FLORIDA STATE UNIVERSITY
COLLEGE OF SOCIAL WORK

APPRAISING CONTEMPORARY SOCIAL WORK RESEARCH:
META-RESEARCH ON STATISTICAL REPORTING, STATISTICAL POWER, AND
EVIDENTIAL VALUE

By
DANIEL JOSEPH DUNLEAVY

A Dissertation submitted to the
College of Social Work
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

2020

Daniel Joseph Dunleavy defended this dissertation on June 16, 2020.

The members of the supervisory committee were:

Jeffrey R. Lacasse

Professor Directing Dissertation

Lenore M. McWey

University Representative

Tomi Gomory

Committee Member

Lisa Schelbe

Committee Member

James J. Clark

Committee Member

The Graduate School has verified and approved the above-named committee members, and certifies that the dissertation has been approved in accordance with university requirements.

For Mary and Lawrence Buckley

ACKNOWLEDGMENTS

This dissertation could not have been completed without the help of my friends and family.

Thank you to Mom and Dad and the rest of my family. Not only have you helped me earn this degree, but you have unconditionally supported and encouraged each of my goals along the way. This document (hopefully) serves to answer the eternal question, “So when *are* you going to graduate, anyway?” ;)

Katie, Makenna, and Melissa, I will always appreciate our countless hours of conversation, meals, aimless walks around town, and coffeeshop work sessions. Your feedback inside and outside the classroom has made me a better scholar and person. I am lucky to have had you all along for the ride and to have gained you as three of my best friends.

Megan, Lane, Aidan, Sandi, and Matt, thank you for your constant encouragement, feedback, empathy, and friendship. I can’t wait to see what great things each of you do.

Many thanks to the late Dr. Jon Tennant. I will miss our collaborations, but most of all your unbridled enthusiasm for open science, your intellectual depth and breadth, and general passion for life. The world lost you too soon.

I am also fortunate to have had an overwhelmingly supportive committee. Thank you to Dr. Lisa Schelbe for going above and beyond at each step of the way. Not only have you been an academic role model, but you’ve provided consistently thoughtful feedback and support. Thank you to Dr. James Clark for your constant encouragement of my scholarship and for always challenging me to do my best work. Thank you to Dr. Lenore McWey, for providing thoughtful feedback throughout the dissertation process; it has undoubtedly helped improve it at each step. Drs. Tomi Gomory and Jeffrey Lacasse deserve special mention. In your own ways you have allowed me the intellectual space to explore and grow as a person and scholar. I will cherish our frequent conversations and debates, your mentorship, and friendship.

Thank you to Drs. Stephen Tripodi, David Cohen, Craig Stanley, Jean Munn, Neil Abell, Dina Wilke, Karen Randolph, James Justus, and Mrs. Fran Gomory for help and guidance along the way. A big thank you as well to all the staff at the College of Social Work, but especially Jennifer Farinella, Alli Dillon, Cassandra Brown, Angela Shrum, Lauren Antista, and Emily Joyce.

Each of my students deserve special mention as well. Thank you for your hard work, patience, curiosity, and for helping me improve as an instructor and scholar.

Above all, thank you to Sandy (and Maddie) for your unconditional love and support. Words won’t do it justice but thank you for all that you’ve done and for the sacrifices you’ve had to make. None of this would have been possible without you. I love you both with all of my heart.

TABLE OF CONTENTS

List of Tables	vii
List of Figures	viii
Abstract	ix
1. INTRODUCTION	1
The Problem: Science in Crisis	1
Is Social Work in a Replication Crisis? Does it Matter?	5
Solutions to the Replication Crisis	6
Contemporary Meta-Research	9
Meta-Research and Social Work	13
Meta-Research and the History of Social Work Research	14
The Benefits of Social Work Meta-Research	15
The Plan Going Forward	15
2. STATISTICAL BACKGROUND: THEORY AND CONCEPTS	16
Schools of Statistical Inference	17
The Frequentist School	17
Alpha (α)	23
Beta and Statistical Power	25
P-Values and Statistical Significance	29
Effect Size	31
Confidence Intervals	31
NHST and Limitations of Frequentist Statistics	33
Alternative Approaches	35
Conclusion	37
3. METHODOLOGY	39
Purpose and Outline of the Study	39
Methodology and Study Design	41
Storage and Availability of Data and Materials	44
Sampling Plan and Sample	44
Inclusion and Exclusion Criteria for Articles	45
Research Questions	46
Research Hypotheses	48
Data Extraction, Analysis, and Reporting	53
Conflicts of Interest	57

4. RESULTS	58
Results from Part 1: Statistical Reporting.....	60
Results from Part 2: Power Survey	64
Results from Part 3: P-Curve	66
5. DISCUSSION	72
Summary and Interpretation of Findings	72
Limitations of the Study.....	77
Implications for Social Work Research	83
Future Directions: Social Work Meta-Research and Statistical Reform	84
APPENDIX – FSU IRB EXEMPTION	85
References	86
Biographical Sketch	110

LIST OF TABLES

1	Four Possible Statistical Outcomes	23
2	Selective Review of Power Surveys (Sorted by Year)	54
3	Sampling Frame by Journal Frequency	58
4	Reason for Exclusion by Type and Frequency	59
5	Journal Frequencies	61
6	Frequencies for Variables	62
7	Frequencies of Explicit Alpha Level	63
8	Frequencies of Implicit Alpha Level	63
9	Frequencies of Type of Power Analysis	63
10	Frequencies of Type of Confidence Interval	64
11	Survey of Power from Studies Published Across Five Social Work Journals Between 2014-2018	67
12	Robustness Test of Power from Studies Published Across Five Social Work Journals Between 2014-2018	68

LIST OF FIGURES

1	Distribution of P-Values	24
2	50% Statistical Power, Simulated 100,000x	28
3	Power for Each Effect Size by Journal	69
4	P-Curve (Primary)	70
5	P-Curve (Robustness Test)	71

ABSTRACT

Background. In recent years, the veracity of scientific findings has come under intense scrutiny in what has been called the “replication crisis” (sometimes called the “reproducibility crisis” or “crisis of confidence”). This crisis is marked by the propagation of scientific claims which were subsequently contested, found to be exaggerated, or deemed false. The causes of this crisis are many, but include poor research design, inappropriate statistical analysis, and the manipulation of study results. Though it is uncertain if social work is in the midst of a similar crisis, it is not unlikely, given parallels between the field and adjacent disciplines in crisis.

Objective. This dissertation aims to articulate these problems, as well as foundational issues in statistical theory, in order to scrutinize statistical practice in social work research. In doing so, it parallels recent work in psychology, neuroscience, medicine, ecology, and other scientific disciplines, while introducing a new program of meta-research to the social work profession.

Method. Five leading social work journals were analyzed across a five-year period (2014-2018). In all 1,906 articles were reviewed, with 310 meeting inclusion criteria. The study was divided into three complementary parts. Statistical reporting practices were coded and analyzed in Part 1 of the study ($n = 310$). Using reported sample sizes from these articles, a power survey was performed, in Part 2, for small, medium, and large effect sizes ($n = 207$). A novel statistical tool, the p-curve, was used in Part 3 to evaluate the evidential value of results from one journal (*Research on Social Work Practice*) and to assess for bias. Results from 39 of the 78 eligible articles were included in the analysis. Data and materials are available at: <https://osf.io/45z3h/>

Results. Part 1: Notably, 86.1% of articles reviewed did not report an explicit alpha level. A power analysis was performed in only 7.4% of articles. Use of p-values was common, being reported in 96.8% of articles, but only 29% of articles reported them in exact form. Only 36.5% of articles reported confidence intervals; with the 95% coverage rate being the most common (reported in 31.3% of all studies). Effect sizes were explicitly reported in the results section or tables in a little more than half of articles (55.2%). Part 2: The mean statistical power for articles was 57% for small effects, 88% for medium effects, and 95% for large effects. 61% of studies did not have adequate power (.80) to detect a small effect, 19% did not have adequate power to detect a medium effect, and 7% a large effect. A robustness test yielded similar but more conservative estimates for these findings. Part 3: Both the primary p-curve and robustness test yielded right-skewed curves, indicating evidential value for the included set of results, and no evidence of bias.

Conclusion. Overall, these findings provide a snapshot of the status of contemporary social work research. The results are preliminary but indicate areas where statistical design and reporting can be improved in published research. The results of the power survey suggest that the field has increased mean statistical power compared to prior decades; though these findings are tentative and have numerous limitations. The results of the p-curve demonstrate its potential as a tool for investigating bias within published research; while suggesting that the results included from *Research on Social Work Practice* have evidential value. In all this study provides a first step towards a broader and more comprehensive assessment of the field.

CHAPTER 1

INTRODUCTION

The Problem: Science in Crisis

In recent years, the veracity of scientific findings has come under intense scrutiny in what has been called the “replication crisis” (sometimes called the “reproducibility crisis” or “crisis of confidence”). This crisis is marked by increased skepticism by researchers and the public (Baker, 2016; National Academy of Sciences, Engineering, and Medicine, 2016, 2019) towards the reliability and reproducibility of results in the biomedical and social sciences. For instance, in 2005 John Ioannidis famously argued that most published research findings are false and demonstrated how numerous biases throughout the research process can drastically increase the likelihood of false positive claims (Ioannidis, 2005, pp. 0699-0670). Similar arguments have been put forth within the field of psychology, which experienced a series of “embarrassing” incidents between 2011 and 2012 (Pashler and Wagenmakers, 2012, p. 528), including data fabrication by an award winning psychologist (Stroebe, Postmes, & Spears, 2012) and the publication of a study claiming to have found evidence for extrasensory perception in a top-ranked psychology journal (i.e. Bem, 2011); among other issues (see Chambers, 2017 for an overview of these events). Still further, Simmons, Nelson, and Simonsohn (2011) demonstrated the multitude of ways (and ease) one can obtain statistically “significant” results in the presence of a null effect, while large teams of investigators have questioned the extent to which findings across the sciences can be independently replicated (Camerer et al., 2018; Open Science Collaboration, 2015). In all, the faith, utility, and reliability of much scientific research has and continues to be brought into question.

Causes of the crisis. Researchers have put forth numerous hypotheses about the causes of the reproducibility crisis (see Ioannidis, 2005, 2008, 2012); none of which completely explain the problem. These include: 1) low statistical power (Button et al., 2013; Cohen, 1992; Sham & Purcell, 2014), 2) impoverished theory (Mischel, 2008; Muthukrishna & Henrich, 2019), 3) reliance on exploratory, rather than confirmatory analyses of data (de Groot, 1969, 1969/2014; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011; Wagenmakers, Wetzels, Borsboom, Kievit, & van der Maas, 2011; Wagenmakers et al., 2012), 4) the idiosyncratic nature and flexibility of data analysis (Gelman & Loken, 2013; Silberzahn et al., 2018; Simmons et al., 2011), 5) reliance and misuse of frequentist statistics (Colling & Szűcs, 2018; Dienes, 2016; Meehl, 1976, 1978; Wagenmakers et al., 2011), 6) engagement in Questionable Research Practices (QRPs)¹ (e.g. Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; Kerr, 1998; Murphy & Aguinis, 2019; Wansink, 2016), 7) publication bias (Ioannidis, Munafò, Fusar-Poli, Nosek, David, 2015; Rosenthal, 1979), and 8) research fraud (e.g. data fabrication).^{2,3} As can be seen, these causes cut across all stages of the research process, from study design (e.g. theoretical grounding; statistical power) to data analysis (e.g. exploratory analysis; use of frequentist statistics to test hypotheses) and the reporting of results (e.g. QRPs, publication bias). Individually and collectively these factors (which I'll henceforth refer to as the “causes of

---

¹ QRPs include, but are not limited to “p-hacking” (e.g. collecting data until results are statistically significant), “HARKing” or hypothesizing after results are known, and selective reporting or “cherry picking” of statistically significant results (see Fraser et al., 2018; Kerr, 1998).

² Though outright fraud is seemingly a rare event (Simohanson, 2013).

³ Other factors include incentivization of poor research practices by academic institutions and funders (Chambers, 2017; Edwards & Roy, 2017) and poor measurement of constructs (Fried & Flake, 2018).

crisis”) undermine the reliability of findings and validity of scientific inferences⁴ and negatively impact socio-professional behavior.⁵

Epistemic consequences. Whatever the true nature and characterization of scientific progress (Campbell, 1987; Gomory, 1997; Kuhn, 1962, 1970; Ioannidis, 2012; Lakatos, 1968 Popper, 1959/1968, 1963, 1970, Reichenbach, 1938), the causes of crisis undermine the testing of hypotheses and the validity of inferences derived from those tests. For instance, low statistical power impedes the discovery of underlying effects (when present) and increases the probability that a significant result is a false positive (Button et al., 2013; Wagenmakers et al., 2015).⁶ Publication bias, selective reporting, and other QRPs may prevent publication of outcomes that would otherwise falsify a statistical hypothesis from entering the published literature (see generally Morey & Lakens, 2016; Popper, 1963) and undercut that essence of approaches like Evidence-Based Medicine/Evidence-Based Practice (EBM/EBP; Ioannidis, 2016), which require grounding in the evaluation of *all* results (i.e. the general scientific and philosophical norm known as the “principle of total evidence”; see Stegenga, 2018). The philosopher of science Deborah Mayo (2018) characterizes tests afflicted by QRPs as producing *BENT* or “bad evidence, no test”. A test that yields positive results in the presence of QRPs is not much of a *test* at all and the purported “evidence” supporting it is thus “lousy”, having little weight. The hypothesis under scrutiny is then, “essentially immune to having its flaws uncovered by the data”

---

⁴ Here and throughout I am primarily talking about quantitative research. However, many of the same issues apply, *in principle*, to qualitative research. For a discussion, see Aguinis and Solarino (2019).

⁵ Bird (2018) argues that failure to replicate results is consistent with good scientific practice and is a consequence of hypotheses having low prior probability of being correct to begin with. While I don’t disagree that considering “the base rate of falsity” is an important contextual aspect of the debate around replication, all things being equal, I believe Bird *underestimates* the impact of Causes of Crisis on scientific progress; particularly as it pertains to the testing and falsification of hypothesis (see also McElreath & Smaldino, 2015; and the next section).

⁶ Button et al. (2013) also note that if low statistical power is widespread, the published literature is more likely to overestimate or exaggerate the size of an effect; with researchers only being able to reliably detect large effects (p. 365).

(p. 5). When causes of crisis are widespread, research is not scientific at all, but merely a form of posturing (Tripodi, 1974, pp. 143-144), leading to programmes of research that have the guises of scientific inquiry, but are epistemologically fruitless or “degenerative” in Lakatosian sense (1968).

Socio-professional consequences. The unreliability of scientific research has several “real world” consequences for researchers and the public, including, perhaps most notably, diminishment of public trust in the scientific enterprise. Proclamations that “most published research findings are false” (Ioannidis, 2005) can encourage public skepticism toward all scientific claims (e.g. Edwards & Roy, 2017; Vazire, 2017); regardless of their inherent truth or falsity. This leaves experts in disrepute and the lay-public ill-equipped to adjudicate the accuracy of any given scientific claim. This problem is further exacerbated by the current socio-political and technological climate (O’Connor & Weatherall, 2020), which facilitates the spread of misinformation and enables the development of conspiracy theories (Hendricks & Vestergaard, 2018). If public trust in science is sufficiently eroded, scientific expertise becomes enfeebled. Goldenberg (2019) makes this clear when discussing distrust of experts within the context of vaccine guidelines, stating: “If nonexperts find themselves unsure of the integrity of expert advice, even the sheer number of signees to the scientific consensus will do little to convince.” (p. e398). Given current skepticism, researchers will need to both repair the internal workings of their own scientific practice and public trust in that work.⁷

Other consequences include wasted resources (e.g. time and money of funders, researchers, participants) and the increased likelihood of promoting ineffective or harmful policies and interventions. Consider the case of Study 329, an industry-sponsored clinical trial,

---

⁷ A point made all the more salient given the current COVID-19 pandemic (O’Connor & Weatherall, 2020).

testing the efficacy and safety of paroxetine (Paxil) (and imipramine) vs. placebo in adolescents with depression. Funded by SmithKline Beecham (what is now GlaxoSmithKline), the study reported that paroxetine was a “generally well tolerated and effective for major depression in adolescents” (Keller et al., 2001, p. 762). Other concerns aside (e.g. ghostwriting, conflicts of interest; Lacasse & Leo, 2010), the study is marked by serious charges of selective reporting. Jureidini, McHenry, and Mansfield (2008) argue that the study authors minimized two of the pre-specified primary outcomes, which did not favor paroxetine, and instead promoted secondary outcomes which did (p. 76). A re-analysis of the trial data by LeNoury et al. (2015), contra Keller et al. (2001), found neither paroxetine nor imipramine were effective treatments for depression, and both had clinically significant increases in harms (i.e. suicidal ideation). This is but one example that shows how, when outcomes are selectively or incorrectly reported (or are misleading) (see Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Gøtzsche, 1990), informed decisions about treatment choice and policy evaluation cannot be made.

Study 329 may seem an outlier, but selective reporting (and other biases) are not uncommon (e.g. Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2004; Gøtzsche, 1990). Nor are the consequences of less “deliberate” actions any less problematic. Stegenga (2016) notes how low statistical power not only can lead researchers to miss important treatment benefits, but also the detection of harms; which may be rare, but severe (Cai, Parast, & Ryan, 2010). Consequently, it may only be after many years (and additional studies/reports) that an ineffective or harmful intervention is stopped or removed (see broadly Herrera-Perez et al., 2019).

Is Social Work in a Replication Crisis? Does it Matter?

We do not know if social work is in the midst of a replication crisis (Howard & Garland, 2015; Yaffe, 2019). The extent to which social workers engage in QRPs, the power of social

work studies, publication bias, etc. has not been systematically investigated. However, neither is there reason to believe that contemporary social work research is performed differently than that in adjacent fields (e.g. psychology, criminology, economics, etc.); especially given historical misuse of research within the profession (see Tripodi, 1974). Likewise, the profession draws many of its research practices from the broader social sciences and largely exists within the same academic incentive structure (Ioannidis, 2012; Nosek, Spies, Motyl, 2012; Zollman, 2019). Therefore, the probability that the profession is in a crisis is nontrivial; which is particularly unsettling, given that most social workers are seemingly unaware of any crisis at all (Yaffe, 2019). In sum, it is plausible that many claims within social work are overstated at best, and at worse false. It is important that the field take precautionary measures (Howard & Garland, 2015), even if not in crisis, to ensure that research is being performed as efficiently and optimally as possible. I turn next to some potential solutions to the replication crisis, focusing on the use of meta-research, which motivates the present dissertation study.

Solutions to the Replication Crisis

Many small and large-scale solutions have been proposed to solve the replication crisis and to rectify the underlying causes of crisis (see broadly Chambers, 2017; Nosek, Spies, & Motyl, 2012). One of the most comprehensive developments is the concept of a Registered Report. Registered Reports (RR) (Chambers, 2013; Chambers, Dienes, McIntosh, Roshtein, Willmes, 2015; Simons, Holcombe, & Spellman, 2014), implemented at the journal-level, require submission of study methods and analytic strategy prior to data collection.⁸ After undergoing peer review, the manuscript may gain “in principle acceptance” for publication at the

---

⁸ I highlight this solution here, because it directly and indirectly affects all the causes outlined earlier (except for, perhaps, the second cause).

journal where it is reviewed. Studies that are accepted as Registered Reports (e.g. Hall et al., 2018), are done so on the merits of methodological rigor, rather than its findings. A study submitted as a Registered Report would undergo a second round of peer review after its results have being written-up, to verify that all facets of the proposed analytic strategy are included. Through the creation of RRs, there is incentive for researchers to follow their pre-specified protocol (i.e. to maintain in principle acceptance) and disincentive to engage in QRPs and other methodological “sins” (Chambers, 2017; Chambers et al. 2015).

