract

There is a push in psychology toward more transparent practices, stemming partially as a response to the replication crisis. We argue that the psychology of religion should help lead the way toward these new, more transparent practices to ensure a robust and dynamic subfield. One of the major issues that proponents of Open Science practices hope to address is researcher degrees of freedom (RDF). We pre-registered and conducted a systematic review of the 2017 issues from three psychology of religion journals. We aimed to identify the extent to which the psychology of religion has embraced Open Science practices and the role of RDF within the subfield. We found that many of the methodologies that help to increase transparency, such as pre-registration, have yet to be adopted by those in the subfield. In light of these findings, we present recommendations for addressing the issue of transparency in the psychology of religion and outline how to move toward these new Open Science practices.

Keywords: Open Science, Researcher Degrees of Freedom, Pre-Registration, Registered Reports, Psychology of Religion.

Introduction

Nelson, Simmons, and Simonsohn (2018) described the field of psychology as being in a reformation period after a "replication crisis". This crisis is defined by many research methods, findings, and publication practices being recognised as less robust than originally perceived. For example, contemporary research is often underpowered (Button, Ioannidis, & Mokrysz, 2013), despite repeated warnings about low sample sizes (Cohen, 1962; Sedlmeier & Gigerenzer, 1989). Journals publish articles with low-quality control; up to 50% of published articles contain a statistical reporting error (Brown & Heathers, 2017; Nuijten, Hartgerink, & van Assen, 2016). Many studies originally seen to be robust cannot be replicated, even with larger sample sizes (e.g., Open Science Collaboration, 2015; Wagenmakers et al., 2016). Finally, publication bias appears to be particularly prevalent in psychology as results almost always favour a researcher's hypothesis (Fanelli, 2010). The reformation movement in psychology, part of the broader Open Science movement, is focused on mitigating these issues. Special issues presenting solutions to these problems for psychology, both as a whole and within specific sub-fields, have proliferated (e.g., Asarnow et al., 2018; Kappenman & Keil, 2017; Rouse, 2018). Here, we focus on one part of this reformation: researcher degrees of freedom.

Researcher degrees of freedom (RDF) are the numerous choices made in the process of conducting research. These choices are often arbitrary and can influence the outcome of significance testing. This, in turn, can push the conclusions into different directions, depending on the decisions that the researcher made (Simmons, Nelson, & Simonsohn, 2011; Wicherts et al., 2016). This is why RDF is an issue: the opportunistic use of this flexibility to tweak data, especially when undocumented in manuscripts and thus without transparency, is ethically problematic. A non-exhaustive list of decisions researchers need to make includes sample size,

exclusion criteria, which measures to report, cut-off points for comparison groups, what covariates to include, and how to calculate outcome variables. Roettger (2018) provides a comprehensive overview of how RDF appear throughout the research timeline and reports a simulation that demonstrates how RDF dramatically increase the false positive rate.

Flexibility in performing data analyses is problematic (Lane et al., 2016), especially when combined with publication bias (Bakker, van Dijk, & Wicherts, 2012). In an exaggerated example, Simmons, Nelson, and Simonsohn (2011) showed how undisclosed flexibility in analytic decision making can lead to nonsensical but statistically significant results. In a real study, participants listening to the Beatles' song "When I'm Sixty-Four" were found to be a year and a half younger than those who listened to "Kalimba". However, this finding only appeared after dropping an additional condition, including participants' fathers' age as a covariate, and because they repeatedly analysed their data while collecting data until they found a significant effect. This process is known as "*p*-hacking", i.e. iterating on an analysis until the result is statistically significant (Head et al., 2015). This procedure is not necessarily nefarious. For example, checking for outliers or covariates is often advisable, but if this process is not outlined before checking the data, then post-hoc rationalisations can be used for determining which definition of 'outlier' to use. In this example, many definitions of 'outlier' would make sense, and so one that changes a non-significant result into a significant result may be chosen based on seemingly valid reasoning to all those involved. We believe that this is an inevitable result of a lack of pre-defined controls over the flexibility in analytic decision making. Indeed, most researchers do not recognize most of these behaviors as questionable (John, Loewenstein, & Prelec, 2012).

The example in Simmons et al. (2011) is exaggerated, but the innocuous nature of flexibility in data analysis can be seen in a more realistic example. Twenty-nine teams of analysts answered the same research question using the same dataset (Silberzahn et al., 2018). Out of the 29 teams, 20 found a significant result. Combined, the teams used 21 unique combinations of covariates. This supports Gelman and Loken's (2013) "garden of forking paths", the variation and flexibility permitted by a complex dataset. This flexibility is often rationalized as making sense of data on the path to answering a research question. If the procedures are fully transparent there is no problem with repeatedly analyzing data. However, these results should be specifically labelled as exploratory (de Groot, 1954/2014).

We argue that RDF are not a problem themselves, but become problematic when there is a lack of transparency surrounding researchers' decisions and their rationale for making them. Undisclosed flexibility in RDF, (also dubbed the opportunistic use of RDF; Wicherts et al., 2016), should be viewed as a cognitive bias with steps put into place to mitigate its effect. Fortunately, we have reached a point in the reformation where various authors have written practical suggestions for change (Nelson et al., 2018), and we can focus on the solutions presented to minimize this bias in the future.

One suggestion to control the opportunistic use of RDF is using a pre-registered analysis plan (van't Veer & Giner-Sorolla, 2016) or a registered report (Chambers, 2013; Chambers, Feredoes, Muthukumaraswamy, & Etchells, 2014; Nosek & Lakens, 2014). Pre-registration is a process where researchers outline their hypotheses, data collection methods, measures, and analyses in advance via an online repository. A registered report is the logical extension of this: the introduction and methods sections are written in advance and submitted to a journal as a stage-one manuscript prior to data collection (Chambers et al., 2014). Editors and reviewers

evaluate the stage-one paper and decide whether or not to provide in-principle acceptance based on the research question and quality of the authors' proposed methods. If so, the final manuscript will be published regardless of whether the researcher's hypotheses were supported, providing the final methods were consistent with the stage-one submission. After data collection, analysis, and write-up, the paper is then submitted as a stage-two manuscript for a second round of peer review. Editors and reviewers check that the authors adhered to their proposed methods, provided sound justification for any deviations, and that the overall quality of the write-up meets the standards of the journal. So long as the methods remained consistent, or changes in the analytical plan were validly justified between stage one and stage two, this second stage of review is meant to refine the paper for publication more than to act as a gatekeeping stage. The provisional acceptance of a paper prior to data collection removes problems related to RDF and publication bias.