Others specific recommendations include, 1) replacement of frequentist statistics with Bayesian statistics (discussed in some detail in the next chapter) (Dienes & Mclatchie, 2018), 2) pre-registration of study protocols and 3) encouragement of confirmatory (rather than exploratory) analyses (Wagenmakers et al., 2012), 4) sharing of data, code/syntax, codebooks, and other materials in public repositories (Lindsay, 2018), 5) large-scale, multi-site collaborations (to among other things, increase statistical power in rare or hard to reach populations) (e.g. Moshontz et al., 2018), 6) the encouragement of independent replication studies (McNeely & Warner, 2015; Nosek & Errington, 2019), 7) meta-research (Ioannidis, Fanelli, Dunne, & Goodman, 2015; Nuijten, 2018), and 8) educational reform (e.g. Dunleavy & Lacasse, 2020). Each of these solutions has been implemented across the sciences to some extent (Chambers, 2017; Spellman, 2015).

What is meta-research? Meta-research is a broad and burgeoning area of study, which concerns the methods, reporting, and reproducibility of research (Ioannidis et al., 2015). Simply put meta-research is “research on research” (Nuijten, 2018). Meta-research can focus on any aspect of the research process; from study planning/design through dissemination/post-publication, as well as the context in which such research is produced (e.g. the broader academic

system). It is a largely a descriptive area of study, seeking to uncover how scientific research is practiced; as well as a prescriptive one, in that it is concerned with identifying and rectifying factors contributing to suboptimal research practices (see Hardwicke et al., 2019; especially section 1.3).

How meta-research fits into scientific practice. Meta-research and the broader “open science” movement (see Spellman, Gilbert, & Corker, 2018; and generally Dunleavy, 2020; Tennant et al., 2020a, 2020b), encourages the identification and correction of suboptimal research practices and the promotion of rigorous (i.e. severe) testing of hypotheses (Mayo, 2018). These efforts can best be viewed as a form of continuous calibration of the overarching scientific ecosystem (its methods, dissemination of results, vetting of claims, etc.) (Hardwicke et al., 2019). By continuously appraising its scientific behavior, a field of research should arrive at more rigorous methods and a more robust, yet still fallible (Popper, 1963), body of evidence (i.e. a progressive programme of research in Lakatos’ terms, 1968).

Further, meta-research and broader reform efforts impact the sociology of science (i.e. the conduct of researchers and other stakeholders within the scientific ecosystem). The Mertonian norms of universalism, communism, disinterestedness, and skepticism (Merton, 1973) are promoted and encouraged by most meta-researchers (e.g. Hardwicke et al., 2019). Research is understood to be a collective effort, whereby the public scrutiny of published work is understood to be the *sine qua non* of scientific progress (Chartier et al., 2018; Vazire, 2017). Indeed, as Bruner (2013) argues, epistemic communities, such as the field of social work, function most efficiently (i.e. increased likelihood of honest reporting; minimization of data fabrication and fraud) when peers *actively* “police” the community of researchers; rather than the passive, secretive, closed model found in the current scholarly publication system (Dunleavy,

Tennant, & Francuzik, 2020). In other words, meta-research “act[s] as a check” (Barnes, TenEyck, Pratt, & Cullen, 2018, p. 8) on a field’s overall functioning.

Contemporary Meta-Research

I turn next to a selective sampling of meta-research studies from across the biomedical and social sciences, highlighting the fields of mental health, neuroscience, psychology, economics, and criminology. Doing so will show what meta-research looks like in practice and demonstrates its utility as a feedback mechanism in each field/area of research; thus motivating its use in the current study.

Mental health. *Key study #1: Turner, Matthews, Linardatos, Tell, & Rosenthal (2008).* *Area: Publication/reporting bias.* Turner et al. (2008) obtained all reviews (n = 74 studies) submitted to the FDA for 12 antidepressant drugs (selective serotonin reuptake inhibitors or SSRIs). They found that 26 studies (31%) were unpublished. Of studies that were found to have *positive* results, all but one was published in a journal (p. 254). In other words, positive results were overrepresented in the published literature, while negative results were greatly underrepresented. The analysis by Turner et al. extends beyond these figures, but as can be seen, the study serves as an important indicator for publication bias (whatever its causes) and has implications for mental healthcare policy, our understanding of the mental health treatment literature, and its synthesis.

Key study #2: Border et al. (2019). Area: Replication of prior findings. Recently, Border and colleagues challenged the validity of an entire program of research in behavioral genetics research. It is commonly believed that Major Depressive Disorder (MDD) has an underlying genetic cause(s). The 5-HTTLPR genotype, for example, has been hypothesized to increase the likelihood of a carrier having MDD or a depressive episode (e.g. Caspi et al., 2003). Increased

risk may be attributed to the genes themselves, the interactions of multiple genes, or gene-environment interaction. Using large, population-based samples, the authors performed pre-registered analyses of 18 depression candidate genes (each of which have been studied ten or more times by genetics researchers), which accounted for hypothesized interactions between gene(s) and environment as well as for potential moderating variables (e.g. childhood trauma; socioeconomic deprivation). The results of this replication did not support previous findings (p. 1) and raised doubt about the veracity of previous findings (e.g. attributing positive results to low statistical power, measurement error, sampling variability, etc., p. 6). Border et al. (2019) do not mince words, concluding:

“Our results mirror those of well-powered investigations of candidate gene hypotheses for other complex traits, including those of schizophrenia... In agreement with the recent recommendations of the National Institute of Mental Health Council Workgroup on Genomics...we conclude that it is time for depression research to *abandon* historical candidate gene and candidate gene-by-environment interaction hypotheses.” (p. 11; emphasis added).

While the analyses employed and conclusions drawn by the authors are certainly subject to debate, the study will unquestionably challenge how we think about psychiatric genetics research. Researchers will be prompted to reflect on why prior depression gene studies which had positive results, failed to replicate here, and could push researchers towards other areas of inquiry.

Neuroscience. *Key study: Button et al. (2013). Area: Study design (i.e. statistical power).*

Button et al. investigated the power of neuroscience research, drawing on the data from 49 meta-analyses. There was an obvious bimodal distribution of power among studies with median power of 21%; and mean power between 8% and 31% (p. 371). The authors highlight the relationship between low power and the replicability of neuroscience findings, as well as the ethical implications for humans and animals in performing studies that have limited scientific value (p.

372). Though optimistic about the future, they concluded that: “Small, low-powered studies are endemic in neuroscience” and “There is now substantial evidence that a large proportion of the evidence reported in the scientific literature may be unreliable.” (p. 374). This study has caused a re-appraisal of previous results within neuroscience and prompted researchers to think more explicitly about statistical power and other design considerations when planning and conducting their studies (de Haas, 2018).

Psychology. *Key study: Open Science Collaboration (2015). Area: Replication of prior findings.* In what may be the most well-known and ambitious meta-research study, the Open Science Collaboration (OSC) conducted replications of 100 studies from three psychology journals. These replications were performed using original study materials and (where possible) with the cooperation or advisement of the original study authors. The OSC evaluated reproducibility based on five metrics (statistical significance, p-values, effect sizes, meta-analysis of combined original and replication effects, and the subjective assessment of replication teams) (pp. 2-3). The OSC found that replicated effects were typically half the magnitude of those found in the original studies and that of the replication studies, only 39% were rated as having verified the results of their original counterparts (p. 943; see especially Table 1).

Economics. *Key study: McCloskey and Ziliak (1996). Area: Study design/statistical reporting.* In many ways McCloskey and Ziliak serve as an exemplar for the current study. They surveyed all, full-length papers using a regression analysis ($n = 182$) in the *American Economic Review*, published during the 1980s. McCloskey and Ziliak asked 19 yes/no questions of each paper, related to the reporting of study design and statistical results. For instance, they asked: if a differentiation was made between statistical and substantive (i.e. real world) significance?; if the

authors considered the power of the test?; if “asterisk economics”⁹ was eschewed?; and so on (p. 110). Notably, they found that 70% of papers failed to distinguish between statistical and practical significance, 69% did not report descriptive statistics, and only 4% considered the power of their statistical tests (p. 106). The study has been cited more than 900 times (Google Scholar), serving as an important example about statistical reporting and as a call for reform.

Criminology. In the last couple of decades, the field of criminology has shown increased concern around the issues of replicability, study design, and reporting (e.g. Pridemore, Makel, & Plucker, 2018). The following studies reflect this concern.

Key study #1: McNeeley & Warner (2015). Area: Replication (i.e. frequency of replication studies). McNeeley & Warner recently surveyed studies published between 2006 and 2010 in criminology and criminal justice journals; finding that only about 2% can be classified as replications; and only 6 of 691 studies (< 1%) being classified as direct replications. They argue that given these unimpressive findings and given the centrality of replication for knowledge development, criminology will need to foster a culture or tradition of replication across the field (e.g. among researchers, journals, education in graduate programs) (pp. 13-14).

Key study #2: Barnes et al. (2019). Area: Study design (i.e. statistical power). Barnes et al. estimated statistical power, using sample sizes and mean effect sizes from 81 criminology meta-analyses. They found that studies had mean power of 0.605 (calculated on the *observed* effect size) and estimate that about half of criminology studies are underpowered (p. 15).

Key study #3: Bushway, Sweeten, & Wilson (2006). Area: Study design/statistical reporting. Bushway et al. modified the questions used by McCloskey and Ziliak (1996) and

---

⁹ “Asterisk economics” is the practice of employing asterisks or “stars” next to p-values, to designate different “degrees” or “levels” of significance. This practice is criticized in the next chapter of the present work.

applied them to articles in three criminology journals ($n = 82$). Similarly, Bushway et al. (2006) found that 60% of papers failed to distinguish between statistical and practical significance, 35% did no report descriptive statistics for all variables reported, and only 30.5% considered the power of their chosen statistical tests.

Summary. This has only been a sampling of meta-research studies performed within the above fields and does not take into account the broader efforts and discussions occurring in fields like nutrition (Ioannidis, 2013, 2018), sociology (Abbott, 2007; Freese, 2007), and ecology (Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; Kelly, 2006). Together, these examples demonstrate the utility of meta-research for assessing the scientific practices of a field. McCloskey & Ziliak (1996) and Bushway et al. (2006) serve as exemplars for the first analysis performed in this study. Button et al. (2013) and Barnes et al. (2019) demonstrate the importance of investigating the statistical power of a body of research; an issue I will focus on in this study's second analysis.

Meta-Research and Social Work

Though meta-research has not become a *formalized* area of inquiry within social work, one can find historical antecedents. In line with the dissertation study performed here, John Orme examined the statistical power of articles published in *Social Work Research and Abstracts* from 1977-1984 (Orme & Combs-Orme, 1986) and for a decade of articles within the *Journal of Social Work Education* (Orme and Tolman, 1986). Crane (1976) examined the power of studies of the effectiveness of social interventions. Epstein (1990) cleverly (and controversial) investigated response bias (as a proxy for publication bias) among social work journals, by submitted articles reporting positive or null findings. Recently, Thyer (2015) sought to determine the number of randomized controlled trials (RCTs) employed in the social work literature.

Elsewhere, social workers have expressed concerns congruent with contemporary meta-researchers (e.g. the shortcomings of peer review; see de Smidt & Forey, 1997; Gambrill, 2018). Together, these examples show that there is precedent for meta-research within social work scholarship. Next, I discuss how meta-research fits within the historical and intellectual development of the profession.

Meta-Research and the History of Social Work Research

Though meta-research is not currently a formal area of study within social work research, it complements the field's overarching historical and intellectual development; one in which the field is repeatedly found to be deficient in some manner and reacts to correct the perceived deficiency. This traces in part to social work's desire to establish itself as a rigorous and mature scientific discipline; a desire at least extending back to Flexner's (1915/2001) infamous assessment of the field. Consequently, the field has continuously sought to improve how research is conducted (and recorded) and how that research is used to inform policy and practice (Kirk & Reid, 2002; see also Richmond, 1917).

This is exemplified by the profession's response to the "crisis of effectiveness" experienced across the early-mid 20th century (Kirk & Reid, 2002; see also Fischer, 1973, 1976). The field strove to demonstrate to both private and public supporters and funders that social work can benefit individuals and communities. This spawned widespread reflection on the field's methods and research practices (see generally Mullen & Dumpson, 1972). For example, a number of articles appeared in the 1970s and 1980s drawing the attention of social work researchers to neglected topics in statistical inference (e.g. statistical power) and research design (e.g. Cowger, 1984; Crane, 1976; LeCroy, 1985; Orme, 1987; Tripodi, 1974), as well as how social worker practitioners understood and utilized research findings (e.g. Glisson & Fischer,

1987; Kirk & Fischer, 1976). The 1990s witnessed concern for the dearth of social work scholarship compared to adjacent fields and for the continued gap between research and practice (see Task Force on Social Work Research, 1991). These concerns primed the profession for the later, widespread, adoption of evidence-based practice (EBP) in the late '90s and early 2000s (Gambrill, 1999; Okpych & Yu, 2014).

The Benefits of Social Work Meta-Research

We do not know if social work is in the midst of a replication crisis. The meta-research study performed here will help appraise the state of contemporary social work research. It is a “first step” towards assessing the “health” of the field. However, even if the field is not in crisis there are numerous benefits to performing meta-research within social work. First, such a study would help ensure that contemporary research practices are appropriately employed and act as a feedback mechanism for social work researchers and journals (Hardwicke et al., 2019). By doing so, social workers could have greater confidence in the field’s research practices and outputs. Second, the study would complement current efforts by meta-researchers to appraise the quality of research; thus bridging the gap between social work and the research trajectories of other fields. Third, it would complement historical efforts to enhance the rigor of research practices (as described above) and introduce a novel area of inquiry to the field.

The Plan Going Forward

In the next chapter, I outline the relevant statistical theory and concepts, focusing on frequentist or “classical” statistics. This will provide context for the problems discussed in the present chapter (e.g. low power, use and misuse of frequentist statistics) and clarifies the concepts under study in this dissertation.

CHAPTER 2

STATISTICAL BACKGROUND: THEORY AND CONCEPTS

How a discipline utilizes methodological tools can greatly impact its development. For instance, when used judiciously statistical analyses can provide a measure of fit for competing models (e.g. Glover & Dixon, 2004) or a measure of evidence for a given hypothesis (e.g. Dienes, 2016; Morey, Romeijn, & Rouder, 2016). Misunderstood or used inappropriately (Greenland et al., 2016) and researchers can arrive at erroneous conclusions, waste finite resources, and pursue fruitless lines of research. Given the central role statistics play in social work research and their utility for performing meta-research, it is imperative that the meaning and utility of statistical tools and concepts be made clear and precise.

In this chapter, I briefly outline the major schools of statistical inference, giving the majority of my attention to frequentist (aka “classical”) statistics. Then, I sketch out two major approaches within the frequentist school: the Fisherian and Neyman-Pearson approaches. The social sciences, including social work, commonly employ a blend of these two approaches, dubbed Null Hypothesis Significance Testing (NHST). I then discuss key statistical concepts employed within frequentist statistics, including: alpha, power, p-values, confidence intervals, and effect size. Clarifying the meaning of these concepts will help elucidate some of the problems discussed earlier and lay the groundwork for the current study.¹⁰ I close by considering some limitations of frequentist statistics as used in NHST and by considering alternative approaches to statistical practice for social work.

---

¹⁰ Recall, misuse of frequentist statistics (and related problems such as low statistical power) are among the main driving forces behind the replication crisis (e.g. Colling & Szűcs, 2018; Dienes, 2016; Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011).

Schools of Statistical Inference

When thinking about the strengths and limitations of a given statistical approach, it is helpful to consider the goals implicit to the endeavor. Biostatistician Richard Royall (1997) provides a useful taxonomy for thinking about these goals. When faced with an observation, Royall states that a researcher may ask themselves one of three overarching questions: 1) What do I believe? 2) What should I do? and 3) What does this observation tell me about hypothesis ‘A’ vs hypothesis ‘B’? (p. 4). These three questions (and their associated goals) correspond to the three dominant schools of statistical inference. Question 1 (belief) corresponds to the Bayesian school. The researcher is tasked with adjusting their epistemic beliefs in light of new evidence. Question 2 (action) corresponds to the frequentist school. The researcher is tasked with making a correct decision in light of the observation. Question 3 corresponds to the Likelihood school (evidence). The researcher is tasked with evaluating the extent to which the observation provides evidence for one hypothesis over another; if it does so at all. Although researchers have interest in each of these questions,¹¹ the scientific landscape (including social work research), has largely been dominated by the frequentist school (Freedman, 1995, p. 20; see also Dienes, 2008; Oakes, 1986; Sober, 2008).

The Frequentist School

Giving a unified account of frequentist statistics is difficult, if not impossible. This is due in part because the major figures within this school (e.g. Ronald A. Fisher and Jerzy Neyman/Egon Pearson) spent much of their careers in heated disagreement over basic principles of statistical inference and methodology (Berger, 2003). Nevertheless, the following four

---

¹¹ Indeed, researchers are often guilty of giving frequentist concepts Bayesian interpretations (e.g. Dienes, 2011, p. 276), a point Bayesians commonly cite in their favor.

features can be said to characterize frequentist statistics. Though no individual feature will precisely align with the practice of statistical inference put forth by either Fisher or Neyman-Pearson, together they form the philosophical underpinnings of current statistical practice (i.e. NHST).¹²

First, frequentist statistics, as the name suggests, is rooted in a specific interpretation of probability: *the frequency theory of probability*;¹³ which traces back to the work of Reichenbach, Venn, and von Mises (La Caze, 2016). The frequency theory defines probability in terms of the *long run* relative frequency of a specific event. Philosopher Alan Hájek (2008) captures this definition well: “[T]he probability of an outcome A in a reference class B is the proportion of occurrences of A within B” (p. 96). In other words, for some particular event, outcome, or property, probability (as defined here) is the number of occurrences within an infinite sequence. Two examples will help illustrate.

First, consider a two-sided coin (heads | tails). We can specify an outcome (A): the coin landing heads. The reference class (B) is coin tosses. Our probability is defined as A/B. The probability of a coin landing heads, *in the limit* of this infinite sequence of coin tosses, is hypothesized to be 50%. To be clear, from the frequentist perspective we are *only* thinking about probabilities in terms of the long run. It would not make sense to say that a single coin has a 50% probability of landing heads on a single toss (or other finite sequence); only that as the number of

---

¹² For example, Fisher (1930) originally favored a fiducial interpretation of probability over the frequentist version discussed here.

¹³ Also known as “the frequency interpretation of probability” (La Caze, 2016) or “long run frequency” (Hacking, 1976, p. 2).

tosses approaches the limit in an infinite sequence, shall the percentage of tosses landing heads approach 50%.¹⁴

Consider a more pragmatic example, inspired by Reichenbach (1951, p. 237). There is a man named Brian. Brian is aged 55 and has cancer (we'll call this cancer "Lacasseemia"). His doctor tells Brian that with treatment 10% of men, just like Brian, survive. How does a frequentist understand this claim? We have an event (A: cancer survival) and a reference class (B: men aged 55 with Lacasseemia). In the long run 10% of men, just like Brian, survive with treatment. However, this is not to say that Brian has a 10% chance of survival. The frequentist notes that Brian will either survive or he will not. From the frequentist perspective, it does not make sense to assign a probability to the individual event (i.e. Brian's survival); probability applies to the long run across the reference class (survival rates for men like Brian), not to individuals therein.

We'll see how the frequency interpretation of probability¹⁵ comes into play below, when discussing error rates of statistical procedures, such as the confidence interval.

The second feature regards the *sample space*. In frequentist statistics, the sample space pertains to *all* possible outcomes of a specific phenomenon (Henkel, 1976, p. 16). For instance, if I toss a fair coin once, the sample space contains two possible outcomes: heads or tail [H, T]. If I toss the same coin *two* times, the sample space would be all possible combinations of those two

---

¹⁴ This may seem counterintuitive. But consider for a moment a case in which you toss a coin three times. It would not make sense to say that, since the probability is 50%, we expect to observe 1.5 heads and 1.5 tails. Rather we will either see 2 heads and 1 tail or 1 head and 2 tails. It is in the long run, that the frequencies converge towards 50%.

¹⁵ A final note on the frequency interpretation before moving on: Statistician, and self-described frequentist, David Freedman (1995) argues that there is a difference between relative frequency and frequency probability (p. 23). That is, that the relative frequency of an event may be different than the long run probability. I think he is mistaken. The example he uses to make his argument appears to rely on a version of frequentist probability known as *finite* frequentist probability; which is different than the *infinite* frequentist probability employed by Mises and Fisher (Hacking, 1976, p. 6; see also La Caze, 2016 for differences between finite and infinite frequentism).

tosses (regardless of the outcome) (Kruschke, 2011, p. 30). In this scenario the sample space is: [H-H, T-T, H-T, and T-H].

In frequentist statistics tools such as the p-value condition on the sample space. That is, p-values are conditioned not only on the observed results, but also on results that did not *but could have* occurred, no matter how unlikely; as defined by the sample space. This is apparent when revisiting the definition of a p-value. A p-value is: “the probability under the null hypothesis *of observing data at least as extreme* as the data that were actually observed” (emphasis added; Wagenmakers, Lee, Lodeqtckx, & Iverson, 2008, p. 184). In other words, the p-value is partly determined by data that “was never observed” (p. 184) in the study. One can begin to see similarities between the frequency theory of probability and the sample space.

Third, frequentism eschews *prior knowledge*. Neither Fisherian significance testing nor Neyman-Pearson hypothesis testing formally assigns prior (or posterior) probability to hypotheses (Mayo, 2003, p. 20). This is in direct contrast to practice within the Bayesian school (Howson & Urbach, 1993), which takes into account a researcher’s beliefs prior to data analysis and incorporates those beliefs into parameter estimation and resulting posterior probabilities (see Etz & Vandekerckhove, 2018 for a gentle introduction).

The fourth feature pertains to the goals of statistical inference. In this case, how one should *act*, in light of some observation or evidence. Fisher and Neyman-Pearson provide two distinct approaches to this goal. I will sketch out the associated concepts in more detail below. For now, I’ll provide a simplified representation of each approach; recognizing that in NHST, these approaches are often blended together (Gigerenzer, 2004). Under the Fisherian approach a researcher articulates a hypothesis to be tested or “nullified”: the *null hypothesis*. A significance level (α) is set and data collected. If the calculated p-value is less than or equal to this pre-

specified threshold (e.g. if $\alpha = 0.05$ and $p \leq 0.05$), then “either the null is true (and something unusual happened), or the null is false” (Hacking, 2001, p. 215). Using a form of probabilistic *modus tollens*,^{16,17} Fisher would then recommend that we “reject” the null hypothesis.¹⁸

The approach developed by Neyman & Pearson (1933) is much more explicitly behavioristic. Neyman (1957) states “[T]he purpose of every piece of serious research is to provide grounds for the selection of one of several contemplated courses of action” (p. 16). Here we can see the clear link to Royall’s second question: “What should I do?”. However, unlike Fisherian significance tests (reject or fail to reject), the Neyman-Pearson is contrastive. It considers not only the rejection and acceptance of a single null hypothesis, but of a second, (explicit) *alternative* hypothesis as well. Neyman and Pearson (1933) argue that having a second hypothesis is a necessity, stating, “We are inclined to think that as far as a *particular* hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of that hypothesis.” (pp. 290-291, emphasis added). In other words, one must have a second hypothesis with which to contrast the truth/falsity of the one being scrutinized.

---

¹⁶ This can be formally represented as (Sober, 2008, p. 49):

$$\begin{array}{l} \text{Pr}(O \mid H) \text{ is very high} \\ \text{not } O \\ \hline \text{not } H \end{array}$$

¹⁷ It should be noted that probabilistic *modus tollens* has been heavily criticized as an inferential procedure (Sober, 2008; see Chapter 1).

¹⁸ Fisher vehemently opposed the behavioristic approach of Neyman-Pearson (Mayo, 1996, p. 380), preferring an evidential interpretation of p-values (Fisher, 1935). He would likely recoil in my characterization of his methods as behavioristic. However, given the questionable ability of p-values to serve as a form of evidence (Royall, 1997; Sober, 2008) and the behavioral tendencies implicit in Fisher’s early work (Mayo, 2018, p. 176), I argue that this is a defensible characterization; whatever Fisher’s misgivings.

With two competing hypotheses, the researcher is faced with a new set of problems. They must consider whether the acceptance or rejection of *either* hypothesis is made in error (see Table 1). The Neyman-Pearson approach seeks to quantify and minimize the probability of these errors (α , β) in the long run. Alpha (α) remains in use as the significance level, as in Fisherian testing, but is not interpreted as a measure of the strength of evidence. Rather it becomes the fixed rate for incorrectly rejecting the null hypotheses over the long run (i.e. Type I error rate). Alpha is set by the researcher, in advance, and is typically placed at the 0.05 threshold; though this is a matter of convention and a different value may be justified depending on the context or needs of the researcher (Lakens et al., 2018). The second error, erroneously accepting the null as true when it is false (i.e. Type II error or beta [β]), is reduced by maximizing the likelihood of finding a true effect, when in fact one exists (i.e. statistical power). Given that the minimization of one error results in the inflation of another the researcher is tasked with considering the costs of each type of error and the trade-off in making one over another.¹⁹ Neyman and Pearson (1933) make this clear:

“From the point of view of mathematical theory all that we can do is to show how the risk of the errors may be controlled and minimised. The use of these statistical tools in any given case, in determining just how the balance should be struck, must be left to the investigator”. (p. 296).

Though these four features of frequentist statistics do not neatly fit into the Fisherian and Neyman-Pearson approaches outlined above (nor should this list be considered exhaustive), they provide sufficient background for the frequentist school and context for its concepts and tools. I

---

¹⁹ This is a bit of a simplification, as there are other types of error one can make (see Gelman, 2014 for discussion of *Type M* [sign] and *Type S* [magnitude] errors).

turn to these tools and concepts next, and then situate them within the NHST framework used in current practice.

Table 1.

Four Possible Statistical Outcomes

Researcher's Decision	State of the World	
	<i>H0 is True</i>	<i>H1 is True</i>
<i>Accept H0</i>	Correct Decision	Type II Error (β)
<i>Reject H0</i>	Type I Error (α level)	Correct Decision

Alpha (α)

Alpha is a *pre-experimental* concept (Berger & Wolpert, 1984, p. 5) used to inform the design of an experiment. It is the long run probability that a researcher will make a Type I error, under the null hypothesis (Dienes, 2008). A Type I error is when the null hypothesis is rejected, when it is in fact true, commonly referred to as a “false positive”. Alpha levels are traditionally set at the 0.05 level, though this is largely a matter of convention (Garland Campbell, 2018). If the same experiment was performed 100,000 times, with an alpha level of 0.05, the number of cases where the researcher would incorrectly reject the null (i.e. Type I errors) is approximately 5,000. This is because, under the null hypothesis all p-values are equally likely (Hung, O’Neill, Bauer, & Köhne, 1997). Since every p-value has the same probability of occurring when the null is true, we can fix the percentage that will be “false positives” by stipulating our alpha level.

Though, it must be noted that we never truly know in advance whether our hypotheses are true or false.

Figure 1 demonstrates this phenomenon by simulating a study where the null is true, 100,000 times (with an alpha of 0.05).²⁰ As can be seen from this figure, each interval (0-0.1, 0.1-0.2, ... 0.9-1.0) has the same density. P-values are distributed uniformly, and we can infer, for example, that the likelihood of observing a p-value of 0.25 is the same as observing a p-value of 0.75, under the null.

Figure 1.

Distribution of P-Values

---

²⁰ This simulation was performed in R using code modified from Daniël Lakens' online statistics course "[Improving Your Statistical Inferences](http://shiny.ieis.tue.nl/MOOC/assignment_1/)". The associated Shiny App is here: http://shiny.ieis.tue.nl/MOOC/assignment_1/

Where a researcher should set alpha is the subject of much debate (this point is discussed below further; see Benjamin et al., 2018; Lakens et al., 2018; Mudge, Baker, Edge, & Houlahan, 2012 for more detailed discussion and debate). If you set alpha at a more strict level, say 0.01 or 0.005, you'll almost always avoid Type I errors (e.g. ~1,000 per every 100,000 iterations of the study where the null is true when using $\alpha = 0.01$). However, the consequent reduction in statistical power (discussed below) means that while you are less likely to make a Type I error, you are also less likely to find a true effect, if one exists. How these trade-offs should be made requires much consideration as to the real-world consequences of each error (see Field et al., 2004 for an example within the field of environmental conservation) and the broader state of research in a particular field.

Misconceptions. Since alpha is the long run error rate, it does not tell you about the probability of making a Type I error in any given study (Greenland et al., 2016). Nor can it tell us you the percentage of published studies that are false positives (Oakes, 1986), due in part to publication bias (Rosenthal, 1979).²¹

Beta and Statistical Power

Beta (β). Whereas alpha is the long run probability of incorrectly rejecting the null, beta is the probability, over the long run, of failing to reject the null when it is false (Dienes, 2018); also known as the Type II error rate (false negative). Like alpha, it is a pre-experimental concept. Unlike alpha, beta is difficult to control as it is affected by one's sample size, alpha level, the

---

²¹ It should also be noted that the meaning of many statistical concepts reviewed here, such as the Type I error rate, are still vociferously debated. Rubin (2019) provides a compelling analysis distinguishing between Fisherian and Neyman-Pearson Type I error rates. He argues that, "the Neyman-Pearson Type I error rate refers to a replicable random sampling procedure that has the potential to sample from all relevant subsets in the target population of interest during a long run of exact replications. In contrast, the Fisherian Type I error probability refers to a hypothetical sampling procedure that is restricted to an imaginary sample-specific population (reference set) that does not contain any recognizable relevant subsets." (p. 13). In making this distinction, Rubin argues that the Fisherian interpretation is the most appropriate for the social sciences.

effect size under the alternative hypothesis, measurement error, and a host of other assumptions²² unique to the statistical model and testing situation. To understand the concept of beta and how it may be minimized, we must discuss its counterpart; statistical power.

Power (1- β). Power is the long run probability of correctly rejecting the null hypothesis when it is false (i.e. $\text{Prob}(\text{reject } H_0 | H_0 \text{ is false})$, Dienes, 2008, p. 63). Alternatively, we may say that power is the long run probability of correctly accepting the alternative hypothesis, when it is true.²³ It is inversely related to beta; and can be calculated by subtracting beta from 1 (i.e. $1-\beta$). Consequently, as beta increases, power decreases. As beta decreases, power increases. The task of the frequentist researcher then, is to set the statistical power of a study to a level where they are comfortable in making a Type II error at a rate of β . This requires the use of a formal strategy called power analysis (Cohen, 1988).

Like beta, power is affected by sample size, alpha level, the effect size under the alternative hypothesis, measurement error, and a host of other assumptions unique to the statistical model and testing situation (Cohen, 1988).^{24,25} As a general rule, as sample size, alpha level, and the hypothetical effect size are increased, so too does the statistical power of a study. Increasing the study sample size reduces error and increases the resemblance between sample and population. When increasing the sample size is not feasible, other strategies can be employed (Lazic, 2018). Since alpha and beta are related concepts, decreasing alpha will reduce power; and

---

²² Random sampling is commonly cited as a prerequisite for valid inference, within the frequentist tradition (Berk & Freedman, 2003; Gorard & White, 2017; White & Gorard, 2017). Though see Gelman, 2017 and Kass, 2011.

²³ Sometimes power is treated as if it is a measure of the strength of evidence (not unlike Fisherian p-values). For a critical discussion of this notion, see Machery (2016).

²⁴ Given that power is a function of alpha, sample size, and effect size, a researcher can solve for any one factor, when the other three are known. Therefore, we can perform three other types of power analyses (as outlined in Cohen, 1988, pp. 14-15), if desired. Though for now, I restrict myself to the most frequently used type; estimating power given sample, alpha, and an effect size (i.e. an a priori power analysis).

²⁵ Power is also affected by design considerations (e.g. between-subjects vs. within subjects; Giner-Sorolla et al., 2019).

increasing alpha will increase power. Finally, the larger the hypothesized effect size, the easier it would be to detect, *if* it exists.

In practice, it is recommended that researchers perform a power analysis prior to data collection (Cohen, 1988) to improve study planning and increase the likelihood that they are able to detect an effect when it exists. A power level of 80% is usually recommended, though this is largely a “convention” (Cohen, 1992, p. 156) and requires further reflection.

In Figure 2, I provide a visualization of power at the 50% level. The figure shows a distribution of p-values for a study where the alternative hypothesis is true; with an alpha level of 0.05 and 50% power; simulated 100,000 times. As you can see the distribution is right-skewed; with half of p-values fall between 0.00 and 0.05 and half fall between 0.05 and 1.0. As a study’s power is increased the skew becomes more extreme (see Hung et al., 1997).

Misconceptions and limitations of power. Like many frequentist concepts, power is often misrepresented as applying to an individual event rather than the long run frequency of a study (Greenland et al., 2016). A significant p-value in a study with 80% power does not mean that there is a 20% chance of error. We cannot be certain, in the individual study, whether the null or alternative hypothesis is true. Similarly, for any pair of *individual* studies it is not the case that the higher-powered study is *always* more informative than the lower-powered study (since power is a *pre-experimental* concept and is conditioned on outcomes beyond those actually observed). It may be the case that the lower-powered study provides stronger evidence in that particular instance; however unlikely this may be over the long run. This is what Wagenmakers et al. (2015) call *the power fallacy*.

Perhaps more jarring are the consequences of measurement error on power itself. Power is calculated under ideal conditions. However, if the tools used to measure the effect size are not

reliable, then a study's power can be drastically reduced (Kanyongo, Brook, Kyei-Blankson, & Gocmen, 2007; Parsons, 2018). For a paired samples t-test, with 80% power, an alpha level of 0.05, an effect size of .5 (Cohen's d), and perfect reliability conditions a researcher would need a sample size of 34. This sample size is nearly triple that needed for a large effect size ($d = .8$); again under ideal conditions. If we reduce reliability of our measures to .85, we suddenly need a sample size of 44 to achieve 80% power for a medium effect. The necessary sample size drastically increases as the size of the effect decreases.^{26,27}

Figure 2.

50 % Statistical Power, Simulated 100,000x

²⁶ These figures are based off of Table 6 in Kanyongo et al (2007, p. 87).

²⁷ One final note on power: Because we never *actually* know the true value of the effect size for a given phenomenon, power analysis is a limited and fallible technique. However, it is possible to estimate a study's power across effect sizes. Morey and Lakens (2016) elaborate: "Although researchers often talk about 'the' power of a test, and indicate the power of a test by a single value (e.g., 90% power), *a test has power to all hypothetically possible true values* of the effect size. Therefore, it is important to think of statistical power as a curve of values". (p. 5, emphasis added). This alternative strategy permits the researcher greater flexibility in study design and conceptualization.

P-Values and Statistical Significance

P-values. P-values are one of the most commonly used frequentist tools and perhaps the most understood (Aschwanden, 2015). What is a p-value? It is the probability of obtaining data that is as extreme (or more extreme) than the observed results, under the null hypothesis (and other assumptions inherent to the statistical model; see Freedman, Pisani, & Purves, 1978, pp. 437-453). Put differently, a p-value is the, “probability of the observed result, plus more extreme results, if the null hypothesis were true” (Goodman, 2008, p. 136). This last point cannot be stressed enough. A p-value is calculated, on the basis of the observed data, under the assumption that the null hypothesis is *true* (e.g. $p(D|H_0)$). Observing data that results in a large p-value (e.g. $p = 0.80$) would not be very surprising under the null model (i.e. the data is compatible with the null hypothesis); whereas data that results in a small p-value (e.g. $p = 0.02$) would be surprising under the null (or “more incompatible” with the null hypothesis) (Wasserstein & Lazar, 2016, p. 131).

Misconceptions. There are numerous misconceptions about p-values (see Goodman, 2008 and Greenland, 2019 for an overview). I present a few of the more notable ones here. First, a p-value is not the probability that the null hypothesis is true. Again, a p-value is derived from $p(D|H_0)$, not $p(H_0|D)$. Therefore, $p = 0.05$ *does not* mean that there is only a 5% chance that the null hypothesis is true. Second, large p-values *do not* prove that the null hypothesis is true. Rather a large p-value merely indicates that the data observed is “not surprising” (Motulsky, 2016, pp. 85). Third, one minus p (i.e. $1-p$) *is not* the probability that the alternative hypothesis is true (If $p = 0.03$, it does not mean that there is a 97% chance that result is a true effect; rather, under the null hypothesis [if true], only 3% of experiments would yield a more extreme result.

Statistical significance. Having defined alpha and the p-value, we can turn our attention to an equally contentious concept: statistical significance. Simply put, a p-value is statistically significant when it is less than or equal to our pre-specified alpha level. If alpha is set at 0.05 and we observe a p-value of 0.04, we are permitted, in the frequentist tradition, to state that our result is statistically significant. In the Neyman-Pearson approach alpha is fixed in advance. If we observe a p-value of 0.009 and had set alpha at 0.05, it is statistically significant with regard to *that* threshold. We cannot say that it is significant at the 0.01 threshold unless that was the alpha level set in advance; though use of multiple alpha-levels is common in current practice (Meehl, 1978; Gigerenzer, 2015).