A selection of RDF that pre-registration can cover are: sample size justification, outlier removal, data analysis plans, materials/scales, and specifying hypotheses. Failing to control for these RDF increases the false positive rate and reduces the reliability of published literature. It is important to have a clear *a priori* plan for how many participants will be included, as repeatedly analyzing data during data collection increases the false positive rate (Simmons et al., 2011). Using an opportunistic combination of data preprocessing techniques, including outlier removal, increases the false positive rate (Morís Fernández & Vadillo, 2019). In addition, when authors perform significance tests the alpha level is normally set to 0.05 for an individual test. However, when multiple tests are performed, the experimentwise error rate is the alpha level multiplied by the number of tests performed. Not controlling for multiple tests, thus modifying the alpha, increases the rate of false positives (Bender & Lange, 2001). Given that these factors can all

increase the rate of false positives, we believe that they are particularly important to constrain during pre-registration.

Why the Field of Psychology of Religion?

The psychology of religion has frequently been criticized due to the inevitability of researchers' predispositions toward the subject matter biasing their research (Ladd & Messick, 2016; Messick & Farias, 2019; Wulff, 1998). In addition, researchers have argued that the field has potentially been slanted in a pro-religious direction due to religiously-motivated organizations funding a substantial amount of the field's research (Ambasciano & Coleman, 2019; Bains, 2011; Wulff, 2003). However, these issues may be less widespread today than they once were. Still, the field has far from removed itself from the shadows of the past. The field is disproportionately composed of individuals who are themselves religious, which often puts those individuals at odds with the wider scientific community. The psychology of religion has, throughout its history, struggled to thrive due to the hostility in which many scientists outside of the field approach religion (Wulff, 1998).

Another area of contention within the field is the lack of consistency in terminology. Categories such as 'religion' and 'spirituality' are polysemantic and resist straightforward definition. This results in constructs, operationalizations, and measurement scales that can vary widely in meaning and the terminology used to describe them (Coleman & Jong, in press). In part, this reflects the challenges associated with defining and measuring all sociocultural phenomena (Coleman & Hood, 2015; de Jager Meezenbroek et al., 2012; Moberg, 2002; Slater, Hall, & Edwards, 2001). However, utilizing Open Science practices to increase transparency of RDF when studying religion and spirituality is especially important due to the amorphous nature of the terms and highly-charged nature of public discussion around these topics.

Although the use of pre-registration and registered reports has increased across psychology (Nosek & Lindsay, 2018), we have anecdotally noticed that these meta-methods appear less frequently in the psychology of religion (Uzdavines, Hill, Coleman, Gibson, & Stauner, 2017). To investigate this observation, we conducted a systematic review of articles published in three major psychology of religion journals to audit the uptake of Open Science methods. We audited across all issues of three psychology of religion journals from 2017. We reasoned that by 2017 enough time had passed since the major events triggering the replication crisis and discussion of reform between 2011 and 2015 for uptake of Open Science methods to begin within the psychology of religion (e.g., Open Science Collaboration, 2015; Simmons et al., 2011). Our hypothesis is that the recommendations for reform have not penetrated the psychology of religion. We believe that if we find very few articles transparently mitigating researcher degrees of freedom, we are justified in proposing how the psychology of religion can embrace Open Science practices.

We present: (1) a systematic review of psychology of religion literature to understand the extent of RDF and (2) ways that researchers within psychology of religion can control for RDF by using Open Science practices. We hope these recommendations will strengthen our discipline and further standardize how we propose and report research, increasing trust of the field among outsiders beyond the psychology of religion.

Methods

We pre-registered the protocol for this systematic review on the Open Science Framework (OSF; https://osf.io/fs67k/?view_only=e19600c738d247ff926c0d5e9378c6f0). This included our hypothesis, eligibility criteria, and review process. Any omissions or deviations from this protocol are highlighted in this article. The nature of this study was exploratory as our

hypothesis was not operationalized. We were looking at various possible RDF to find what can be learned from the sampling frame as a basis to make recommendations to others in the field.

Sampling Frame

In an effort to capture a broad perspective on the state of the psychology of religion, we reviewed all of the articles published in the 2017 issues of three major journals: *Archive for the Psychology of Religion* (*APR*)*, International Journal for the Psychology of Religion* (*IJPR*)*,* and *Psychology of Religion and Spirituality* (*PRS*). These three journals represent the most widely read psychology of religion journals focusing purely on psychology and religion, without other sub-field specialization. Other journals that we considered, but discounted, were '*Mental Health, Religion, and Culture*', '*Journal of Religion and Health*', '*Journal of Psychology and Theology*' and '*Journal of Psychology and Christianity*'. We dismissed them for focusing on theological matters, social anthropology, or Christianity rather than being generally representative of the field.

For inclusion, an article had to be (1) empirical, (2) include quantitative methods in at least one part of the article , and (3) published in the year 2017. Despite growing interest in Open Science practices in qualitative research (Tamminen & Poucher, 2018), our review and subsequent recommendations focus on quantitative methods. Before applying our exclusion criteria, the sampling frame consisted of 14 (*APR*), 21 (*IJPR*), and 45 (*PRS*) articles. The final sample included 8, 14, and 31 articles respectively.

Data Extraction

The primary outcome we coded for was the proportion of papers that reported constraints on RDF. Constraints were based on the pre-registration checklist by Wicherts et al. (2016). This checklist was designed for authors to use in publishing research, therefore we focused on the

RDF constraints that would be most easily observed. Our codes included: whether the study was pre-registered, whether sample size was justified *a priori* (e.g., through power analysis), and whether any measures were modified without validation, also known as a questionable measurement practice (Flake, Pek, & Hehman, 2017). Most categories were coded using a dichotomous present or not present rating with additional options available for when a code was not applicable to the article or when the presence of a category was unclear. When the two coders compared results and discussed any discrepancies, we determined that some of our initial codes were too interpretative and dropped them to focus on RDF less prone to differences in interpretation. For example, our original coding scheme included whether the definition of a construct was consistent with the measure the author used. This is highlighted as a deviation from our pre-registered coding scheme, as the data were recorded and are available on the OSF, but are not included in the final manuscript. As a result, we did not code for some data analysis RDF highlighted in Wicherts et al.'s checklist, including justifications for correcting or discarding data, dealing with assumption violations, and outlier removal. We also did not code for certain specialized statistical RDF. The full coding scheme is available on the OSF (https://osf.io/yd7k9/?view_only=e19600c738d247ff926c0d5e9378c6f0).