Misconceptions and controversies. The value of statistical significance and the use of significance tests is subject to much debate (Rasnick, 2017; Morrison & Henkel, 1970; see also Machery, 2019 and Morey 2017, 2018 for critical discussion). Some have argued that setting alpha at 0.05 leads to high rates of false positive findings and encourages results that are of low evidential value to be reported as “significant” (e.g. Benjamin et al., 2018). Consequently, Benjamin et al. (2018) suggest moving the threshold from 0.05 to 0.005 for novel findings. Others (e.g. Lakens et al., 2018) have argued that uniform thresholds (however large or small) are arbitrary and can have drastic consequences for scientific practice. Instead, Lakens et al. (2018) argue that researchers should *justify* their alpha level in advance, based on the unique research context; a sentiment consistent with Fisher (1956), who states “[N]o scientific worker has a *fixed* level of significance at which, from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.” (p. 42, emphasis added). Still others have suggested that statistical significance is not a useful concept, lacks direct bearing on clinical/practical significance of a finding (McCloskey,

1985, 1995; Ziliak & McCloskey, 2008), and thus should be retired (Amrhein, Greenland, & McShane, 2019), abandoned (McShane, Gal, Gelman, Robert, & Tackett, 2019; Wasserstein, Schirm, & Lazar, 2019), or even banned (Trafimow & Marks, 2015). The debate about statistical significance appears to linger on and researchers are largely continuing to use such concepts in everyday practice (see below on NHST).

Effect Size

An effect size is best conceptualized as being the “magnitude of some measure” or effect under study (Nickerson, 2000, p. 280; e.g. the degree of association between two or more variables). An estimated effect size is integral for performing an a priori power analysis (Cohen, 1988). By performing a power analysis, the researcher is tasked with considering what size effect they truly care about (Lakens, 2017). Once the study is performed, an *observed* effect size (e.g. association between two variables or of group indices) can provide information about magnitude (see Ferguson, 2009 for an overview). Though effect sizes can be categorized by size (e.g. the small, medium, and large designations of Cohen, 1962), size does not necessarily correspond to clinical or practical significance; or what Ziliak and McCloskey (2008, 2014; McCloskey, 1995) call “oomph”. Additionally, effect sizes can be presented in raw or standardized forms. Common examples include: Pearson’s r , Cohen’s d , odds ratios, and eta-squared (η^2); among numerous others.

Confidence Intervals

Confidence intervals have been put forth as one solution to the problems surrounding p-values (e.g. Gardner & Altman, 1986; Rothman, 1978; Fidler, Thomason, Cumming, Finch, & Leeman, 2004). Though potentially useful in practice, confidence intervals have also been difficult for researchers to interpret (Greenland et al., 2016; Hoekstra, Morey, Rouder, &

Wagenmakers, 2014; Poole, 1987). To clarify then, a confidence interval is an interval estimate, calculated from sample data, for some population parameter (Cumming & Finch, 2005). Values within the interval are believed to be more “plausible” values for the parameter (p. 176) than those residing outside of the interval.²⁸ In this regard, a confidence interval can be said to give an indication to the *precision* of an estimate (Naimi & Whitcomb, 2020). In general, the wider the range of values, the less precise the estimate. The narrower the range of values, the more precise the estimate.

Great care should be taken when interpreting a given confidence interval. Typically, researchers will construct a “95% confidence interval” around some parameter estimate. The percentage here (i.e. 95%) merely refers to the characteristics of the procedure created to generate the interval (see Neyman, 1937). It *does not* refer to the probability that an individual confidence interval contains the estimated population parameter.²⁹ 95% refers to the probability over the long run. In other words, the *performance* of the procedure. We cannot say whether any individual interval contains that parameter (Morey, Hoekstra, Rouder, Lee, & Wagenmakers, 2016), nor determine the probability that it does for any individual interval. We only know that according to our procedure that out of 100,000 intervals, approximately 95,000 would contain the interval and 5,000 would not.³⁰ To estimate the probability that a population parameter falls within a specific interval range would require use of Bayesian *credibility intervals* (Dienes, 2008).

---

²⁸ Greenland et al. (2016) would contend that this claim of “plausibility” rests on a series of assumptions about the statistical tests performed and assumptions underlying the calculation of the confidence interval.

²⁹ Though even Nobel Prize winning economist Milton Friedman has been guilty of this error (Giles, 2011).

³⁰ I say “approximately” because in any finite set of intervals, the probability may not be 95%.

NHST and Limitations of Frequentist Statistics

NHST in practice. The dominant statistical method used in the biomedical and social sciences (including social work) is Null Hypothesis Significance Testing or NHST (e.g. Bushway, Sweeten, & Wilson, 2006; Schwab, Abrahamson, Starbuck, & Fidler, 2011; Stang, Poole, & Kuss, 2010). NHST, rooted in frequentist theory, is an amalgamation of the Fisherian and Neyman-Pearson approaches (Perezgonzalez, 2015). The two approaches are often blended, unbeknownst to the researcher; even when doing so is theoretically inconsistent³¹ (Gigerenzer, 1993) or harmful to the reliability of scientific research.

Researchers are often unaware of the limitations of NHST (Schwab et al., 2011), leading to it frequently being applied in the following way (see Cohen, 1994; Szűcs & Ioannidis, 2017; Nickerson, 2000 for a more detailed overview):

- 1) A null hypothesis is set by the researcher. Typically, the hypothesis posits “no effect”. This is also known as the “*nil* hypothesis”.
- 2) No alternative hypothesis is proposed or specified [Note: not unlike in Fisherian testing].
- 3) Whenever p is less than 0.05, report the finding as a statistically significant result and “reject the null hypothesis” in favor of an effect [Note: Not entirely dissimilar from Neyman-Pearson testing; though see point #2 above].
- 4) Publish results, highlighting statistically significant findings; down-playing or minimizing non-significant findings.

---

³¹ Mayo (2018) defends a blended version of Fisherian and Neyman-Pearson statistics, centered around Popper’s “severe test” principle (1979).

Other features:

- 5) Alpha is not commonly pre-specified; and is “flexible” depending on the size of the observed p-value (Meehl, 1978).
- 6) Power analyses are not routinely performed or are done so retrospectively.
- 7) Effect sizes and confidence intervals are inconsistently reported or eschewed altogether (Schwab et al., 2007).

This may seem a harsh or unfair characterization of current practice, but the body of social science research over the last 50 years largely suggests otherwise. It may be a caricature of formal statistical inference (as described by Fisher and Neyman-Pearson), but is not therefore an inaccurate representation of current statistical practice (e.g. Cohen, 1994; Gigerenzer, 2004; Meehl, 1978; Wetzels et al., 2011; Ziliak & McCloskey, 2008).

Limitations of frequentist approaches. It would be inappropriate to clarify the nature and use of frequentist statistics and use of NHST without briefly discussing their limitations. This is not the place for a full exposition, but some important highlights can be sketched (see Oakes, 1986; Royall, 1997; and Sober, 2008; Wagenmakers et al., 2008 for an overview of limitations).

At a theoretical level, frequentist statistics have difficulty mapping onto “everyday practice”. The frequency interpretation of probability does not permit inferences to single studies, but rather the long run behavior of a procedure. This may be of no concern to the quality control manager in a large manufacturing center, but such a procedure lacks obvious utility for the types of studies performed within the field of social work. Further, social science research is not the same as cases involving the tossing of a coin or the rolling of a die. We can envision the long run probabilities of the latter events, but know *a priori* that a single study will never be

repeated in the exact same way, except hypothetically; a point that brings into question what the appropriate reference class for an empirical study within social work (Howson & Urbach, 1993).

Second, whatever the limitations of Bayesian statistics, the eschewal of prior probability within frequentist statistics would seem naïve. The prior probability of our null and alternative hypotheses is essential for understanding the results of a study. A positive result would not appear to mean much if the prior probability of it occurring is extremely low. Further, the concept of using prior knowledge/beliefs to inform study design and updating those beliefs after testing is arguably one of the essential functions of the scientific enterprise.³²

Third, frequentist statistics do not provide direct evidence for hypotheses (Wagenmakers, 2004), but rather serve as a guide for long run behavior. Neither can tools such as the p-value, provide support for the null hypothesis (Wagenmakers et al., 2008). If one of our primary goals, as researchers, is to produce evidence that directly bears upon our research hypotheses, then frequentist statistics is (at least partially) inadequate.

Alternative Approaches

My current stance on these issues is as follows: Given that frequentist statistics are ubiquitous in social work research, reform of current practices may be worthwhile while alternative methods (e.g. Bayesian methods) are developed and integrated. I do not envision that Bayesian methods will replace and supersede frequentist approaches, but rather exist alongside them. Their use will depend on the research question under investigation, what the researcher wishes to gain from the experiment (i.e. as captured by Royall's three questions), and the best analysis for answering it. This should not be considered a form of statistical pluralism (or

---

³² This point is gaining greater traction. The Food and Drug Administration (2010), for example, has started to embrace the use of Bayesian statistics in clinical trial design and analysis.

anarchism), but perhaps as a version of statistical pragmatism (Kass, 2011). The choice of statistical approach should be justified in advance, as should other statistical decisions (i.e. alpha; Lakens et al., 2018).

Reforming NHST. Frequentist statistics have limitations, but so long as we are using them we may as well do so as rigorously as possible.³³ This would entail improving how such concepts are taught to students (e.g. Cassidy et al., 2019; Dunleavy & Lacasse, 2020) and assessing how they are being interpreted by researchers (Gigerenzer, 2018; Hoekstra et al., 2014; Oakes, 1986). If nothing else, improving our understanding of the meaning of frequentist concepts should increase our understanding of their strengths and limitations. Given the institutional and professional inertia around learning new methods,³⁴ this seems a viable path.

This effort can be complemented by other types of methodological reform, such as the improvement of study conceptualization and design (Wasserstein & Lazar, 2016). Journal editors, funders, and institutional leadership can encourage practical changes to the way we conduct and report our research (Shlonsky, D'Andrade, & Brookhart, 2002; Smaldino & McElreath, 2016). Researchers would be encouraged to routinely: 1) engage in sample-size planning, 2) perform *a priori* power analyses, 3) pre-register statistical analyses, 4) submit Registered Reports, 5) make code, analyses, and data available in a public, online repository (e.g. discussed by Chambers, 2017), and 6) explicitly consider the role of theory for the phenomenon under study (e.g. Borsboom, 2013; Gambrill, 1997; Gomory, 2001a, 2001b; Muthukrishan & Henrich, 2019; Nugent, 1987) (see also proposals suggested earlier in this dissertation).

---

³³ This includes understanding when inferential statistics should be avoided altogether (e.g. Lacasse, 2008).

³⁴ Stamey, Sherr, and Williams (2011) is one of the few (if not only) explicit calls within Social Work for use of Bayesian methods.

Expanding our statistical toolbox. Shifting the collective behavior of a profession is difficult and daunting. To ease or “nudge” the profession towards new methods, instructors, mentors and editors/reviewers could encourage students and researchers to use complementary statistical tools (when appropriate) in addition to traditional ones (Ioannidis, 2019). Dienes, for example, suggests providing, “a B for every P” (Dienes & Mclatchie, 2018, p. 216), meaning a Bayes factor for every p-value (see Benjamin & Berger, 2019; Lakens et al., 2018; Wetzels et al., 2011).³⁵ Quintana and Eriksen (2017) concur. In addition to improving author interpretation of results, they argue that “...this will also assist readers and reviewers [to] better understand the benefits of Bayesian statistics.” (p. 7) (see also Gigerenzer, 2004; and as a counterpoint Fidler et al., 2004). This prescription to introduce complementary tools corresponds well with Kuhn’s earlier work (1977), which argues that scientists and disciplines must necessarily balance traditional views and methods with iconoclastic ones in order to be successful and innovative (see especially Chapter 9).

Conclusion

As stated at the beginning of this chapter, how a discipline utilizes methodological tools can greatly impact its development; for better or worse. Used inappropriately or without forethought, inferential statistics can provide the “image of objectivity” (Tripodi, 1974), without making substantive advances in our knowledge of social problems. However, when used thoughtfully (see Cohen, 1994), inferential statistics can shed light on our overarching research questions (Hand, 1994; see also Glass, 2010) and help build fruitful programmes of research (Dienes, 2008; Lakatos, 1968). In reviewing the frequentist school and its underlying concepts

---

³⁵ See Ly (2017) for a historical and pragmatic overview of Bayes factors.

and tools, I have provided a firm backdrop for exploring the state of contemporary, quantitative social work research. In the next chapter I describe the methodology of this dissertation study.

CHAPTER 3

METHODOLOGY

Purpose and Outline of the Study

The purpose of this meta-research study is to assess statistical practices in contemporary social work research. It parallels recent work in psychology (Nuijten, 2018), neuroscience (Szucs & Ioannidis, 2017), medicine (de Vries, Roest, Turner, & de Jonge, 2018; Herrera-Perez et al., 2019), ecology (Fraser, Parker, Nakagawa, Barnett, & Fidler, 2018; Kelly, 2006), and other scientific disciplines (see Ioannidis, Fanelli, Dunne, & Goodman, 2015); while introducing a formal program of inquiry to the social work profession. The study is comprised of three complementary parts, focusing on the 1) statistical reporting practices, 2) statistical power, and the 3) evidential value of a subsection of research studies. Together, these three parts provide a nuanced appraisal of contemporary social work research.

Part 1 evaluates statistical reporting practices of published studies. This includes examining the frequency and manner in which standard statistical concepts (i.e. alpha, power, p-values, and confidence intervals, effect sizes) are presented in published articles. Adequate reporting is necessary for interpreting and conveying research findings; and permits potential replication and reproduction of study results by reviewers, journal editors, and other researchers (Thyer, 1989, 1991). Incomplete reporting of statistical concepts may mislead readers (Altman, 1994), resulting in inappropriate conclusions about the rigor of study design and/or the validity of study results (Glasziou et al., 2014). For example, Blier et al. (2010) reported the results of a randomized controlled trial comparing the use of two antidepressants (i.e. combination therapy) to antidepressant monotherapy in the treatment of major depression. The authors concluded that combination therapy was more effective but failed to report confidence intervals for their

analyses. Lacasse (2011), reanalyzed the data and calculated confidence intervals for the reported Number Needed to Treat and argues that, in fact, there is “substantial uncertainty” (p. 152), and therefore combination therapy cannot be recommended on the basis of this study. Further, problems in statistical reporting may cause substantive claims to become unfalsifiable (see generally Popper, 1959/1968), among other issues.³⁶ In sum, Part 1 provides an important indicator how social work research is reported within the published literature, as well as serving as the foundational component for the second and third analyses.

Part 2 surveys the statistical power of published studies. Statistical power, the long-run probability of correctly rejecting the null hypothesis when it is false, is an indispensable feature of contemporary (frequentist) statistical practice. Performing an a priori power analysis, as a part of study planning, is one means by which researchers can help increase the reliability of inferences made from sample data. This survey of power informs us about the chances a study from the sample had of finding a statistically significant effect, if one indeed exists (i.e. the population effect is not zero). Or put differently, this survey will help answer “What kind of chance...investigators have of rejecting false null hypotheses?” (Cohen, 1962, p. 145). In doing so, it builds upon the seminal program of research developed by psychologist Jacob Cohen (1962, 1988, 1992, 1994; see also Sedlmeier & Gigerenzer, 1989); and expands upon the work of earlier social worker researchers (e.g. Orme & Combs-Orme, 1986; Orme & Tolman, 1986), who performed power surveys of the field over three decades ago. The results of this survey fill a glaring hole in the social work literature, permit us to answer whether studies within the

---

³⁶ Altman (1982) notes that inadequate research and reporting may also result in risk or inconvenience to current and future research subject/patients, wasted resources (e.g. time, money), lead to the adoption of inappropriate or ineffective interventions, and lead to the premature adoption or termination of various research programs (pp. 66-67); consequences that can be deemed unethical (Altman, 1980).

profession have larger statistical power than in previous decades—where power was particularly poor (see LeCroy, 1985; Orme, 1987), and perhaps lead to methodological reform in study design/planning, where necessary.

Part 3 utilizes a novel statistical technique, called the “p-curve” to assess evidential value (Simonsohn, Nelson, & Simmons, 2014). A p-curve graphs the distribution of reported, *significant* p-values (i.e. between 0.01 and 0.05) in a set of studies. The shape/skewness of this distribution can inform us about the possible presence of bias via Questionable Research Practices (QRPs) (e.g. p-hacking, selective reporting of results, publication bias) (see Chambers, 2017; Epstein, 2004; Fanelli, 2012; Fiedler & Schwarz, 2016; Franco, Malhotra, & Simonovits, 2014; Fraser et al., 2014; Ioannidis, 2014). In the absence of QRPS, the set of results may be indicative of evidential value (Simonsohn et al., 2014, pp. 535-536).

Given the distribution of p-values under the null and alternative hypotheses (e.g. Hung, O’Neill, Bauer, & Köhne, 1997) we can expect a set of significant findings, with no QRPs to be right-skewed (see Figures 2 and 3 of the current study). A distribution that is left-skewed may be indicative of QRPs, like p-hacking and publication bias; or no evidential value if flat. Since p-curves are a relatively new technique and the approach is novel to the profession, I have performed a pilot analysis of a single journal’s results. This is discussed in greater detail below.

Methodology and Study Design

Defining the sample. A sample of articles was retrieved from five leading academic social work journals: 1) *British Journal of Social Work* (BJSW), 2) *Research on Social Work Practice* (RSWP), 3) *Social Service Review* (SSR), 4) *Social Work* (SW), and 5) *Social Work Research* (SWR). These journals were chosen because they are representative of the profession (Aubele & Perruso, 2017) and have a long history of publishing social work research. All five

were ranked within the first tier (top 10) of social work journals according to h-index and g-index metrics in a study by Hodge & Lacasse (2011). A more recent survey of social work faculty found that all five journals were ranked within the top seven social work journals, by mean overall quality score (see Hodge, Yu, & Kim, 2019, p. 5). The same survey found that these journals also represented five of the top six journals in terms of mean prestige score (p. 6).

Further, each journal does well on traditional scientometric tools, such as the impact factor (see study materials); with the BJSW (1.57), SW (1.42), and RSWP (1.27) ranking as the top three social work journals for 2018. SWR and SSR were ranked 11th (0.95) and 15th (0.73) respectively. Given the above considerations and their history of publishing an early form of meta-research (i.e. the publication of power surveys by Orme & Combs-Orme, 1986 in *Social Work Research and Abstracts*³⁷ and Orme & Tolman, 1986, in SSR), the latter two journals were chosen for inclusion in the sample. As a whole, these five journals are deemed to be among the most prestigious discipline-specific outlets and serve as a valid sampling frame from which to assess the state of the social work literature.

Social work journals, as opposed to individual researchers, were chosen as the sampling frame for several reasons. First, prior meta-research studies (e.g. power surveys) have analyzed published articles using journals as their sampling frame (e.g. Brewer, 1972; Cohen, 1962; McCloskey & Ziliak, 1996; Orme & Combs-Orme, 1986; Orme & Tolman, 1986; Sedlmeier & Gigerenzer, 1989; Szucs & Ioannidis, 2017). While largely based on convention (e.g. Orme & Orme-Combs, 1986, p. 6) this approach facilitates comparisons of the published output among journals and provides the opportunity to see how a journal's content has changed over time. For example, by analyzing SWR (see below), one can determine whether there has been an increase

---

³⁷ SWR since 1995.

in the mean power of its published studies over time (e.g. as done by Sedlmeier & Gigerenzer, 1989). Additionally, for journals not previously analyzed in this manner (e.g. BJSW), this can serve as a baseline for future comparisons.

Secondly, using social work journals as a sampling frame may increase the likelihood of capturing discipline-specific knowledge. Focusing on research conducted by social workers, regardless of where it is published, may increase the chances of evaluating studies non-specific to the field and/or shaped by interdisciplinary collaborators. While social work journals may not fully capture the research performed by social workers *in toto*, or specifically omit contributions by non-social workers, it provides a sense of the research specific to the field, which is vetted by senior members of the field (i.e. most reviewers and members of the editorial board for these journals are social workers).³⁸

An alternative approach would have been to investigate articles authored by leading social work researchers, regardless of publication outlet. This strategy would help evaluate the quality of some of social work's best researchers, there are a few reasons for caution. While epistemically warranted, this approach may invite unwarranted social and professional blowback. Although criticism is a key element of scientific progress, it is not always welcome within the scientific community (e.g. Fiske, 2016a, 2016b; Sabeti, 2018), even if well-intentioned (see Heathers, 2017). A random sample of top social work researchers would by no means "single out" any given researcher, but could be perceived as such by any such researcher included in the sample; especially given that all materials, data, and analyses for this study will be made publicly available.

---

³⁸ Further, these journals are ostensibly read by social work academics, students, and practitioners. For example, the National Association of Social Workers touts its journal *Social Work* as being the premier journal for social workers. See <https://www.naswpress.org/publications/journals/sw.html>.

Finally, if the analyses performed in this study yield results indicating suboptimal reporting and practices, reform may be more likely to occur at the level of social work journals, whose standards can directly influence the quality of publication (e.g. Schelbe & Thyer, 2019; see also Rothman, 1998), rather than among sampled researchers. In conjunction with the limitations noted above, analyzing articles by journals was chosen as the most practical option. Future studies may devise strategies for analyzing the work, attitudes, and behaviors of individual social work researchers, and/or investigate the possibility of differences in reporting practices/statistical power, between “top” social work journals and “lower” ranked journals.

Storage and Availability of Data and Materials

All data and code are stored and made available to reviewers on the project’s Open Science Framework (OSF) page: <https://osf.io/45z3h/>. All data and code will be made public upon acceptance of the prospective manuscript(s) in a scholarly journal and shared under a CC0 license.

Sampling Plan and Sample

Articles were sampled from five social work journals (listed above) for the years 2014 through 2018 (a five-year timeframe). SSR, SW, and SWR each publish issues quarterly. RSWP published a varying number of issues per volume (between 6 and 8, of any type, during this study period). The BJSW publishes eight issues per volume. Additionally, the analytic sample was restricted to traditional issues (i.e. issues that were not special collections or supplements). To ease overrepresentation by BJSW and RSWP, during years where more than four traditional issues were published, the four issues from each with the *greatest* number of articles was included. The decision to exclude special issues and articles from “special sections” within traditional issues, was done to prevent the inclusion of articles that may deviate from the norm

within the field. For example, articles within special issues may be the outgrowth of an academic conference or symposium, or invited by a guest editor; and thus may not meet traditional peer review criteria.

Justification of sample timeline. The years 2014-2018 were chosen for two reasons. First, it marks the first volumes of social work research published since the contemporary “replication crisis” began (somewhere between 2011 and 2013; see generally Table 1 in Spellman, 2013).³⁹ Second, it is comparable in size and scope to contemporary meta-research projects (e.g. Camerer et al., 2018; Dumas-Mallet et al., 2017; Szucs & Ioannidis, 2017) and historical antecedents within social work and the social sciences (e.g. Brewer, 1972; Cohen, 1962; Crane, 1976; Orme & Combs-Orme, 1986; Orme & Tolman, 1986; Penick & Brewer, 1972).

Inclusion and Exclusion Criteria for Articles⁴⁰

Inclusion criteria. Articles were selected for inclusion if they had met the following criteria: 1) Are published in one of the five target journals (listed above), 2) are published between the years 2014 and 2018, 3) are written in the English language, 4) report on an empirical study, 5) employ and report the results from frequentist inferential statistical tests (see variables section), and 6) have full-text access.

Exclusion criteria. Articles were excluded from the study if they: 1) Are not published in one of the five target journals, 2) are published before 2014 or after 2018, 3) are in-

---

³⁹ Of course, as many have pointed out (e.g. Spellman, 2013), the identification of problems related to replication, statistical inference (e.g. power, NHST), openness/transparency in research, publication bias, etc. goes back decades (Greenwald, 1975; Rosenthal, 1979; Tversky & Kahneman, 1971), if not centuries (see Babbage, 1830 and the seminal text *Methodologie* by de Groot, 1960, 1969).

⁴⁰ Note: Articles included in analyses for Part 2 and Part 3 must meet the above criteria AND criteria necessary for their respective analyses (for Part 2 articles must have reported sample size; for Part 3 articles must have reported a significant p-value in the study’s primary analysis).

press/ahead of print articles, or are not assigned to a specific volume/issue, 4) are a part of a special issue or are designated to a “special section” within an issue, 5) are not written in the English language, 6) are any one of the following types of articles: a) editorial, b) commentary or opinion-piece, c) systematic review, d) meta-analysis, e) qualitative study, f) psychometric/validation study, g) book review, h) letter/letter-to-the-editor, 7) exclusively employ Bayesian, likelihood, or model-fit (e.g. AIC-based) statistical approaches to statistical inference, or 8) if the full-text of the article is not available.

Research Questions

Overarching research question. This study is seeking to answer the broad question, “What is the state of contemporary social work research in light of the modern ‘replication crisis’?” (see generally Yaffe, 2019). In the last decade, many popular claims in biomedical and social science have been brought under scrutiny (Bem et al., 2014; Button et al., 2013; Herrera-Perez et al., 2019; Kahneman, 2012; Ioannidis, 2005; Open Science Collaboration, 2015; van der Zee, Anaya, & Brown, 2017). Among the underlying contributions to this trend has been inappropriate statistical design/methodology (i.e. Button et al., 2013; Ioannidis, 2005; Meehl, 1990) and the use of Questionable Research Practices (QRPs) (e.g. Fiedler & Schwarz, 2016; Kerr, 1998). The field of social work has been relatively silent (or unaware) of the replication crisis and the problems that have contributed to it; save for Howard & Garland (2015) (and more recently Chartier et al., 2018; Yaffe, 2019).

Given that social work has a high prior probability of experiencing the same problems as neighboring fields (see Yaffe, 2019) and given that it is prudential to take precautionary measures to ensure rigorous research, even if the field is not in crisis (Howard & Garland, 2015, p. 194), then it is therefore critical that the social work research literature be audited and

appraised. This study will partly fill this need by reviewing a sub-section of contemporary social work research. By analyzing statistical reporting practices, estimated the statistical power of the published literature, and in investigating the evidential status of studies, a clearer picture of its overall utility and value will be ascertained. Doing so can help improve the rigor and value of social work research for informing policy and practice and lead to any necessary reform efforts within the profession (Hardwicke et al., 2019).

Research question #1. *How frequently and in what manner are common statistical concepts reported in the sampled articles?*

Sample size

- A. Was the sample size reported? Yes/no
- B. If yes, what is the reported sample size?
- C. If yes (for A) it reported as being predetermined? Yes/no
- D. Is that sample reported as being randomly drawn?⁴¹ Yes/no

Alpha

- A. Is alpha explicitly reported? Yes/no
- B. If yes, what level is it reported at? 0.01, 0.05, 0.10, etc.?
- C. If no (for A), is there a, single, implicit alpha level (e.g. 0.05)?
- D. If no (for C), are multiple alpha levels employed (e.g. * = Significant at 0.05, ** = Significant at 0.01, *** Significant at 0.001 or less)?

Power

- A. Was a power analysis explicitly reported? Yes/no
- B. If yes, was it an a priori, post-hoc, or sensitivity, or other type of power analysis?

---

⁴¹ Here I mean any type of probability sampling procedure.

- C. If yes (for A) what was the reported power level for the analyses? (Direct quote from text where possible).

P-values

- A. Are p-values reported? Yes/no
- B. If yes, are they reported in exact/precise form?

Confidence intervals

- A. Are confidence intervals reported? Yes/no
- B. If yes, at what coverage level are they reported? (e.g. 95%)

Effect size

- A. Is an observed effect size explicitly reported? Yes/no

Reported sample size from this section will help inform the analysis for Part 2.

Research question #2. *What percentage of studies in the sampled articles are adequately powered ($\text{pow} \geq .80$) to detect small, medium, and large effect sizes?*

Secondary question. *Has the mean statistical power of social work research increased since Orme & Combs-Orme, 1986?*

Research question #3. *Is there indication of bias (e.g. Questionable Research Practices) within published studies in the journal Research on Social Work Practice?*

Secondary question. *Is there indication that social work research in the journal Research on Social Work Practice has evidential value?*

Research Hypotheses

Formal hypothesis for Part 1.

H1-1. I hypothesized that sample size will be pre-specified in less than half of articles.

- *Hypothesis supported if:* Less than 50% of articles explicitly state the sample size was pre-specified.
- *Hypothesis disconfirmed if:* More than 50% of articles explicitly state that the sample size was pre-specified.

H1-2. I hypothesized that alpha will be explicitly reported in less than half of articles.

- *Hypothesis supported if:* Less than 50% of articles explicitly state an alpha level.
- *Hypothesis disconfirmed if:* More than 50% of articles explicitly state an alpha level.

H1-2a. I hypothesized that when alpha is specified, it will be done so at the 0.05 level, more than half the time.

- *Hypothesis supported if:* More than 50% of articles explicitly state an alpha level at 0.05.
- *Hypothesis disconfirmed if:* Less than 50% of articles explicitly state an alpha level at 0.05.

H1-2b. Of studies that do not explicitly state an alpha level, I hypothesized that more than half will use multiple cut-offs (e.g. * = Significant at 0.05, ** = Significant at 0.01, *** = Significant at 0.001 or less)?

- *Hypothesis supported if:* More than 50% of articles use multiple significance cut-offs.
- *Hypothesis disconfirmed if:* Less than 50% of articles use multiple significance cut-offs.

H1-3. I hypothesized that a formal *a priori* power analysis will be reported in less than 25% of articles.

- *Hypothesis supported if:* Less than 25% of articles report an a priori power analysis.
- *Hypothesis disconfirmed if:* More than 25% of articles report an a priori power analysis.

H1-4. I hypothesized that when reported, p-values will be reported in exact form less than half the time.

- *Hypothesis supported if:* Less than 50% of articles report exact p-values; when p-values are reported at all.
- *Hypothesis disconfirmed if:* More than 50% of articles report exact p-values; when p-values are reported at all.

H1-5. I hypothesized than more than half of articles will report confidence intervals; and do so at the 95% level.

- *Hypothesis supported if:* More than 50% of articles report confidence intervals at the 95% level.
- *Hypothesis disconfirmed if:* Less than 50% of articles report confidence intervals and/or do not do so at the 95% level.

H1-6. I hypothesized that an effect size will be reported in more than half of articles.

- *Hypothesis supported if:* More than 50% of articles report an effect size.
- *Hypothesis disconfirmed if:* Less than 50% of articles report an effect size.

Formal hypotheses for Part 2.

H2-1. I hypothesized that a less than half of articles will have adequate power ($\text{pow} \geq .80$) to detect a *small* effect size (where small, medium, and large refer to the contrastive designations described in Chapter 2, above).

- *Hypothesis supported if:* Less than 50% of articles have power of .80 or greater to detect a small effect size.
- *Hypothesis disconfirmed if:* More than 50% of articles have power of .80 or greater to detect a small effect size.

H2-2. I hypothesized that less than half of articles will have adequate power ($\text{pow} \geq .80$) to detect a *medium* effect size.

- *Hypothesis supported if:* Less than 50% of articles have power of .80 or greater to detect a medium effect size.
- *Hypothesis disconfirmed if:* More than 50% of articles have power of .80 or greater to detect a medium effect size.

H2-3. I hypothesized that less than 75% of articles will have adequate power ($\text{pow} \geq .80$) to detect a *large* effect size.

- *Hypothesis supported if:* Less than 75% of articles have power of .80 or greater to detect a large effect size.
- *Hypothesis disconfirmed if:* More than 75% of articles have power of .80 or greater to detect a large effect size.

H2-4. I hypothesize that there will be no more than a 5% (.05) increase in the mean statistical power of social work studies for each effect size, when compared to those reported by Orme & Combs-Orme (1986) (small = .31, medium = .76, large = .92).^{42,43}

---

⁴² This study was chosen as the reference point because of its focus on social work research, rather than research on social work education, as found in Orme & Tolman (1986).

⁴³ This hypothesis is associated with the second research question for Part 2.

- *Hypothesis supported if:* Mean power for articles in this study is between .31 and .36 for a small effect size; between .76 and .81 for a medium effect size, and between .92 and .97 for a large effect size.
- *Hypothesis disconfirmed if:* Mean power for articles in this study is greater than .36 for a small effect size; .81 for a medium effect size; and .97 for a large effect size.

Justification for hypotheses. I made these conjectures (H1-H4) based on the results of previous power surveys (see Table 2).⁴⁴ In Orme & Combs-Orme's (1986) survey of social work research, they found that the sample of articles had mean power values of .31, .76, and .92 for small, medium, and large effect sizes.

My hypotheses may seem conservative, given the technological ease in which power analyses can be performed and the methodological advances social work has experiences in the last thirty years (see Lindsey & Kirk, 1992 and contrast with Soydan, 2010). However, I believe institutional and intellectual incentives (Heesen, 2018; Ioannidis, 2012, 2016; Young, Ioannidis, & Al-Ubaydli, 2008) and inertia (Cohen, 1990; Feynman, 1974; Maxwell, 2004; Sedlmeier & Gigerenzer, 1989), as in other fields, has prevented power analysis from being more commonly utilized by social workers; thus tempering any expected gains in power due to improved planning, design, measurement, etc. (LeCroy, 1985; Orme, 1987).

---

⁴⁴ The mean statistical power for a small effect ranged from .14-.31; for medium effects a range of .48-.76; and .71-.92 for large effects.

Hypothesis for Part 3.

H3-1. I hypothesized that the significant p-values analyzed from *Research on Social Work Practice* will be left-skewed, which is indicative of Questionable Research Practices (e.g. p-hacking).

- *Hypothesis supported if:* The p-curve shows an obvious left-skew.
- *Hypothesis disconfirmed if:* The p-curve shows an obvious right-skew or if the p-curve appears flat.

H3-2. I also hypothesized that therefore, as a result of a left-skewed p-curve, studies analyzed in this p-curve will lack apparent evidential value.

- *Hypothesis supported if:* The p-curve shows an obvious left-skew.
- *Hypothesis disconfirmed if:* The p-curve shows an obvious right-skew or if the p-curve appears flat.

Justification for hypotheses. My hypotheses for Part 3 are not based on prior research on the topic, within social work. In fact, the p-curve analysis being performed here is likely the first of its kind within the profession. I merely presume that social work research is afflicted by QRPs, as is found in related research disciplines (see broadly: Ioannidis, 2005; Kerr, 1998, Head et al., 2015; Simonsohn et al., 2014).

Data Extraction, Analysis, and Reporting

Data extraction. All article information, for each issue of each journal included in the sampling frame, has been uploaded to the OSF page. Data from articles was extracted into CVS spreadsheets, which are also available on the OSF page.

Analysis of data. All analyses have been prespecified. Any unplanned analyses based on the variables above will be clearly noted as being exploratory.

Table 2.

Selective Review of Power Surveys (Sorted by Year)

Study (Year)	Journals Investigated	Years	Number of Articles Analyzed	Mean Effect Size (S, M, L)
Cohen (1962)	Journal of Abnormal and Social Psychology	1960	70	.18, .48, .83
Brewer (1972)	American Educational Research Journal	Nov 1959-May 1971	47	.14, .58, .78
Penick & Brewer (1972)	Journal of Research in Science Teaching	Vol 6, 7 1969-1970	66	.22, .71, .87
Chase & Tucker (1975)	Nine Journals: Speech/Communication	1973	46	.18, .52, .79
Chase & Chase (1976)	Journal of Applied Psychology	1974	121	.25, .67, .86
Orme & Combs-Orme (1986)	Social Work Research & Abstracts	1977-1984	79	.31, .76, .92
Orme & Tolman (1986)	Journal of Social Work Education	1976-Sp. 1985	64	.20, .68, .88
Sedlmeier & Gigerenzer (1989)	Journal of Abnormal Psychology	1984	54 (64 experiments)	.21, .50, .84
Szucs & Ioannidis (2017)	18 Journals: Cognitive Neuroscience & Psychology	Jan 2011 to Aug 2014	3,801 (Across the 3 domains)	.17, .49, .71

Note: Studies by Crane (1976), Button et al. (2013), Dumas-Mallet et al. (2017), Ioannidis et al. (2017) are excluded from the table above, as they have a different unit of analysis and/or report different outcomes than those proposed here and reviewed in this table.

Part 1 Analytic Strategy. This survey of statistical reporting examined the frequency that each statistical concept is reported. Descriptive statistics are reported using the open-source statistical software JASP (<https://jasp-stats.org/>) version 0.11.1.0

Part 2 Analytic Strategy. This power survey relies on results from Part 1. Using statistical software (e.g. G*Power) (Faul, Erdfelder, Lang, & Buchner, 2007) and referencing common power tables as necessary (e.g. Cohen, 1988; Murphy & Myers, 1998), power for each primary analysis, was calculated based on the sample size for each test, to detect a standardized small, medium, and large effect size (SML). Where sample size was unclear, the overall study sample size was used. SML effect sizes were defined using commonly accepted benchmarks,⁴⁵ which vary depending on the statistical test employed. An alpha-level of 0.05⁴⁶ and non-directional tests were used where applicable.

The overall unit of analysis for this survey was articles (as in previous power surveys; e.g. Cohen, 1962; Orme & Combs-Orme, 1986; Orme & Tolman, 1986; Sedlmeier & Gigerenzer, 1989). Average statistical power for articles from each journal, and the mean and median power for all articles in the sample for each effect size was analyzed.

Only *primary analyses* were included in the power survey. Where one inferential statistical test was employed, that was designated as the primary analysis. Where multiple analyses were employed, the analysis that directly related to the primary research question was considered as the primary analysis. Tests of assumptions (e.g. tests of homogeneity), manipulation checks, omnibus tests, unplanned or post-hoc comparisons, were excluded.

---

⁴⁵ Here I follow Cohen's (1988) canonical benchmarks for interpreting S/M/L ESs; though I acknowledge the limitations of this approach in general (e.g. Correll, Mellinger, McClelland, & Judd, 2020)

⁴⁶ A 0.05 alpha level is a reasonable standard for this analysis, as it is the conventional cut-off for significance in the social sciences; despite efforts reform (see Benjamini et al., 2018; Lakens et al., 2018).

Nonparametric tests were included if there is a clear method for assessing its power. Otherwise, as other researchers (e.g. Cohen, 1962, p. 149) have done, nonparametric tests were estimated using the associated parametric equivalent. All cases were noted where the article was excluded from analysis either due to the reasons listed above or if there is not a clear method for estimating power.

Part 3 Analytic Strategy. The p-curve analysis relies on results from Part 1. Version. 4.06 of the p-curve application (found at: <http://www.p-curve.com/app4/>) was used. All results are reported in full (e.g. binomial and continuous tests for evidential value; power of tests in analysis); including full- and half- p-curves. Study results that were ambiguous, with regard to whether they should be included in the p-curve analysis are noted and recorded (see column O in the spreadsheet). P-curve robustness checks were performed (with and without those results), so that readers can determine the extent to which study results depend on ambiguous cases (see Simonsohn et al., 2014, p. 539).