Two of the authors (KJM and TJC-III) independently completed our coding scheme before comparing results. SJC compared the results and calculated the interrater reliability. We used Gwet's AC1 (Gwet, 2008) instead of the commonly used Kappa coefficient as its statistical properties are more favourable. AC1 provides a less biased estimate of the 'true' interrater reliability (Gwet, 2008) and avoids underestimating interrater reliability when there is an imbalance in the marginal totals of the table (Wongpakaran et al., 2013). We decided to use a cut-off of .80 as the accepted rule of thumb for a 'high' value of Kappa-like measures (Bajpai,

Bajpai, & Chaturvedi, 2015, p. 26). If the value was below .80, SJC provided her input and the group came to a final agreement. If the value was .80 or above, a discussion was held between KJM and TJC-III to discuss which categories the differences fell within with SJC moderating the discussion.

Results

Inter-Rater Reliability

Once both KJM and TJC-III completed their coding independently, SJC calculated interrater reliability for each of the 15 measures that were relevant to RDF in the psychology of religion (see Table 1). The raw data and R (R Core Studio, 2017) script are available on the OSF (https://osf.io/b6n8t/?view_only=7585611b8ffa4fe387abe8caca278cc4). Due to the low level of agreement across multiple measures, SJC provided input to help the group come to a consensus. We acknowledge here that what constituted a 'low level of agreement' was not defined in the pre-registration. We used the .80 value for AC1 that Bajpai, Bajpai, and Chaturvedi (2015) provide due to its general acceptance. We added this comment to increase the transparency of our own RDF, as a guide to others on how to include such comments in their own work.

[Insert Table 1 here]

Systematic Review - Overview

Of the 80 journal articles across the three journals of interest, 27 papers did not meet our inclusion criteria due to being review papers, purely qualitative papers, or theoretical papers. We included 53 papers for full coding. The completed coding scheme is available on the OSF (https://osf.io/mj7sp/?view_only=2612658d15d14c84a559a082d847b488). In addition, see Table 2 for a summary of codes and percentages of papers matching each code following consensus.

[Add Table 2 Here]

No researchers included a pre-registration with their studies, although two authors did mention that they should have done so and would seek to do so in the future.

RDF in Defining and Rationalizing Measures

Definitions of variables should always clearly match up with the operationalization of those variables. A discrepancy between the definition of a construct and how it is operationalized into a measure causes problems with the validity of the data and analytical claims. This is especially true in the psychology of religion, where abstract concepts like religion and spirituality are regularly studied. Not only are these concepts difficult to define, they are difficult to measure. We identified inconsistencies between how R/S constructs were defined and how they were measured in instances where either the definition was vague or non-existent. We also identified cases where a definition is clearly addressing a different aspect of spirituality or religion than what the scale was measuring. Of the included papers, 45 (85%) provided definitions for all their Religious/Spirituality (RS) measures and one (2%) had no RS measures. The remaining seven (13%) used RS terminology, but did not provide a definition for all RS measures. For the clarity of these definitions, we coded six (11%) for using only vague definitions for their RS measure(s). Three (6%) papers used more than one RS measure with a mix of vague/undefined and clear definitions. Thirty-nine (73%) of the papers included a clear definition for all of their RS measure(s). Four (8%) papers were among those that did not define their RS measures. One article (2%) had no RS measures at all.

We also coded for whether or not the RS measures authors used matched the definitions of the RS variables they described measuring. We coded two (4%) papers as having non-matching RS measure definitions, 40 (75%) had matching definitions for all their RS measures, 10 (19%) had a mixture of matching and non-matching between their variable definitions and

measures, and one (2%) paper had no RS measures at all. Notably, however, our interrater agreement was very low. This result should be interpreted with caution. Some of our difficulty with agreement within this area stems from difficulty picking apart RDF from issues of validity in measurement selection.

For rationales behind the RS measure(s) used, 50 papers (94%) provided a full rationale for the authors' decision while three (6%) included partial rationales. The authors of four (8%) papers created their own measures. In papers that used previously published RS measures, three (6%) modified all of their measures in some way, 14 (26%) modified at least one measure, and 32 (60%) did not modify their measures. Of the 17 papers that modified their measures in some way, all provided an explanation for cut-offs for these scales, where applicable. Only three did not provide a rationale for modifying the measure. However, we had very low levels of agreement regarding this criteria and these results should be interpreted cautiously.

RDF in Statistical Analysis

We coded for 1) if a study was exploratory or confirmatory, 2) if there was a rationale for the sample size, and 3) to what extent Type I and Type II errors were accounted for to explore the use of RDF in statistical analysis. The research approach of 19 (36%) papers was exploratory in nature, while 25 papers (47%) used a confirmatory approach. Exploratory studies were defined as testing a non-directional hypothesis, while confirmatory studies were defined as testing a directional hypothesis. The remaining nine (17%) mixed exploratory and confirmatory analyses. Twelve (23%) papers provided a rationale for sample size, but only two of these used an *a priori* power analysis. The remaining 41 (77%) articles did not include any justification for their sample sizes. For Type I and Type II error correction, four (8%) papers corrected for Type I error and 43 (81%) papers did not mention whether they corrected their results for possible Type

I errors. One (2%) paper explained why they did not include any corrections. Type I error correction was not applicable in the remaining five (9%) papers (e.g., if only one test was conducted).

RDF in Reporting Results

There are a number of ways RDF can appear in how authors report their results. We coded for whether or not authors provided exact *p*-values or reported non-significant results as *trending* or *marginally significant* to guide the narrative in a direction that still supports the author's claims. During our initial coding, we found that 25 (47%) of the 53 papers included in our analysis gave exact *p*-values throughout the text of the paper. However, we then decided to include tables in our coding scheme. Of the 25 papers we initially coded as providing exact *p*-values throughout the text, nine (17%) of these followed the correlation table convention of using an asterisk to mark gradations of *p*-value (e.g., * indicates $p < .05$, ** indicates $p < .01$) instead of providing exact *p*-values in their tables. After deliberation between coders, we decided that this variation in reporting *p*-values between text and tables qualified as not providing exact *p*-values throughout the entirety of the paper. We acknowledge this was not included in our pre-registration as we did not anticipate this scenario. Thus after our final coding, only 16 (30%) papers provided exact values for every *p*-value, 20 (38%) did not provide exact *p*-values and 16 (30%) papers varied in whether, and how, they reported exact *p*-values throughout their results sections. This was not applicable for one (2%) of the papers as it only included descriptive statistics.

We also coded for the presence of RDF related to reporting non-significant results as "trending" or "marginal" and then considering these non-significant results within the discussion

in a way that implied the authors' believed them to be "almost-significant". We found that nine (17%) papers considered non-significant results "marginally" significant and/or considered results above the authors' alpha value, usually of $p < .05$, as significant in some way. The remainder (83%) did not do so. Of the nine papers, five (9%) over-inflated their results in their discussion/conclusion sections (e.g., making a "marginally significant" effect one of the key findings of the article). The remaining articles (91%) did not over-inflate their results. While other psychology disciplines found higher rates for reporting $.05 < p < .10$ as "marginally" significant, 30.1% in clinical psychology and 45.4% organizational psychology (Olsson-Collentine, van Assen, & Hartgerink, 2019), the rate in our sampling frame is still higher than ideal.

Discussion

In this systematic review, we investigated the transparency of RDF within the psychology of religion. We did this after noticing, anecdotally, that the signs of movement toward using Open Science practices seen in other parts of the psychological sciences had yet to appear within the field (Uzdavines et al., 2017). We conducted a systematic review of all papers published by three major psychology of religion journals in 2017. After completing the review, we found many instances where RDF could have been better controlled or accounted for, demonstrating a lack of transparency in the use of RDF and making it difficult to know exactly what process was followed by the authors of the published studies. As such, we believe that making recommendations regarding how the psychology of religion can better handle and report RDF is warranted.

Systematic Review - Deeper Analysis

Papers with unaccounted for RDF were spread across all three journals. This suggests that these practices exist across the psychology of religion, as opposed to being journal-specific, and that there is a problem with transparency in the use of RDF throughout the discipline. Our review demonstrated the lack of adoption of pre-registration, as we found no pre-registered articles. This means we cannot know to what extent researchers changed their hypotheses after data analysis. Comparatively, 19 articles were pre-registered in the journal *Psychological Science* alone in the year 2017 (Nosek & Lindsay, 2018). Given five years have passed since psychological research methods and practices first received scrutiny for increasing RDF (see, Nelson et al., 2018 for a brief history), it is surprising that no articles were pre-registered in any of the three psychology of religion journals. Between the beginning of the reformation and the 2017 articles, many calls moving to pre-registration have been made in social and psychological sciences (e.g., Bakker et al., 2012; Miguel et al., 2014; Wagenmakers et al., 2012).

We also found a lack of Type I error correction, or even the mention of it, within many of the articles within our sampling frame. These omissions bring the reliability of results presented in these articles into question. Significant findings might be an artifact of a lack of correction for multiple tests rather than true effects. As previously discussed, it is unlikely that any of these authors omitted Type I error correction intentionally, rather this is an example of unaccounted RDF.

Another issue is the lack of sample size justification within most of the sampling frame, which may have led to underpowered studies. Only 12 of 53 (23%) articles provided any justification for the sample sizes used in their research and only two (4%) of these studies reported a power analysis. Similar audits of sample size calculations have been performed in other fields, with no studies reporting a power analysis in a selection of psychophysiology

articles (Larson & Carbine, 2017) and only a single study reporting one in fMRI research (Guo et al., 2014). Researchers often use rules of thumb to determine their sample size and consistently overestimate the statistical power provided by their samples (Bakker, Hartgerink, Wicherts, & van der Maas, 2016). This review suggests that research within the psychology of religion is still conducted without statistical power in mind, despite repeated warnings of the implications of underpowered studies (Cohen, 1962; Sedlmeier & Gigerenzer, 1989; Button et al., 2013). This is a double-barrelled issue, as the effect of the opportunistic use of RDF is more severe in underpowered studies (Bakker et al., 2012). Moreover, we cannot know something that is not included in a paper. As such, if authors are not being transparent about RDF, we cannot know about the unstated actions or intentions of the author. This extends to sample sizes. Without a clearly justified sample size, it becomes more difficult to evaluate whether or not researchers engaged in data peeking, a sub-type of *p*-hacking.

Finally, we coded for the presence of confirmatory and/or exploratory research but few papers made this distinction explicit. We coded for this in order to help better contextualize the presentation of the results of the papers in our sampling frame. The interpretation of results can be skewed by whether or not research is exploratory or confirmatory in nature. When research is exploratory, the results should be interpreted as a basis for new ideas, hypotheses and further research. On its own, exploratory research has less evidential value (de Groot, 1954/2014). For confirmatory research, a specific hypothesis should be tested in a clear way with an *a priori* analytical plan (Wagenmakers et al., 2012). However, the line between exploratory and confirmatory research is often blurred in psychological research, with the majority of publications purporting to test and support a researcher's hypothesis (Fanelli, 2010). Advances in

research are achieved through a cycle of exploratory and confirmatory research, but it is crucial that the distinction between the two is explicit.

Practical Recommendations

Based on the results of our systematic review and other articles on reducing RDF (e.g., Wicherts et al., 2016), we believe these recommendations will help improve the level of transparency surrounding RDF and the quality of research in the psychology of religion.

Journal article policy. One reason for the lack of pre-registered articles found in this review could be that those who did pre-register their studies published somewhere that provides an incentive for doing so. As of the beginning of 2017, none of the journals within our sampling frame actively promoted Open Science practices. Commonly raised challenges for adopting Open Science practices among researchers include a lack of incentive and/or opportunity from the gatekeepers of scientific publishing such as journal and grant funding bodies (Allen & Mehler, 2018; Munafò et al., 2017). If Open Science practices have a steep learning curve, but are not part of the journal submission guidelines or if the journal researchers want to publish in does not accept registered reports, why would authors commit to them? Therefore, it is important for journals to encourage and facilitate the adoption of Open Science practices. Laudably, there have been recent developments in this area. In December 2017, the *IJPR* began to offer badges for publications demonstrating Open Science practices (Elk, Rowatt, & Streib, 2017). In June 2018, the *APR* began offering Open Science badges and accepting submissions of registered reports. Open Science publication badges are like scout patches and indicate that the article includes Open Science practices such as pre-registering the study's hypotheses and methodology prior to beginning data collection and/or openly sharing data and materials alongside the

published article. This is an attempt to incentivize Open Science practices as the badges signal more transparent and robust research practices.