The p-curve technique is used to determine the shape of the distribution of significant p-values (p-values between 0.01 and 0.05) for articles published in *Research on Social Work Practice* for years 2014-2018. This analysis helps us understand whether or not there are bias (e.g. Questionable Research Practices) present in the studies selected and if not, whether or not there is purported evidential value.

Study selection rule (for the p-curve analysis). Articles meeting inclusion/exclusion criteria (see above) in the journal *Research on Social Work Practice* for the years 2014-2018 that contain *significant* p-values in their respective primary analyses.

Justification of selection rule. As Simmons, Nelson, and Simonsohn (2017) note, a study selection rule is provided:

“To ensure that studies are not selected in a biased manner, and more generally to help readers and reviewers detect possible errors, the set of studies included in *p*-curve must be determined by a *predetermined* rule. The rule, in turn, should be concrete and precise enough that an independent set of researchers following the rule would generate the same, or virtually the same, set of studies.”

The selection rule chosen above meets these criteria: It has been chosen in advance (predetermined) and is explicit and precise. The justification for the content of the rule, however, remains open to discussion. Previous meta-research studies (e.g. Motyl et al., 2017) have used scholarly journals as the unit of analysis; though journals are not the only thing that can be analyzed using a *p*-curve (see Simohnson et al., 2014; as well as Carbine & Larson, 2019; Cuddy et al., 2018; Lakens, 2017). I have chosen this journal for analysis because it is widely regarded as one of the field’s top journals, it has a long publication record, and it has an explicit focus on quantitative studies (Thyer, 2015a). As this is the first known *p*-curve analysis in social work, it was determined that a single journal was an appropriate starting point to pilot such an analysis.

Conflicts of Interest

I have been a reviewer and Consulting Editor for the journal *Social Work* since 2019. I have been an ad hoc reviewer for *Research on Social Work Practice* (2012) and as of Spring 2020 have served on its Editorial Board. There are no other known personal, professional, or financial conflicts of interest in relation to this work. An ICMJE Disclosure Form for Potential Conflicts of Interest is available at study OSF page: <https://osf.io/45z3h/>

CHAPTER 4

RESULTS

There were 1,906 total articles eligible for inclusion based on the sampling frame (see Table 3).⁴⁷ The *British Journal of Social Work* comprised 847 of these (44.4%), *Research on Social Work Practice* 466 (24.4%), *Social Work* 285 (15.0%), *Social Service Review* 178 (9.3%) and *Social Work Research* 130 (6.8%). Of these, 411 (21.6%) articles were excluded as a result of the predetermined sampling plan (i.e. to prevent overrepresentation of BJSW and RSWP content within the sample). 1,185 (62.2 %) were removed for meeting specific exclusion criteria. The most common type of articles excluded were book reviews (n = 241, 12.6%), non-empirical articles (n = 197, 10.3%), and qualitative research (n = 179, 9.4%). 310 articles (16.3%) met inclusion criteria for the present study. For a list of all exclusions, by both type and frequency, see Table 4.

Table 3.

<i>Sampling Frame by Journal Frequency</i>			(N = 1,906)
Journal	N	% of Total	Cumulative %
British Journal of Social Work	847	44.4 %	44.4 %
Research on Social Work Practice	466	24.4 %	68.9 %
Social Service Review	178	9.3 %	78.2 %
Social Work	285	15.0 %	93.2 %
Social Work Research	130	6.8 %	100.0 %

⁴⁷ This excludes general journal materials (e.g. a journal's table of contents, call for papers, editorial board list, etc.)

Table 4.

Reason for Exclusion by Type and Frequency

(N = 1,906)

Included (N = No, Y = Yes); Reason?	N	% of Total	Cumulative %
N; *Excluded From Sample via Sampling Plan	411	21.6 %	21.6 %
N; Book Review	241	12.6 %	34.2 %
N; Commentary	69	3.6 %	37.8 %
N; Content Analysis	5	0.3 %	38.1 %
N; Descriptive Statistics Only	39	2.0 %	40.1 %
N; Editorial	75	3.9 %	44.1 %
N; Ethnography	20	1.0 %	45.1 %
N; Guest Editorial	6	0.3 %	45.4 %
N; Invited Article	29	1.5 %	47.0 %
N; Letter	13	0.7 %	47.6 %
N; Meta-Analysis	4	0.2 %	47.8 %
N; Misc.	2	0.1 %	48.0 %
N; Misc. (Case Study)	4	0.2 %	48.2 %
N; Misc. (Classification Study)	1	0.1 %	48.2 %
N; Misc. (Forecasting Study)	1	0.1 %	48.3 %
N; Misc. (Science Mapping)	1	0.1 %	48.3 %
N; Misc. (Single-Subject Design)	3	0.2 %	48.5 %
N; Non-Frequentist Approach	1	0.1 %	48.5 %
N; Non-Empirical Article	197	10.3 %	58.9 %
N; Psychometric Study	59	3.1 %	62.0 %
N; Qualitative Research	179	9.4 %	71.4 %
N; Review	20	1.0 %	72.4 %
N; Special Issue	115	6.0 %	78.4 %
N; Special Section	84	4.4 %	82.8 %
N; Systematic Review	17	0.9 %	83.7 %
Y; Empirical Article	310	16.3 %	100.0 %

Results from Part 1: Statistical Reporting

For complete information about results from Part 1, the reader is directed to Tables 6-10.

Sampling. Nearly all articles reported sample size 97.7% (303 out of 310). Sample size was *not* explicitly said to have been predetermined in 93.2% of articles. This latter result supports Hypothesis (H) 1-1 ($< 50\%$).⁴⁸ Only 11% of studies mentioned some form of random sampling.⁴⁹

Alpha. Alpha was *not* explicitly reported in 86.1% of articles. This result supports H1-2 ($< 50\%$). 12.9% of articles explicitly reported using an alpha level of 0.05; this corresponds to 88% of cases *when* it is explicit, thus supporting H1-2a ($> 50\%$). One article explicitly used a more stringent threshold (i.e. alpha set at 0.01). Two others used a more liberal threshold (alpha set at 0.10). 62.9% of articles appeared to use multiple alpha levels; as evidenced by inclusion of asterisks at different thresholds (e.g. * = Significant at 0.05, ** = Significant at 0.01, *** Significant at 0.001), which supports H1-2b ($> 50\%$). Finally, there were 19 cases (6.1%) where some form of alpha-level correction (e.g. Bonferroni correction test) was used.⁵⁰ In these cases, the original alpha-level was recorded, if implicit or explicit, as alpha is a pre-experimental concept; whereas alpha-level adjustment generally occurs after testing has occurred.

Power. Researchers explicitly reported performing a power analysis (of any type) in only 7.4% of cases ($n = 23$). The most common form was the a priori power analysis; which was mentioned in 16 articles (5.2% of the sample); supporting H1-3 ($< 25\%$). This was followed by post-hoc power analysis in 1.9% of articles, and 1 case of a sensitivity analysis (0.3%). Specific

---

⁴⁸ Hypotheses are explicitly described above in Chapter 3.

⁴⁹ In order to provide as conservative of an estimate as possible, any type of sampling using randomization (e.g. simple random sampling, stratified random sampling, etc.) was counted here if mentioned.

⁵⁰ Note: This variable was not included in the methodology prior to data collection, but arose as a variable of interest during coding. It should be noted as being exploratory in nature.

details about the reported power analyses in each individual article are not discussed here, but can be found in the supplemental materials.

P-Values. P-values were almost universally used as a part of statistical analysis, being reported in 96.8% of articles. However only 29% of articles reported p-values in precise (i.e. exact) form; supporting H1-4 (< 50%).

Confidence intervals. Confidence intervals were less commonly used than p-values, being reported in 36.5% of articles; disconfirming H1-5 (> 50%). The most common type were intervals with a 95% coverage rate (31.3% of total articles). A handful of studies reported intervals at the 90% level (2.6%, n = 8). Additionally, 8 articles (2.6%) reported using confidence intervals, but did not specify the level which they were at. While it is presumed that this was done at the 95% level, no such assumptions were made here.

Effect sizes. Effect sizes were explicitly reported in the results section or tables in a little more than half of articles (55.2%); supporting H1-6 (> 50%).

Table 5.

Journal Frequencies

(N = 310)

Journal	N	% of Total	Cumulative %
British Journal of Social Work	76	24.5 %	24.5 %
Research on Social Work Practice	79	25.5 %	50.0 %
Social Service Review	43	13.9 %	63.9 %
Social Work	47	15.2 %	79.1 %
Social Work Research	65	21.0 %	100.0 %

Table 6.

Frequencies for Variables

(N = 310)

Variable	Reported? (N)	% of Total
Sample Size Reported	No = 7 Yes = 303	2.3% 97.7%
Sample Size Predetermined	No = 289 Yes = 14 N/A = 7	93.2% 4.5% 2.3%
Random Sample	No = 269 Yes = 34 N/A = 7	86.8% 11.0% 2.2%
Explicit Alpha Reported	No = 267 Yes = 43	86.1% 13.9%
Multiple Alpha Levels	No = 115 Yes = 195	37.1% 62.9%
Power Analysis	No = 287 Yes = 23	92.6% 7.4%
P-Values	No = 10 Yes = 300	3.2% 96.8%
Precise (P-Values)	No = 210 Yes = 90 N/A = 10	67.7% 29.0% 3.2%
Confidence Intervals	No = 197 Yes = 113	63.5% 36.5%
Effect Size	No = 139 Yes = 171	44.8% 55.2%
*Bonferroni or Other Alpha- Level Adjust.	No = 291 Yes = 19	93.9% 6.1%

N/A = Not Applicable

*This variable was not pre-specified in the analytic plan.

Table 7.

Frequencies of Explicit Alpha Level

(N = 310)

Alpha Level	N	% of Total	Cumulative %
0.01	1	0.3 %	0.3 %
0.05	40	12.9 %	13.2 %
0.1	2	0.6 %	13.9 %
0	267	86.1 %	100.0 %

0 = Not Explicit

Table 8.

Frequencies of Implicit Alpha Level

(N = 310)

Alpha Level	N	% of Total	Cumulative %
0	213	68.7 %	68.7 %
0.05	53	17.1 %	85.8 %
0.1	1	0.3 %	86.1 %
N/A	43	13.9 %	100.0 %

0 = Not Implicit; N/A = Not Applicable

Table 9.

Frequencies of Type of Power Analysis

(N = 310)

Power Analysis	N	% of Total	Cumulative %
A Priori	16	5.2 %	5.2 %
Post Hoc	6	1.9 %	7.1 %
Sensitivity	1	0.3 %	7.4 %
N/A	287	92.6 %	100.0 %

N/A = Not Applicable

Table 10.

Frequencies of Type of Confidence Interval

(N = 310)

Interval Type	N	% of Total	Cumulative %
0	8	2.6 %	2.6 %
90%	8	2.6 %	5.2 %
95%	97	31.3 %	36.5 %
N/A	197	63.5 %	100.0 %

0 = Confidence Interval reported, but type not specified. N/A = Not Applicable

Results from Part 2: Power Survey

Overall survey. Of the 310 articles used in Part 1, 298 were eligible for inclusion in the power survey performed in Part 2. 12 articles were removed for not clearly reporting (or omitting altogether) study sample size. Of the 298 eligible articles, 207 were included in the full power survey. Exclusions were made where there was insufficient information about a test to compute power, the test used in the article met exclusion criteria, and/or there was no known method for computing power (see the supplemental materials for more details on how these decisions were made on a case-by-case basis).

Alpha was set at 0.05; and a two-tailed test was employed where possible. The mean statistical power (see Table 11) for the 207 included articles was 57% for small effects, 88% for medium effects, and 95% for large effects. 61% of articles did not have adequate power (i.e. power of .80) to detect a small effect, 19% did not to detect a medium effect, and 7% did not to detect a large effect. These results support H2-1 (> 50% for small effect size) and disconfirm H2-2 (> 50% for medium effect size) and H2-3 (> 25% for large effect size). If a higher power level of .90 is chosen, as is sometimes recommended in clinical trial research (Center for Drug

Evaluation and Research, 2001, p. 28), power is universally lower, with 67% of articles do not have adequate power to detect a small effect, 23% did not to detect a medium effect, and 11% did not to detect a large effect.

Robustness test. Given the numerous assumptions made in the power survey performed above (see Chapter 5), a post-hoc robustness test was performed. Power for small, medium, and large effect sizes were computed for the 96 (of 207) articles for which the power calculation was straightforward. That is, articles where the primary analysis could be run in G*Power, without converting it to its closest statistical equivalent which has effect size benchmarks (see Table 12 for full details). Again, alpha was set at 0.05; and a two-tailed test was employed where possible. The robustness test yielded similar but more conservative results than the overall survey.

The mean statistical power was 43% for small effects, 80% for medium effects, and 92% for large effects. 80% of articles did not have adequate power (i.e. power of .80) to detect a small effect, 30% did not to detect a medium effect, and 13% did not to detect a large effect. Though the robustness test was not pre-specified, to the extent that it is sound analytically, the results again support H2-1 and disconfirm H2-2 and H2-3. If a higher power level of .90 is chosen, 86% of articles do not have adequate power to detect a small effect, 35% did not to detect a medium effect, and 18% did not to detect a large effect.

Comparison with Orme & Combes-Orme (1986). The overall power survey reported mean power of 57% for small effects, 88% for medium effects, and 95% for large effects. Each of these values is larger than those reported in Orme and Combes-Orme (1986) (31%, 76%, and 92% respectively). Since the mean power for small and medium effects was outside the bounds predicted, these results disconfirm the *conjunctive* hypothesis presented in H2-4. Notably, for the values for the robustness test, there was greater power in the current sample for small effects

(43%) and medium effects (80%), and the same mean power for large effects (i.e. 92%); results that would likewise disconfirm H2-4.

Comparisons between journals. *Social Service Review* had the highest mean power for small, medium, and large effects (93%, 99%, and 99% respectively). *Research on Social Work Practice* had the lowest mean power for each effect size (37%, 77%, and 91%). These values are for the overall power survey. Figure 3 offers a visualization of these comparisons.

Results from Part 3: P-Curve

The p-curve analysis was performed on articles from the journal *Research on Social Work Practice*. There were 79 eligible articles in the sample. Of these one was excluded for not reporting p-values. Of the 78 articles, 39 had results which were able to be included in the analysis. Reasons for exclusion included insufficient reporting of test statistics in the original article and non-significance of findings; factors automatically excluding a result from analysis in the software.

Primary curve. The primary p-curve analysis (Figure 4) yielded a right-skewed curve, indicating evidential value for the included set of results, and no evidence of bias. Though 39 results were input into the analysis, re-computation of test statistics identified two as being non-significant; and thus both were excluded. Binomial ($p = .0001$) and continuous tests (Full p-curve: $z = -13.32$, $p < 0.0001$; Half p-curve: $z = -13.61$, $p < 0.0001$) indicate that the results included in the p-curve analysis have evidential value. Similarly, the inadequacy test (i.e. the null of 33% power test) *did not* indicate that evidential value was inadequate or absent; as shown by the results of the associated Binomial ($p = .9493$) and continuous tests (Full p-curve: $z = 8.65$, $p > .9999$; Half p-curve: $z = 13.83$, $p > .9999$). Visual inspection of the p-curve (Figure 4) supports these interpretations.

Table 11.

Survey of Power from Studies Published Across Five Social Work Journals Between 2014-2018 (N = 207)

Power	<u>Small Effects</u>		<u>Medium Effects</u>		<u>Large Effects</u>	
	Frequency	Cumulative %	Frequency	Cumulative %	Frequency	Cumulative %
.99	50	100	137	100	171	100
.95-.98	10	76	12	33	10	17
.90-.94	9	71	8	27	4	13
.80-.89	12	67	9	23	7	11
.70-.79	9	61	7	19	7	7
.60-.69	12	57	8	16	2	4
.50-.59	8	51	5	12	1	3
.40-.49	10	47	5	10	1	2
.30-.39	15	42	8	8	1	2
.20-.29	28	35	4	4	1	1
.10-.19	31	21	2	2	2	1
.05-.09	13	6	2	1	—	0
N	207	—	207	—	207	—
M		57		88		95
Mdn		56		99		99
SD		35		22		13
Q1		23		93		99
Q3		56		99		99

Table 12.

Robustness Test of Power from Studies Published Across Five Social Work Journals Between 2014-2018 (N = 96)

Power	<u>Small Effects</u>		<u>Medium Effects</u>		<u>Large Effects</u>	
	Frequency	Cumulative %	Frequency	Cumulative %	Frequency	Cumulative %
.99	6	100	50	100	67	100
.95-.98	5	93	5	46	7	29
.90-.94	2	88	6	41	4	22
.80-.89	6	86	5	35	5	18
.70-.79	5	80	3	30	6	13
.60-.69	6	75	5	27	1	7
.50-.59	4	69	5	22	1	6
.40-.49	7	65	3	17	1	5
.30-.39	8	58	6	14	1	4
.20-.29	15	50	4	8	1	3
.10-.19	20	34	2	4	2	2
.05-.09	12	13	2	2	—	0
N	96	—	96	—	96	—
M		43		80		92
Mdn		31		99		99
SD		32		28		17
Q1		15		64		96
Q3		70		99		99

Figure 3.

Power for Each Effect Size by Journal

Robustness test. The robustness test p-curve analysis (Figure 5), was computed using 39 test results, and also yielded a right-skewed curve, indicating evidential value for the included set of results, and no evidence of bias. Binomial ($p = .0025$) and continuous tests (Full p-curve: $z = -11.78$, $p < 0.0001$; Half $z = -12.94$, $p < 0.0001$) indicate that the results included in the p-curve analysis have evidential value. Similarly, the inadequacy test (i.e. the null of 33% power test) *did not* indicate that evidential value was inadequate or absent; as shown by the results of the associated Binomial ($p = .7194$) and continuous tests (Full p-curve: $z = 7.22$, $p > .9999$; Half $z = 13.01$, $p > .9999$).

Note: The observed p -curve includes 37 statistically significant ($p < .05$) results, of which 30 are $p < .025$. There were 2 additional results entered but excluded from p -curve because they were $p > .05$.

Figure 4.

P-Curve (Primary)

Note: The observed p -curve includes 38 statistically significant ($p < .05$) results, of which 28 are $p < .025$. There was one additional result entered but excluded from p -curve because it was $p > .05$.

Figure 5.

P-Curve (Robustness Test)

Both the primary curve and robustness test disconfirm H3-1 (left-skew indicating presence of QRPs/bias) and H3-2 (left-skew indicating absence of evidential value). Visual inspection of the p -curve (Figure 5) supports these interpretations.

CHAPTER 5

DISCUSSION

Summary and Interpretation of Findings

This study examined articles published in five social work journals between 2014 and 2018. The aim was to assess the status of contemporary social work research in light of the replication crisis; and specifically, to assess issues surrounding statistical practice. The analyses comprised three distinct, but interrelated parts. Part 1 used descriptive statistics to describe how key statistical concepts were reported in articles. Part 2 calculated power for small, medium, and large effect sizes for each article and reported descriptive statistics for the overall sample and for individual journals. Part 3 utilized a technique called the p-curve to investigate whether articles published in *Research on Social Work Practice* had indications of bias or evidential value. These analyses were contextualized with a discussion about the causes and proposed solutions to the replication crisis and foundational issues in statistical inference.