Offering badges has been shown to improve the adoption of Open Science practices (Kidwell et al., 2016; Rowhani-Farid, Allen, & Barnett, 2017). Badges may offer journals' editorial boards a relatively simple and low-cost solution to improve submissions. Ultimately, publishing is important for the careers of scientists and journals are the gatekeepers of science that control what research is published. For randomised controlled trials, there is evidence that journals endorsing reporting guidelines have higher article quality (Turner, Shamseer, Altman, Schulz, & Moher, 2012). Consequently, if all psychology of religion journals endorse reporting guidelines which are designed for the social and behavioural sciences, such as the TOP guidelines (TOP Guidelines Committee, 2016; the *IJPR* is already a signatory), the reporting quality and evidentiary value of our research should increase. Hopefully, all journals in the psychology of religion can offer badges and registered reports, helping to lead the way in promoting Open Science practices. An editorial published, after our analyses had been performed, in the *APR* outlined a new vision for the journal with an emphasis on transparency and utilizing registered reports (Ladd, 2019).

Pre-registration. Regardless of journals adopting badges or registered reports, researchers can begin making RDF in research more transparent by pre-registering their studies. Over time, this will help pre-registration become a new norm in the psychology of religion. Pre-registering design features such as the primary outcome and sample size has been mandatory for clinical trials since 2005 in an effort to reduce publication bias and outcome switching (Viergever & Li, 2015). In addition, pre-registering hypotheses and analytical plans would create a clearer boundary between exploratory and confirmatory research. This distinction would help

reduce the over-inflation of conclusions made from exploratory results. It would also allow for more robust conclusions to be drawn for pre-registered confirmatory experiments, further increasing transparency and certainty in the authors' findings. During the replication crisis, pre-registration was proposed as an initiative for psychological research and has since seen a large uptake with over 8,000 pre-registrations being recorded on the OSF alone as of March 2018 (Nosek et al., 2018). Adopting this practice into the psychology of religion would help modernize the field, allowing both researchers across disciplines and the general public to put greater trust in the field's research.

Addressing concerns about pre-registration. We are aware that there may be reservations from some researchers who worry pre-registration may limit their research in some way or could be difficult to implement. We will briefly outline and address some of these concerns to provide a better understanding of the pre-registration process, which we hope will increase adoption of the practice.

One of the first concerns broached by researchers learning about pre-registration is that, by registering their analytical plan in advance, they are restricting their ability to explore issues they had not considered prior to data collection. While this concern is understandable given the appearance of rigidity that pre-registration brings with it, it is important to note that changes from a pre-registration are allowed and common (for a good example of how this can be done transparently, see Skinner et al., 2017). So long as both the modifications and justifications for them are described in the final paper, changes to the analytical plan or including additional research questions are well within the accepted scope of pre-registered articles. Importantly, this creates a clear boundary between exploratory and confirmatory research. Transparently

published work can contain a mix of confirmatory and exploratory analyses without authors' creativity being stunted.

A second common concern is that the pre-registration process itself may be difficult to learn and that researchers who do not have experience with pre-registration may not do it correctly. This concern can be broken into two parts: (1) learning and using new methodologies comes with an increased risk of making errors, which no researcher wants in their papers and (2) making these errors may cause authors more issues in getting published than not pre-registering their study at all. While we understand researchers' trepidation to adopt the practice, there are guides and primers on how to efficiently conduct a pre-registration (see, Simmons, Nelson, & Simonsohn, 2017; van't Veer & Giner-Sorolla, 2016; http://help.osf.io/m/registrations). Taking the pragmatic approach and gradually adopting more Open Science practices to incrementally increase the level of transparency in your research is still better than not adopting any at all (Klein et al., 2018; Nuijten, 2019). Use of any Open Science practice should be commended on making steps in the right direction, not condemned for not going further.

A third concern often raised by researchers is that pre-registration could waste valuable time better spent on research itself. It may be the case that the design phase of the study takes longer when it is carefully pre-registered (Allen & Mehler, 2018). However, we argue that the time taken to pre-register work ahead of data collection is a time-saving measure in the long run, not a time-cost. Given that, anecdotally, substantial delays between data collection and analysis in large projects are common, thorough documentation would likely help many in the field resume projects more swiftly. In particular, when an analysis plan has been pre-registered, researchers can conduct data analysis at a much faster rate relative to first needing to re-invent their original plan.

It is also important to highlight what pre-registration cannot solve. For example, individual authors can still use measures which have issues with validity or flexibly report their outcomes and procedures. One such issue can be highlighted by the pre-registration of this article. We pre-registered our hypothesis - that adoption of Open Science methods would be 'rare'. Using the term 'rare' without providing a specific rate or quantity is an example of providing a non-operationalized hypothesis. This cannot be addressed by pre-registration itself, but requires authors to clearly operationalise their hypotheses. Moreover, in clinical trials where study registrations have a longer history and are a formal requirement, outcome measures can still be switched or misrepresented (Goldacre et al., 2019). There is, ultimately, nothing stopping researchers from emphasizing a secondary outcome or dropping a condition. Thankfully, pre-registering the study makes this visible, whereas without pre-registration, this process may remain undetected.

Relatedly, a second limitation is the question of who takes on the burden of comparing the pre-registration protocol and the final paper. One answer is that the peer reviewers should appraise the consistency of the final paper with the pre-registration protocol when it is submitted for publication. While this has been raised as an additional burden on the voluntary services of the peer reviewer, we believe that, together with the journal editor, they are best-suited as gatekeepers to what is published. Editors and reviewers hold the power to ensure the article is published as faithfully to the pre-registration protocol as possible and that any deviations are transparently addressed. We benefited from this scrutiny, as one helpful reviewer closely compared our pre-registration protocol with the manuscript and highlighted issues that we addressed.

Registered reports. One solution to many of the above limitations is conducting and publishing a study following the registered report format. This way, the introduction and methods are submitted to the journal and peer-reviewed before data are collected. When the final article is submitted, the peer reviewers can identify if the study has been faithful to the stage-one submission. The study would only be published if it is consistent and the task of checking the consistency of the final article with the stage one submission is built into the publication process. This removes the problem of responsibility for self-archiving a pre-registration protocol. Nevertheless, there are still common misconceptions and perceived barriers, but these have been addressed elsewhere (see, Chambers et al., 2014). Pre-registration and/or registered reports would help to mitigate a large number of RDF.

Sample size justification. Justifying a sample size is an important practice that can be included in a pre-registration protocol. We found that 41 of 53 (77%) papers did not provide a justification for their sample size. Of the remaining 12, only two reported an *a priori* power analysis. Experiments with fewer participants than required for the predicted effect size are considered underpowered. Researchers often neglect justifying their sample size, which provides some insight into why contemporary research is commonly underpowered (Button et al., 2013). This is a major concern, because low statistical power means the results may not be replicable. Take the Open Science Collaboration (2015) as an example. They failed to replicate the original effects in over 60% of their replication studies. One explanation is that, because the original studies were underpowered, the replicating studies based their power analyses on inflated effect sizes. Inflated effect sizes cause power analysis calculations to be skewed, meaning they have less power to detect smaller, more realistic, effect sizes (Etz & Vanderkerckhove, 2016; Morey & Lakens, 2016). That pre-registered studies tend to report smaller effect sizes than non pre-

registered studies (Schäfer & Schwarz, 2019), suggests that non-registered studies may provide slightly inflated results. A key strength of science is that it is a cumulative process, thus it is critically important to perform both well-powered original and replication studies.

Specifically, our next recommendation is for future research to be designed with statistical power in mind. There are currently many tools available online (mostly free) that allow researchers to conduct power analyses. For example, the 'pwr' R package (Champely, 2018) is a free, open-source implementation to conduct power analyses in R. For those not familiar with the R coding language, G*Power (Faul et al., 2007) is a free application dedicated to providing sample size and power calculations for many research designs. In addition, www.powerandsamplesize.com (HyLown Consulting LLC, 2018) is a website where sample size and power can be calculated online. Alternatively, sequential analysis allows for interim analyses to be conducted with error corrections in place to prevent an increase in false-positive rate (Lakens, 2014). These tools make the design of well-powered experiments easy and efficient. As researchers use these tools more frequently, this should increase the number of well-powered studies, thus permitting further trust in psychological research on religion and increasing the transparency of RDF in the field.

Type I error correction. Having a well-powered study allows for researchers to be sure that they are likely to find an effect where there is one (to avoid a false negative). However, false positives should also be avoided. Whether or not a study uses a correction method to account for false positives is an often overlooked form of RDF. To help reduce these RDF, data analyses should include some form of Type I error correction, such as the Bonferroni correction (Bland & Altman, 1995), or False Discovery Rate correction (Benjamini & Hochberg, 1995). This also applies to factorial ANOVA where uncorrected multiplicity is often overlooked (Cramer et al.,

2016). In our review, 43 of 48 (90%) of the papers where Type I error correction could have been applicable did not state whether a correction was performed. While this issue affects other fields as well (see, Lachlan & Spence, 2005), the decision not to mention accounting for Type I error correction in any way in such a large proportion of papers demonstrates low levels of transparency within the psychology of religion. Consequently, we recommend that future researchers include whether or not their results are corrected for Type I error. A rationale for (not) applying a correction should also be provided.

Limitations

There are several limitations to our systematic review. As we coded for the presence or absence of RDF based on what was reported in the paper, we cannot definitively rule out that decisions such as sample size justification were outlined *a priori*. However, this means researchers can improve their reporting of important study design details by including this information. Secondly, we based our coding scheme on the pre-registration checklist by Wicherts et al. (2016) which is designed to be used throughout the process of a study. Our team selected items that we thought could be observable in published articles. In hindsight, our coding scheme could have been better. Some of our criteria were unfocused and open to interpretation, such as whether authors modified measures without theoretical justification or validation. This may have led to our low inter-rater reliability. As outlined in the pre-registration protocol, any disagreements were rectified by SJC and group consensus. Because SJC erred on the side of coding an ambiguous article as more transparent, these findings may also represent an optimistic estimate of the prevalence of RDF within psychology of religion. As demonstrated by our attempt at documenting RDF, they are an elusive assortment of problems that are difficult to identify and measure after studies have been conducted. The coding scheme could also have

been more comprehensive; we overlooked important RDF such as justifying outlier removal techniques and the specificity and directionality of hypotheses. Since most RDF will not be mentioned within the context of a research paper, it is unlikely that any study could show the "true" prevalence of RDF. Therefore, our results may underestimate the extent and diversity of RDF. Future studies could explore the use of RDF in the formative stages of research, then follow their use throughout the research process (also see, Latour & Woolgar, 1979/2013) to get a better grasp on when and where RDF occur that might not be reported in the context of published articles, or when authors think about their use of RDF retrospectively.

Conclusion

We propose three main recommendations for researchers that we believe will improve research within the psychology of religion: (1) research hypotheses and analysis plans should be pre-registered to help control RDF. This will help increase levels of transparency and reliability of psychology of religion research. (2) Sample sizes should be outlined in advance of data collection and justified in the manuscript to help reduce the number of underpowered studies and increase the evidential value of the findings. (3) Type I error correction should be conducted where applicable. Authors should discuss their reasons for not doing so to help improve the transparency and reliability of research. We also recommend that all academic journals within the psychology of religion work to encourage these Open Science practices in some way, be it via a 'badge' system or by offering the ability to publish registered reports. We are happy to see that some journals have started to move in this direction and we are optimistic that others will do the same. We believe that, by applying these recommendations along with continued self-reflection on better research practices, the psychology of religion will benefit from increased trust in the field's results and contribute to the ongoing reformation of science methodologies.

References

Allen, C. P. G., & Mehler, D. M. (2018). Open Science challenges, benefits and tips in early career and beyond. Unpublished manuscript, https://doi.org/10.31234/osf.io/3czyt

Ambasciano, L., & Coleman, T. J. III. (2019). History as a canceled problem? Hilbert lists, du Bois-Reymond's enigmas, and the scientific study of religion. *Journal for the American Academy of Religion*. doi: 10.1093/jaarel/lfz001

Asarnow, J., Bloch, M.H., Brandeis, D., Burt, S.A., Fearon, P., Fombonne, E., Green, J., … Zeanah, C.H. (2018). Special editorial: Open science and the journal of child psychology & psychiatry – next steps?, *Journal of Child Psychology and Psychiatry, 59* (7), 826–827.

Bains, S. (2011). Questioning the integrity of the John Templeton Foundation. *Evolutionary Psychology*, *9*(1), 147470491100900111.

Bajpai, S., Bajpai, R. C., & Chaturvedi, H. K. (2015). Evaluation of inter-rater agreement and inter-rater reliability for observational data: An overview of concepts and methods'. *Journal of the Indian Academy of Applied Psychology*, *41*(3), 20.