The results from Part 1 supported seven of eight pre-specified hypotheses. In all, articles frequently failed to report having engaged in oft recommended pre-study design practices (e.g. specification of an alpha level, performing an a priori power analysis, justifying the sample size, etc.). Additionally, it was not uncommon for articles to omit common statistical results (e.g. precise p-values, confidence intervals, effect sizes). Though researchers are constrained with regards to certain features of statistical practice (e.g. how much power their tests have), there is little reason to not fully report the full findings of a study. This includes both pre-study analytic decisions *and* all aspects of performed statistical tests.⁵¹ The consequences for omitting such

---

⁵¹ This would also include making clear when certain actions have *not* been taken (e.g. randomization, random sampling, power analysis, etc.), rather than merely omitting discussion of them.

information are clear (see Chapter 1; Altman, 1980, 1994) and may be interpreted as violating professional ethical standards (NASW, 2017).⁵²

The results of this study do not yield reasons for these discrepancies. However, returning to our earlier discussion in Chapter 2, it could be posited that, in part, researchers have continued to engage in the norm of Null Hypothesis Significance Testing, despite *longstanding* criticisms (Gigerenzer, 1993; Meehl, 1978; Morrison & Henkel, 1970), by virtue of professional and scholarly “inertia” (Meehl, 1986, p. 331). The dearth of pre-specified alpha levels, power analysis, and precisely reported p-values; along with high frequency in use of multiple alpha levels, supports this interpretation. If NHST is the norm, there may be little pressure to report statistical concepts that have minimal bearing on a manuscript’s likelihood of publication.⁵³ In other words, if educators, journal editors and reviewers, and others professional stakeholders fail to encourage, promote, and/or impose a set of research practices, there will be little incentive to integrate them into practice.⁵⁴ Another explanation could be that researchers did not find omitted concepts useful for interpretation of the study results or did not understand their utility (e.g. Gigerenzer, 2004; Hoekstra et al., 2014).

---

⁵² This, of course, goes without considering whether use of inferential methods is even justified. If certain assumptions are not met (e.g. random sampling; which was not commonly reported in this study), the conclusions drawn from such statistical tests may be at a minimum, questionable (Berk & Freedman, 2003, p. 235).

⁵³ If in fact, authors, journal editors, and reviewers place high value on significance/non-significance, then there is little reason to report precise p-values or to adhere to a single alpha level. Similarly, if there is no norm of performing power analysis or reporting confidence intervals, then what might reporting it add to the manuscript’s chances of getting published? In such a scenario (i.e. where NHST is the dominance form of practice), it be would important to acknowledge that statistical concepts can become embedded with epistemic value and instrumental value; features which may exist distinctly, in harmony, or in conflict, depending on the broader scientific context.

⁵⁴ Meehl (1986) offers a bold solution. “Editors ought to arbitrarily reject such papers [i.e. where data reporting is incomplete], so that members of the profession would come to take it for granted that measures of overlap and effect size must be presented.” (p. 334). See also Rothman (1998) for a discussion of one journal’s attempt to shift the ingrained practices, as well as Osborne (2008).

The results from Part 2 were mixed. The hypothesis that most articles would not have adequate statistical power to detect a small effect size was supported, while the corresponding hypotheses about medium and large effects was disconfirmed. Likewise, the conjunctive hypothesis contrasting mean power for the sample to the only comparable study of power on social work by Orme and Combs-Orme (1986) was disconfirmed.

There are a few possible explanations for these results; which are somewhat of an aberration compared to previous power surveys (Table 2; see also Sedlmeier & Gigerenzer, 1989). First, overall statistical power, of social work studies, may *actually* be increasing over time. This would be a welcomed occurrence and would mark the continued scientific maturation of the profession; as has arguably happened throughout its history (see Chapter 1). Second, researchers may be taking advantage of new techniques and opportunities, that prior generations of social science (and social work) researchers did not have. In the sample analyzed here, it was not uncommon to see repeated-measures designs, hierarchical/multilevel models, and use of online sampling. Each of these tools can help increase the overall power of a test. Third, the results of this power survey may have been inflated. As will be discussed in the limitations section below, every effort has been made to interpret the results analyzed here charitably. In doing so, mean power may have been higher than otherwise calculated. While none of these explanations should be considered definitive, they are all plausible. Neither should they be considered mutually exclusive, as they are perfectly compatible with one another.

To the extent that the results of Part 2 are accurate (or even overestimated), there is still room for improvement in social work research. If 61% of social work articles do not, in fact, have adequate power to detect a small effect size, then as researchers, we are missing small effects more than half the time (on average). Additionally, almost half of studies (47 of 96) did

not have power above .30 to detect small effects; which means that, at best, in the long run such studies would only find a true effect (when one exists) every 3 out of 10 times. Again, such findings are not uncommon in the social sciences generally. However, given the high-stakes nature of many problems studied in social work research and the often high vulnerability of those studied, such figures are cause for concern, as many effects in social work research may ultimately turn out to be small; thus rendering underpowered research wasteful and potentially harmful.

Finally, the results from Part 3 disconfirm the hypotheses positing that articles from RSWP would purportedly lack evidential value or contain evidence of bias. These results are partially incongruent with the conjecture that the profession is in the midst of a crisis. There are a few explanations for these results. First, RSWP may differ from other social work journals. As it was the only journal to have been analyzed using the p-curve technique, we cannot be sure that this is the case. However, it is a plausible explanation. RSWP (Thyer, 1991, 2015b) has long emphasized that research published in the journal will adhere to strict (or rather, contemporary) research standards. Second, as will be discussed in the limitations section below, the choice of results to be input into the p-curve analysis may be, to some extent, idiosyncratic. As will be noted below, despite having created a pre-analysis plan, there was still a high degree of “researcher degrees of freedom” when it came to analytical decisions (Simmons et al., 2011). Third, the results may simply be misleading. The p-curve is not a definitive test of evidential value or bias (Ederfield & Heck, 2019). It can only provide an indication that either is present in the set of results. It may very well be the case that the data used contained undetected levels of bias (e.g. via QRPs).

Personal considerations. It would be remiss to leave unacknowledged my own (current) beliefs about this study's findings. Presenting them here will hopefully make clearer my own role in this study, above and beyond the obvious analytic and methodological decisions strewn throughout. Additionally, presenting them here (hopefully) will add a level of transparency and allow me (if nothing else) to update them in the future.

The hypotheses listed in Part 2 were strongly informed by prior research on the topic (see Table 2); particularly Orme and Orme-Combs (1986). The hypotheses for Parts 1 and 3, however, were largely informed by my own reading of the literature on statistics, the broader replication crisis, and social work research generally. My prior probabilities⁵⁵ were high that the statistical reporting in social work articles would be poor. Likewise, I had high certainty that the p-curve analysis of RSWP would yield a left-skewed or flat curve. Having now seen the data and considered explanations for it, it would seem appropriate to consider how my beliefs may be updated in light of new evidence.⁵⁶

The results from Part 1 supported the proposed hypotheses and align with parallel literature on statistical reporting (e.g. Weissgerber et al., 2018). I do not see any reason to doubt that statistical reporting practices within social work research need improvement. The results from Part 2 were by and large surprising to me. I expected mean power levels to be lower than was reported for small and medium effects. The mean power levels for large effect was not dramatically different than my own expectations. While I personally believe these results are inflated (see discussion above; and in the limitations section below), I am now more prone to believe I had underestimated, to some extent, statistical power within the sampled literature.

---

⁵⁵ If the reader will permit a brief “whiff” of Bayesianism.

⁵⁶ ***To be clear, this personal musing is distinct, yet complementary, from the conclusions drawn above.

Finally, the results from Part 3 disconfirmed my hypotheses. However, there seem to be plausible explanations abound. For instance, I would not be surprised if RSWP contained a higher degree of evidential value and lower degree of QRPs than other journals in the field.

Limitations of the Study

"[T]he idea is to try to give *all* of the information to help others to judge the value of your contribution; not just the information that leads to judgment in one particular direction or another."

Feynman, 1974, p. 11

The current study contains important limitations which must be acknowledged and made explicit. Only then can the utility of this work be properly ascertained (Ross & Bibler Zaisi, 2019). While every attempt has been made to catalogue all possible limitations, it is noted that still others may apply. I start by noting some general limitations about the study. Then I detail limitations for Parts 1-3.

First, regards the *reliability* of the results. The coding of this study was performed by a single researcher. Though every precaution has been taken to ensure that information gathered from articles was accurately coded, a second rater would permit the calculation of interrater reliability for each variable in the study. Unreliability of coding may undermine the conclusions drawn from the study.

Second, the *generalizability* of the results is unknown. The current study analyzed a subsection of social work journals over a five-year time period (i.e. 2014-2018). It may be the case that there are inherent differences between social work journals; whether across content area (e.g. generalist vs. subject-specific social work journals) or ranking (e.g. journals with high vs.

low impact factors). Further research is necessary to determine the extent to which the findings generalize across the field. This might entail, for example, selecting additional subsections of journals (e.g. in the top 10), and reviewing articles published between 2014-2018, and/or analyzing how reporting, power, and evidential value have changed over a longer period of time (e.g. 1990-2020).

Additionally, the approach used here does not permit generalizations about social work researchers themselves or social work research *in toto*, but rather about *research* published in social work journals.⁵⁷ It may be the case that some social work researchers publish, primarily or even exclusively, in interdisciplinary, area-specific journals (e.g. journals specifically about medicine, behavioral health, criminology, gerontology, etc.). These researchers could be said to still contribute to the overall corpus of social work knowledge, even if their work is not being published in social work-specific journals. Similarly, researchers who are not formally trained as social workers are not precluded (on the basis of their professional identity alone), from publishing in the journals analyzed in this study; or social work journals more broadly. Therefore, the results presented here can only be said to apply to the research published *in* social work journals.

Limitations of Part 1. The most important limitation for Part 1 pertains to reliability. This basic issue is described above. However, here, it is noted that unreliability may disproportionately impact the coding of some variables more than other. Whether sample size, for example, is reported, is a fairly straightforward question. Other concepts are not so clear-cut. Discerning whether an article reports an effect size (i.e. if it is not explicitly discussed by the

---

⁵⁷ This doesn't take into account the assessment of social work research that is conducted, but remains unpublished; whether because it is rejected or abandoned, or turned into white papers, government reports, etc.

authors) can be a difficult for even the skilled quantitative researcher. This owes in part to the vast number of indices that comprise the effect size concept.⁵⁸ Coding of other variables may similarly require making inferential leaps from the published results; such as when confidence intervals are provided, but their coverage rate (e.g. 90%, 95%) is not specified. This limitation is, of course, speculative and can only be investigated with the addition of a second rater.

Limitations of Part 2. Several limitations impact the findings of the power survey performed in Part 2. First and foremost, a number of articles (see results) were excluded from the analysis. Some were excluded because they contained uncommon statistical tests (e.g. biometric analysis; classification tree models), for which there was no known method for computing power and/or were not supported by the G*Power software. Others included complex statistical analyses for which power analysis could, in principle, be performed but for which either the data would need to be accessible and/or simulation studies would need to have been performed (e.g. generalized linear mixed models; structural equation models). Both actions are beyond the scope of the current study.

Still further, some statistical tests did not have known effect size benchmarks (see Cohen, 1988). In such cases, a note was made, and the analysis was converted to its closest statistical equivalent for which there were benchmarks. For example, logistic regressions were converted into multiple regression tests. The rationale for this decision and potential impact are described above. Even if effect size benchmarks were available for each test, this approach can be considered a limitation of its own (Ellis, 2010, pp. 41-43). Cohen (1988) readily admits that the designations of “small”, “medium”, and “large” are merely “conventions”; and further that they

---

⁵⁸ The type of effect size used can vary wildly from researcher to researcher. This is due in part to training and familiarity, the type of data used, analysis, and software used, and conventions within the field of study.

are “arbitrary” to some extent (p. 12). The decision to use benchmarks risks lapsing into the type of “mindless” statistical thinking that critics have pointed out regarding NHST and the use of p-values generally (Gigerenzer, 2004).⁵⁹ With that being said, I believe the utility of such an approach (e.g. providing comparisons with past power analyses) outweighs the limitations of using benchmarks.

When power analyses were performed for individual articles, it was not always clear the number of predictors or covariates used (e.g. multiple regressions, analysis of covariance). Therefore, conservative estimates of predictors (e.g. 2 predictors) and covariates (e.g. 1 covariate) were used throughout. A minimum of 2 time points was specified for repeated-measures analyses, if the number could not be discerned from the published article. When articles did not report the number of participants/observations in the analysis, the full sample size was used for the power analysis. Where participants/observations per group needed to be specified, it was assumed that they were equally distributed. These decisions likely *overestimated* the overall power of studies presented here; but resemble the types of analytic decisions made in previous power surveys (e.g. Orme & Combs-Orme, 1986).

Limitations of Part 3. Generalizability of the results for the p-curve analysis is particularly limited. The analysis was performed on tests from a single scholarly journal (i.e. RSWP) and may be best described as a pilot analysis of a novel statistical tool. Any findings, including those of evidential value described here, must be described as provisional.

While a robustness analysis was performed, several limitations may impact its utility; even if only for drawing inferences about RSWP. First, it was not always clear what a study’s

---

⁵⁹ See Correll, Mellinger, McClelland, & Judd (2020) for an overview; see Morey & Lakens (2016) for an alternative approach to power, especially pp. 17-18.

research questions and hypotheses were. Consequently, it was not always clear which tests comprised the primary analysis. This is further compounded when multiple outcomes (i.e. dependent variables) were present in a study. When this was the case, I took the outcome that was reported first, by consulting the methods and results sections. If the result was non-significant (a requirement for the p-curve analysis), the next outcome was chosen. Whether there was ambiguity or not, a second set of results was included (where possible) in order to perform a robustness test. Unfortunately, when test values and degrees of freedom were not explicitly stated, results could not be included in the p-curve analysis.⁶⁰ Overall, this analysis is limited by a high level of ambiguity in the materials analyzed, presence of my own researcher degrees of freedom, and narrowness in scope.

The p-curve method itself has its own limitations. Its creators (Simmons, Nelson, & Simmons, 2014) believe that it can provide a measure of evidential value. While I agree, critiques of the method merit response. For instance, Bishop and Thompson (2016) describe how “ghost variables” prevent us from *definitively* proving that a set of results has evidential value. Ghost variables are dependent variables that are omitted from the manuscript by the authors because no test of them resulted in statistical significance. Using simulations, they show that when dependent variables (i.e. reported and ghost types) are highly intercorrelated the shape of the p-curve can be skewed to the left. As a result, if the underlying data and its properties are unknown to the analyst, interpretation of a p-curve analysis can become muddled. It is quite possible that some studies analyzed in the current study have ghost variables. Though this can be rectified through use of Registered Reports or pre-registration of analyses, such practices are not

---

⁶⁰ This is an issue that was describe in Part 2 as well. Unfortunately, omitting key statistical information is not uncommon in published articles. See Weissgerber et al. (2018) for an investigation on how a subset of t-tests and ANOVAs were reported.

currently the norm in social work; even if they are encouraged (Harrison & Mayo-Wilson, 2014). Additionally, Erdfelder and Heck (2019) argue that inferences from left-skewed and right-skewed p-curves are ambiguous, with regard to determinations of p-hacking and evidential value, respectively. They argue that left-skewed curves can occur even with selective nonreporting of *significant* results and that right-skewed curves can occur even in when the set of studies contains no effect. After considering these arguments, I agree that the p-curve method is not without fault. However, the fact that it, and other methods (e.g. tests of excess significance, p-uniform, trim-and-fill, etc.), cannot *definitively* and/or directly offer proof of bias (Renkewitz & Keiner, 2019) does not mean it is without utility. Even indirect and flawed tools can have value towards scientific progress (e.g. Berk et al. 2013; Berk et al., 2014), if used with care.

Mitigating study limitations. There are three ways the rigor of this study can be improved. Prior to submission for publication a second rater will be used to ascertain the reliability of coding. This person (to be determined) will be offered authorship, in part, based on these contributions. Consequently, the chances of finding any major discrepancies in coding and/or systematic biases will be improved. Likewise, all materials will be uploaded to the study's public OSF page, in order to increase transparency. This will improve the likelihood that peer reviewers (and post-publication reviewers) may find inaccuracies in the data. The third strategy has already been performed. Robustness analyses were included for Parts 2 and 3 to minimize researcher degrees of freedom (Simmons et al., 2011) by providing a (simplified) type of multiverse analysis of the results (see Dragicevic et al., 2019; Steegan et al., 2016). These are discussed in more detail above.

Implications for Social Work Research

This study, despite its limitations, is an opportunity for critical self-reflection by the profession. We have made great progress over the last century and there is still more work to be done. Our statistical toolbox has become a core and common means for generating discipline-specific knowledge. We should continuously strive to improve its use, while attending to theory development and testing, research design, implementation, and other aspects of the research process.

To the extent that these findings are valid and generalizable, we should be cautiously optimistic that social work research can reliably detect most medium-large effects (when present) and that our results (may) actually have evidential value. With that said, we should be wary of the interpretability of those results, where statistical reporting is inadequate, especially when such information is used to inform policy and practice decisions. Journal editors in social work may consider more stringent adoption of reporting guidelines, such as the CONSORT (e.g. Turner, Shamseer, Altman, Schulz, & Moher, 2012) and STROBE statements (von Elm et al., 2014); among other variants (Shlonsky, D'Andrade, & Brookhart, 2002; Thyer, 2017) and/or use of the Registered Report format (Chambers et al., 2015). Doing so may help change norms in statistical practice from the top-down (without necessarily having to go to the extremes of Meehl, 1986). Social work researchers may do well to consider study-planning in more detail, using pre-registration and Registered Report formats, and review of relevant reporting guidelines. As an ethics-based profession, we owe it to ourselves, our participants/clients, and other researchers who build off of our work, to improve the interpretability and utility of our work by reporting our results in as much detail as possible (Altman, 1980; NASW, 2017; Osborne, 2008).

Future Directions: Social Work Meta-Research

Future meta-researchers would do well to conduct a more robust and ranging analysis of the social work literature. This could mean analyzing a longer time period, a greater number of journals, or findings across different levels of analysis (e.g. individual researchers or institutions). Having generated new information about statistical reporting and power, the current study could serve to inform more nuanced hypotheses and research questions. Social workers may also decide to explore other areas of meta-research, including whether or not results are accurately reported (compared with independent analysis of the data), how social workers are understanding statistics concepts and interpreting research results, among many other topics.

APPENDIX
FSU IRB EXEMPTION

FLORIDA STATE UNIVERSITY
OFFICE of the VICE PRESIDENT for RESEARCH

NOT HUMAN RESEARCH

November 1, 2019

Daniel Dunleavy
[REDACTED]

Dear Daniel Dunleavy:

On 11/1/2019, the IRB staff reviewed the following submission:

Title of Study:	Appraising contemporary social work research: Meta-research on reporting, statistical power, and evidential value
Investigator:	Daniel Dunleavy
Submission ID:	STUDY00000615
Study ID:	STUDY00000615
Funding:	None
IND, IDE, or HDE:	None
Documents Reviewed:	• Dunleavy_IRB_Protocol, Category: IRB Protocol;

The IRB staff determined that the proposed activity is not research involving human subjects as defined by DHHS and/or FDA regulations.