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., Hartgerink, C. H. J., Wicherts, J. M., & ver der Maas, H. L. J. (2016). Researchers' intuitions about power in psychological research. *Psychological Science*, *27*(8), 1–9.

Bender, R., & Lange, S. (2001). Adjusting for multiple testing - When and how? *Journal of Clinical Epidemiology*, 54(4), 343–349

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 289-300.

Bland, J. M., & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *BMJ*, *310*(6973), 170.

Brown, N. J. L., & Heathers, J. A. J. (2017). The GRIM Test: A Simple Technique Detects Numerous Anomalies in the Reporting of Results in Psychology. *Social Psychological and Personality Science*, 8(4), 363–369.

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

Camerer, C.F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., ... Chambers, C.D. (2013). Registered reports: A new publishing initiative at Cortex. *Cortex, 49*(3), 609–610.

Chambers, C.D., Feredoes, E., Muthukumaraswamy, S.D., & Etchells, P. (2014). Instead of "playing the game" it is time to change the rules: registered reports at AIMS neuroscience and beyond, *AIMS Neuroscience, 1*, 4–17

Champely, S., Ekstrom, C., Dalgaard, P., Gill, J., Weibelzahl, S., Anandkumar, A., Ford, C., ... De Rosario, M. H. (2018). Package 'pwr'.

Coleman, T. J. III., & Hood, R. W. Jr., (2015). Reconsidering everything: From folk categories to existential theory of mind. [Peer commentary on the paper "From Weird Experiences to Revelatory Events" by A. Taves]. *Religion and Society: Advances in Research*, 6 (1), 18–22.

Coleman, T. J. III., & Jong, J. (in press). Counting the nonreligious. In A. L. Ai, K. A. Harris, and P. Wink (Eds.) *Assessing spirituality and Religion in a Diversified World: Beyond the mainstream perspective*. New York: Springer

Cramer, A. O. J., van Ravenzwaaij, D., Matzke, D., Steingroever, H., Wetzels, R., Grasman, R. P. P. P., … Wagenmakers, E. J. (2016). Hidden multiplicity in exploratory multiway ANOVA: Prevalence and remedies. *Psychonomic Bulletin and Review*, 23(2), 640–647

de Groot, A.D. (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.

de Jager Meezenbroek, E., Garssen, B., van den Berg, M., Van Dierendonck, D., Visser, A., & Schaufeli, W. B. (2012). Measuring spirituality as a universal human experience: A review of spirituality questionnaires. *Journal of Religion and Health*, 51(2), 336-354.

Elk, M. van, Rowatt, W., & Streib, H. (2018). Good dog, bad dog: Introducing open science badges. *The International Journal for the Psychology of Religion*, 28(1), 1–2.

Etz, A., & Vandekerckhove, J. (2016). A Bayesian perspective on the reproducibility project: Psychology. *PLoS ONE*, *11*(2), 1–12.

Fanelli, D. (2010). "Positive" results increase down the hierarchy of the sciences. *PLoS ONE, 5*(4), 1-10

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: current practice and recommendations. *Social Psychological and Personality Science*, *8*(4), 370–378.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "*p*-hacking" and the research hypothesis was posited ahead of time. *Department of Statistics, Columbia University.*

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., … Mahtani, K. R. (2019). COMPare: a prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials*, *20*(1), 118.

Guo, Q., Thabane, L., Hall, G., McKinnon, M., Goeree, R., & Pullenayegum, E. (2014). A systematic review of the reporting of sample size calculations and corresponding data components in observational functional magnetic resonance imaging studies. *NeuroImage*, 86, 172–181.

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, *61*(1), 29–48.

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.

HyLown Consulting LLC. (2018). Power and Sample Size. Retrieved from http://www.powerandsamplesize.com

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.

Kappenman, E.S. & Keil, A. (2017). Introduction to the special issue on recentering science: Replication, robustness, and reproducibility in psychophysiology. *Psychophysiology, 54*(1), 3–5.

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

Klein, O., Hardwicke, T. E., Aust, F., Breuer, J., Danielsson, H., Mohr, A. H., … Frank, M. C. (2018). A Practical Guide for Transparency in Psychological Science. *Collabra: Psychology*, 4(1), 20.

Lachlan, K., & Spence, P. R. (2005). Corrections for type I error in social science research: A disconnect between theory and practice. *Journal of Modern Applied Statistical Methods*, 5(2), 490–494.

Ladd, K. L. (2019). The Archive for the Psychology of Religion: Editorial principles, practices, and practicalities, *Archive for the Psychology of Religion,* 41(1).

Ladd, K. L., & Messick, K. J. (2016). A brief history of the psychological study of the role(s) of religion. In W. Woody, R. Miller, & W. Wozniak (Eds.), *Psychological specialties in historical context: Enriching the classroom experience for teachers and students* (pp. 204-216). Division 2, American Psychological Association. Retrieved from https://teachpsych.org/ebooks/psychspec

Lakens, D. (2014). Performing high-powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, 44(7), 701–710.

Lane, A., Luminet, O., Nave, G., & Mikolajczak, M. (2016). Is there a publication bias in behavioural intranasal oxytocin research on humans? Opening the file drawer of one laboratory. *Journal of Neuroendocrinology, 28*(4), 1-15.

Larson, M. J., & Carbine, K. A. (2017). Sample size calculations in human electrophysiology (EEG and ERP) studies: A systematic review and recommendations for increased rigor. *International Journal of Psychophysiology*, *111*, 33–41.

Latour, B., & Woolgar, S. (1979/2013). *Laboratory life: The construction of scientific facts* (2nd ed.). Princeton: Princeton University Press.

Lo Martire, R. (2017). rel: Reliability Coefficients. R package version 1.3.1. https://CRAN.R-project.org/package=rel

Messick, K., & Farias, M. (in press). The psychology of leaving religion. *Brill Handbook of Leaving Religion*. Brill.

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... & Laitin, D. (2014). Promoting transparency in social science research. *Science, 343*(6166), 30-31.

Moberg, D. O. (2002). Assessing and measuring spirituality: Confronting dilemmas of universal and particular evaluative criteria. *Journal of Adult Development*, *9*(1), 47-60.

Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. *Retrieved from https://github. com/richarddmorey/psychology_ resolution/blob/master/paper/response. pdf.*

Fernández, L. M., & Vadillo, M. A. (2019). Reaction times: Many ways of inadvertently obtaining a false positive. *Retrieved from* https://doi.org/10.31219/osf.io/d4yqz

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

Nelson, L.D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*(17), 1–24.

Nosek, B.A. & Lakens, D. (2014). Registered reports. *Social Psychology, 45*(3), 137–141.

Nosek, B. A., & Lindsay, D. S. (2018). Preregistration becoming the norm in psychological science. *APS Observer*, *31*(30/3).

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, 201708274.

Nuijten, M. B. (2019). Practical tools and strategies for researchers to increase replicability. *Developmental Medicine & Child Neurology*, *61*(5), 535–539. https://doi.org/10.1111/dmcn.14054

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.

Olsson-Collentine, A., van Assen, M. A., & Hartgerink, C. H. (2019). The prevalence of marginally significant results in psychology over time. *Psychological Science*, 0956797619830326.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), 1–8.

R Core Team. (2017). R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing. [Computer software]. Retrieved from https://www.R-project.org/.

Roettger, T. B. (2018). Researcher degrees of freedom in phonetic sciences. Retrieved from www.psyarxiv.com/fp4jr

Rowhani-Farid, A., Allen, M., & Barnett, A. G. (2017). What incentives increase data sharing in health and medical research? A systematic review. *Research Integrity and Peer Review*, *2*(1), 4.

Rouse, S.V. (2018). Introduction to the special issue on open science practices: badges of honor. *Psi Chi Journal of Psychological Research, 23*(2), 94–97.

Schäfer, T., & Schwarz, M. A. (2019). The Meaningfulness of Effect Sizes in Psychological Research: Differences Between Sub-Disciplines and the Impact of Potential Biases. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.00813

Skinner, I. W., Hübscher, M., Moseley, G. L., Lee, H., Wand, B. M., Traeger, A. C., Gustin. S.M. & McAuley, J. H. (2017). The reliability of eyetracking to assess attentional bias to threatening words in healthy individuals. *Behavior Research Methods,* 50(5), 1778-1792.

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.

Simmons, J.P., Nelson, L.D., & Simonsohn, U. (2017, November 6). How To Properly Preregister A Study [blog post]. Retrieved from http://datacolada.org/64

Slater, W., Hall, T. W., & Edwards, K. J. (2001). Measuring religion and spirituality: Where are we and where are we going?. *Journal of Psychology and Theology*, *29*(1), 4-21.

Tamminen, K.A. & Poucher, Z.A. (2018). Open science in sport and exercise psychology: review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise* 36, 17–28.

Turner, L., Shamseer, L., Altman, D. G., Schulz, K. F., & Moher, D. (2012). Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. *Systematic Reviews*, *1*(1), 60.

Uzdavines, A., Hill, P., Coleman, T.J. III, Gibson, N. J., & Stauner, N. (2017). Open science ideals, practices, and dissemination within the international psychology of religion community. Presentation conducted at the 2017 International Association for the Psychology of Religion World Congress, Hamar, Norway.

van't Veer, A. E., & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2-12.

Viergever, R. F., & Li, K. (2015). Trends in global clinical trial registration: An analysis of numbers of registered clinical trials in different parts of the world from 2004 to 2013. *BMJ Open*, 5(9), e008932

Wagenmakers, E. J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams Jr, R. B., ... & Bulnes, L. C. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science*, *11*(6), 917-928.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632-638.

Wongpakaran, N., Wongpakaran, T., Wedding, D., & Gwet, K. L. (2013). A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: A study conducted with personality disorder samples. *BMC Medical Research Methodology*, 13(61), 1-7.

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, 1-12*.

Wulff, D. (1998). Rethinking the rise and fall of the psychology of religion. In A. Molendijk & P. Peels (Eds.), *Religion in the making: The emergence of the sciences of religion* (pp. 181-202). Leiden, Netherlands: Brill.

Wulff, D. (2003). A field in crisis: Is it time for the psychology of religion to start over?. In P. Roelofsma, J. Corveleyn & J. Van Saane (Eds.), *One hundred years of psychology and religion: Issues and trends in a century long quest* (pp. 11-32). Amsterdam: VU University Press.

Table 1

Inter-Rater Reliability using Gwet's AC1 for types of RDF

Coding Category	*AC1*	Lower	Upper
Pre-Registered	1*	1	1
RS Defined	.483	.293	.674
Clarity of RS Definition	.371	.182	.560
RS Definition Matched Measure	.152	.000**	.336
RS Measure Rationale	.323	.147	.498
Measures Modified	.000**	.000**	.139
Exploratory or Confirmatory	.427	.246	.609

Sample Size Justification	.326	.033	.619
Type I Correction	.897*	.806	.988
Exact *p*-values	.367	.152	.581
Reporting of p >0.05 as meaningful	.843*	.730	.956
Over-Inflation of Results	.836*	.725	.948
Cut-Off Explained	.533	.368	.697
Non-RS Definition Consistency	.341	.137	.545
Unclear Coding Procedures	.462	.288	.635

Confidence intervals provided at the 95% level
* Indicates that there was a high level of agreement
** Indicates that the values appeared below zero

Table 2

Codes and Final Percentages After Reconciliation

Coding Categories	*Codes (Number, Percentage of Papers)*
Pre-Registered	*Yes (0, 0%); No (53, 100%)*
RS Defined	*Yes (45, 85%); No (7, 13%); NA (1, 2%)*
Clarity of RS Definition	*Clear (39, 73%); Vague (6, 11%); Mixed (3, 6%); No Definition Provided (4, 8%), NA (1, 2%)*

RS Definition Matched Measure	*Yes, All (40, 75%); Mixed (10, 19%); No, All (2 4%); NA (1, 2%)*
RS Measure Rationale	*Yes (50, 94%); No (3, 6%)*
Measures Modified	*Yes, All (3, 6%); Yes, Some (14, 26%); No Changes (32, 60%); Created Measure (4, 8%)*
Exploratory or Confirmatory	*Exploratory (19, 36%); Confirmatory (25, 47%); Both (9, 17%)*
Sample Size Justification	*Yes (12, 23%); No (41, 77%)*
Type I Correction	*Yes (4, 8%); Did not mention (43, 81%); No, mentioned (1, 2%); Not applicable (5, 9%)*
Exact p-values	*Throughout (16, 30%); Varied (16, 30%); Did not provide (20, 38%); NA (1, 2%)*
Reporting of p >0.05 as meaningful	*Yes (9, 17%); No (44, 83%)*
Over-Inflation of Results	*Yes (5, 9%); No (48, 91%)*