IRB review and approval by this organization is not required. This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these activities are research involving human subjects in which the organization is engaged, please submit a new request to the IRB for a determination. You can create a modification by clicking **Create Modification / CR** within the study.

Sincerely,

Human Subjects Research Office
humansubjects@fsu.edu

REFERENCES

- Abbott, A. (2007). Notes on replication. *Sociological Methods & Research*, 36(2), 210-219.
- Aguinis, H., & Solarino, A. M. (2019). Transparency and replicability in qualitative research: The case of interviews with elite informants. *Strategic Management Journal*.
<https://doi.org/10.1002/smj.3015>
- Altman, D. G. (1980). Misuse of statistics is unethical. *British Medical Journal*, 281, 1182-1184.
- Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine*, 1(1), 59-71.
- Altman, D. G. (1994). The scandal of poor medical research. *British Medical Journal*, 308, 283-284.
- Amrhein, V., Greenland, S., & McShane, B. (2019). Retire statistical significance. *Nature*, 567, doi: 10.1038/d41586-019-00857-9
- Aschwanden, C. (2015, November 24). Not even scientists can easily explain p-values. *FiveThirtyEight*. Retrieved from: <https://fivethirtyeight.com/features/not-even-scientists-can-easily-explain-p-values/>
- Aubele, J. & Perruso, C. (2017). Toward a sustainable method for core journal lists: A test case using journals in social work. *The Serials Librarian*, 73(2), 89-106.
- Babbage, C. (1830) *Reflections on the decline of science in England and on some of its causes*. London, England: B. Fellowes. Retrieved from:
<https://archive.org/details/reflectionsonde00babbgoog/page/n5>
- Baker, M. (2016). Is there a reproducibility crisis? *Nature*, 533, 452-454.
- Barnes, J. C., TenEyck, M. F., Pratt, T. C., & Cullen, F. T. (2019). How powerful is the evidence in criminology? On whether we should fear a coming crisis of confidence? *Justice Quarterly*, 1-27. Retrieved from: <https://doi.org/10.1080/07418825.2018.1495252>
- Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100, 407-425.
- Bem, D., Tressoldi, P. E., Rabeyron, T., & Duggan, M. (2016). Feeling the future: A Meta-analysis of 90 experiments on the anomalous anticipation of random future events [version 2; peer review: 2 approved]. *F1000Research*, 4, 1188.
<https://doi.org/10.12688/f1000research.7177.2>
- Benjamin, D. J., & Berger, J. O. (2019). Three recommendations for improving the use of p-values. *The American Statistician*, 73, 186-191.

- Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R. A., ... Johnson, V. E. (2018). Redefine statistical significance, *Nature Human Behavior*, 2(1), 6-10.
- Berger, J. O. (2003). Could Fisher, Jeffreys, and Neyman have agreed on testing? *Statistical Science*, 18(1), 1-32.
- Berger, J. O., & Wolpert, R. L. (1984). *The likelihood principle* (2nd Ed.). Institute of Mathematical Statistics.
- Berk R. A., Brown, L., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2013) What you can learn from wrong causal models. In S. Morgan (Ed.) *Handbook of causal analysis for social research* (pp. 403-424). Springer: Dordrecht.
- Berk, R. A., & Freedman, D. A. (2003). Statistical assumptions as empirical commitments. In T. G. Blomberg and S. Cohen (Eds.), *Law, punishment, and social control: Essays in honor of Sheldon Messinger* (pp. 235-254). New York, NY: Aoldine de Gruyter.
- Berk, R., Brown, L., Buja, A., George, E., Pitkin, E., Zhang, K., & Zhao, L. (2014). Misspecified mean function regression: Making good use of regression models that are wrong. *Sociological Methods & Research*, 43(3), 422–451.
- Bird, A. (2018). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*. Retrieved from: <https://doi.org/10.1093/bjps/axy051>
- Bishop, D. V., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*. <https://peerj.com/articles/1715/>
- Blier, P., Ward, H. E., Tremblay, P., Laberge, L., Hébert, C., & Bergeron, R. (2010). Combination of antidepressant medications from treatment initiation for major depressive disorder: A double-blind randomized study. *American Journal of Psychiatry*, 167(3), 281–288. doi:10.1176/ appi.ajp.2009.09020186
- Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No support for historical candidate gene or candidate gene-by-interaction hypotheses for major across multiple large samples. *American Journal of Psychiatry*, 176(5), 376-387.
- Borsboom, D. (2013, November 20). Theoretical amnesia. *Center for Open Science*. Retrieved from: <http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/>
- Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal*, 9(3), 391-401.
- Bruner, J. P. (2013). Policing epistemic communities. *Episteme*, 10(4), 403-416.

- Bushway, S. D., Sweeten, G., & Wilson, D. B. (2006). Size matters: Standard errors in the application of null hypothesis significance testing in criminology and criminal justice. *Journal of Experimental Criminology*, 2(1), 1-22.
- 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, 365-376.
- Cai, T., Parast, L., Ryan, L. (2010). Meta-analysis for rare events. *Statistics in Medicine*, 29(20), 2078-2089.
- Camerer, C. F., Breber, a., Holzmeister, F., Ho, T-H., Huber, J., Johnnesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behavior*, 2, 637-644.
- Campbell, D. T. (1987). Evolutionary epistemology. In G. Radnitzky & W. W. Bartley (Eds.), *Evolutionary epistemology, rationality, and the sociology of knowledge* (pp. 47-89). LaSalle, IL: Open Court.
- Carbine, K. A., & Larson, M. J. (2019). Quantifying the presence of evidential value and selective reporting in food-related inhibitory control training: A p-curve analysis. *Health Psychology Review*, 13(3), 318-343.
- Caspi, A., Sugden, K., Moffitt, T. E., Taylor, A., Craig, I. W., Harrington, H., ... Poulton, R. (2003). Influence of life stress on depression: Moderation by a polymorphism in the 5-HTT gene. *Science*, 18(301), 386-389.
- Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R., & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*. <https://doi.org/10.1177/2515245919858072>
- Center for Drug Evaluation and Research. (2001). Statistical approaches to establishing bioequivalence. *U.S. Food and Drug Administration*. Retrieved from: <https://www.fda.gov/media/70958/download>
- Chambers, C. (2013). Registered Reports: A new publishing initiative at Cortex. *Cortex*, 49, 609-610.
- Chambers, C. (2017). *The 7 deadly sins of psychology: A manifesto for reforming the culture of scientific practice*. Princeton, NJ: Princeton University Press.
- Chambers, C., Dienes, Z., McIntosh, R. D., Roshtein, P., & Willmes, K. (2015). Registered Reports: Realigning incentives in scientific publishing. *Cortex*, 66, A1-A2. Retrieved from: doi: 10.1016/j.cortex.2015.03.022.

- Chan, A., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *JAMA*, 291(20), 2457-2465.
- Chartier, C. R., Kline, M., McCarthy, R. J., Nuijten, M. B., Dunleavy, D. J., & Ledgerwood, A. (2018, December). The cooperative revolution is making psychological science better. *APS Observer*, 31(10). Retrieved from: <https://www.psychologicalscience.org/observer/the-cooperative-revolution-is-makingpsychological-science-better>
- Cohen, J. (1962). The statistical power of abnormal-social psychological research. *Journal of Abnormal and Social Psychology*, 65(3), 145-153.
- Cohen, J. (1973). Statistical power analysis and research results. *American Educational Research Journal*, 10(3), 225-229.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304-1312.
- Cohen, J. (1992). A power primer. *Psychological Bulletin*, 112, 1, 155-159.
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychologist*, 49(12), 997-1003.
- Colling, L. J., & Szűcs, D. (2018). Statistical inference and the replication crisis. *Review of Philosophy and Psychology*. <https://doi.org/10.1007/s13164-018-0421-4>
- Correll, J., Mellinger, C., McClelland, G. H., & Judd, C. M. (2020). Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Science*, 24(3), 200-207.
- Cowger, C. D. (1984). Statistical significance tests: Scientific ritualism or scientific method? *Social Service Review*, 58(3), 358-372.
- Crane, J. A. (1976). The power of social intervention experiments to discriminate differences between experimental and control groups. *Social Service Review*, 50(2), 224-242.
- Cuddy, A. J. C., Schultz, J., & Fosse, N. E. (2018). P-curving a more comprehensive body of research on postural feedback reveals clear evidential value for power-posing effects: Reply to Simmons and Simonsohn (2017). *Psychological Science*, 29(4), 656-666.
- Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist*, 60(2), 170-180.
- de Groot, A. D. (1961). *Methodologie (Methodology)*. Den Haag: Mouton.

- de Groot, A. D. (1969). *Methodology*. The Hague, The Netherlands: Mouton.
- de Groot, A. D. (1969/2014). The meaning of “significance” for different types of research. *Acta Psychol*, 148, 188–194. [translated by E-J. Wagenmakers, D. Borsboom, J. Verhagen, R. Kievit, M. Bakker, A. Cramer, D. Matzke, D. Mellenbergh, & H. L. J. van der Maas]
Retrieved from: <http://dx.doi.org/10.1016/j.actpsy.2014.02.001>
- de Haas, B. (2018). How to enhance the power to detect brain-behavior correlations with limited resources. *Frontiers in Human Neuroscience*, 12, 421. Doi: <https://dx.doi.org/10.3389%2Ffnhum.2018.00421>
- de Smidt, G. A., & Gorey, K. M. (1997). Unpublished social work research: Systematic replication of a recent meta-analysis of published intervention effectiveness research. *Social Work Research*, 21(1), 58-62.
- de Vries, Y. A., Roest, A. M., Turner, E. H., & de Jonge, P. Hiding negative trails by pooling them: A secondary analysis of pooled-trials publication bias in FDA-registered antidepressant trials. *Psychological Medicine*, 1-7, DOI: 10.1017/S0033291718002805.
- Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference*. London, UK: Palgrave Macmillan.
- Dienes, Z. (2011). Bayesian versus orthodox statistics: Which side are you on? *Perspectives on Psychological Science*, 6(3), 274-290.
- Dienes, Z. (2016). How Bayes factors chance scientific practice. *Journal of Mathematical Psychology*, 72, 78-89.
- Dienes, Z., & Mclatchie, N. (2018). Four reasons to prefer Bayesian analyses over significance testing. *Psychonomic Bulletin & Review*, 25(1), 207-218.
- Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). *Increasing the transparency of research papers with explorable multiverse analyses*. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.
<https://doi.org/10.1145/3290605.3300295>
- Dumas-Mallet, E., Button, K. S., Boraud, T., Gonon, F., & Munafò, M. R. (2017). Low statistical power in biomedical science: A Review of three human research domains. *Royal Society Open Science*, 4, 160254.
- Dunleavy, D. J. (2020). Coronavirus as impetus for a lasting change in research culture. SocArXiv. doi:10.31235/osf.io/2ryt3
- Dunleavy, D. J., & Lacasse, J. R. (2020). Reproducible social work research [Course Syllabus]. <https://doi.org/10.17605/OSF.IO/PX62B>

- Dunleavy, D. J., Tennant, J., & Francuzik, W. (2020, June 2). Scientific practice in the time of COVID-19: It's time to commit to change. *Zenodo*.
<https://doi.org/10.5281/zenodo.3873631>
- Edwards, M. A., & Roy, S. (2017). Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. *Environmental Engineering Science*, 34(1), 51-61.
- von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & the STROBE Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: Guidelines for reporting observational studies. *Journal of Clinical Epidemiology*, 61(4), 344-349.
- Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, UK: Cambridge University Press.
- Epstein, W. M. (2004). Confirmational response bias and the quality of editorial processes among American social work journals. *Research on Social Work Practice*, 14(6), 450-458.
- Erdfelder, E., & Heck, D. W. (2019). Detecting evidential value and p-hacking with the p-curve tool: A word of caution. *Zeitschrift für Psychologie*, 227, 249-260.
<https://doi.org/10.1027/2151-2604/a000383>
- Etz, A., & Vandekerckhove, J. (2018). Introduction to Bayesian inference for psychology. *Psychonomic Bulletin & Review*, 25(1), 5-34.
- Fanelli, D. (2012) Negative results are disappearing from most disciplines and countries. *Scientometrics*, 90, 891–904.
- Faul, F., Erdfelder, E., Lang, A., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175-191.
- Feynman, R. P. (1974). Cargo cult science. *Engineering & Science*, 37, 10–13.
- Fidler, F., Thomason, N., Cumming, G., Finch, S., & Leeman, J. (2004). Editors can lead researchers to confidence intervals, but can't make them think: Statistical reform lessons from medicine. *Psychological Science*, 15(2), 119-126.
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1), 45-52.
- Field, S. A., Tyre, A. J., Jonzén, N., Rhodes, J. R., & Possingham, H. P. (2004). Minimizing the cost of environmental management decisions by optimizing statistical thresholds. *Ecology Letters*, 7(8), 669-675.

- Fischer, J. (1973). Is casework effective? A review. *Social Work*, 18, 5-20.
- Fischer, J. (1976). *The effectiveness of social casework*. Springfield, IL: Charles Thomas.
- Fischer, J. (1981). The social work revolution. *Social Work*, 26(3), 199-207.
- Fisher R. A. (1956). *Statistical methods and scientific inferences*. Hafner: New York.
- Fisher, R. A. (1930). *Statistical methods for research workers* (3rd Ed.). Oxford, ENG: Oliver & Boyd.
- Fiske, S. T. (2016a). Mob rule or wisdom of crowds? *APS Observer* [Draft submission]. Retrieved from: <https://web.archive.org/web/20180407065201/http://datacolada.org/wp-content/uploads/2016/09/Fiske-presidential-guest-column-APS-Observer-copy-edited.pdf>
- Fiske, S. T. (2016b). A call to change science's culture of shaming. *APS Observer*, 29(9). Retrieved from: <https://www.psychologicalscience.org/observer/a-call-to-change-sciences-culture-of-shaming>
- Flexner, A. (1915/2001). Is social work a profession? *Research on Social Work Practice*, 11, 152-165.
- Food and Drug Administration. (2010). Guidance for the use of Bayesian statistics in medical device clinical trials. *U.S. Department of Health and Human Services*. Retrieved from: <https://www.fda.gov/regulatory-information/search-fda-guidance-documents/guidance-use-bayesian-statistics-medical-device-clinical-trials>
- Franco, A., Malhotra, N., & Simonovits, G. (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science*, 345 (6203), 1502-1505.
- Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PLoS ONE*, 13(7), e0200303.
- Freedman, D. (1995). Some issues in the foundation of statistics. *Foundations of Science*, 1, 19-39.
- Freedman, D., Pisani, R., & Purves, R. (1979). *Statistics*. New York, NY: W.W. Norton & Company, Inc.
- Freese, J. (2007). Replication standards for quantitative social science: Why not sociology? *Sociological Methods & Research*, 36(2), 153-172.
- Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, 31(3). Retrieved from: <https://www.psychologicalscience.org/observer/measurement-matters>

- Gambrill, E. (1997). *Thinking about knowledge and how to get it*. Paper presented at the Ninth Symposium on Doctoral Research in Social Work [April 18, 1997]. College of Social Work, The Ohio State University, Columbus, OH, USA.
- Gambrill, E. (1999). Evidence-based practice: An antidote to authority-based practice. *Families in Society: The Journal of Contemporary Social Services*, 80(4), 341-350.
- Gambrill, E. (2018). The promotion of avoidable ignorance in the British Journal of Social Work. *Research on Social Work Practice*, 29(4), 455-469.
- Gardner, M. J., & Altman, D. G. (1986). Confidence intervals rather than p values: Estimation rather than hypothesis testing. *The British Medical Journal*, 292, 746-750.
- Garland Campbell, K. (2018). Errors in statistical decision making. In B. Glaz & K. M. Yeater (Eds.), *Applied statistics in agricultural, biological, and environmental sciences*. DOI: 10.2134/appliedstatistics.2016.0007
- Gelman, A. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science*, 9(6), 641-651.
- Gelman, A. (2017, December 13). Yes, you can do statistical inference from nonrandom samples. Which is a good thing, considering that nonrandom samples are pretty much all we've got. *Statistical Modeling, Causal Inference, and Social Science* [Blog]. Retrieved from: www.statmodeling.stat.columbia.edu/2017/12/13/yes-can-statistical-inference-nonrandom-samples-good-thing-considering-nonrandom-samples-pretty-much-weve-got/
- Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem even when there is no "fishing expectation" or "p-hacking" and the research hypothesis was posited ahead of time*. Unpublished manuscript. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf
- Gigerenzer, G. (1993). The superego, the ego, and the id in statistical reasoning. In G. Karen & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues*. Hillsdale, NJ: Hove & London.
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33, 587-606.
- Gigerenzer, G. (2015). Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41(2), 421-440.
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198-218.
- Giner-Sorolla, R., Schoemann, A. M., Montoya, A. K., Conrique, B. G., Carpenter, T. P., Reifman, A., ... Bostyn, D. H. (2019). Power to detect what? Considerations for planning and evaluating sample size. <https://osf.io/jnmya/>

- Glass, D. J. (2010). A critique of the hypothesis, and a defense of the question, as a framework for experimentation. *Clinical Chemistry*, 56(7), 1080-1085.
- Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., ... Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *Lancet*, 383, 267-276.
- Glisson, C., & Fischer, J. (1987). Statistical training for social workers. *Journal of Social Work Education*, 3, 50-58.
- Glover, S., & Dixon, P. (2014). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11(5), 791-806.
- Glover, S., & Dixon, P. (2014). Likelihood ratios: A simple and flexible statistic for empirical psychologists. *Psychonomic Bulletin & Review*, 11(5), 791-806.
- Goldenberg, M. J. (2019). Vaccines, values and science. *CMAJ*, 191, e397-e398. doi: 10.1503/cmaj.181635
- Gomory, T. (1997). Social work and philosophy. In M. Reisch, & E. Gambrill (Eds.), *Social work in the 21st century* (pp. 300-310). Thousand Oaks, CA: Pine Forge Press.
- Gomory, T. (2001a). A fallibilistic response to Thyer's theory of theory-free empirical research in social work practice. *Journal of Social Work Education*, 37(1), 26-50.
- Gomory, T. (2001b). Critical rationalism (Gomory's blurry theory) or positivism (Thyer's theoretical myopia): Which is the prescription for social work research. *Journal of Social Work Education*, 37(1), 67-78.
- Goodman, (2008). A dirty dozen: Twelve p-value misconceptions. *Seminars in Hematology*, 45, 135-140.
- Gorard, S., White, P. (2017). Still against inferential statistics: Rejoinder to Nicholson and Ridgway. *Statistics Education Research Journal*, 16(1), 74-79.
- Gøtzsche, P. C. (1990). Bias in double-blind trials. *Danish Medical Bulletin*, 37, 329-336.
- Greenland, S. (2019). Valid p-values behave exactly as they should: Some misleading criticisms of p-values and their resolution with s-values. *The American Statistician*, 73(supplement 1), 106-114. <https://doi.org/10.1080/00031305.2018.1529625>
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N. & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350.

- Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, 82(1), 1–20.
- Hacking, I. (1976). *Logic of statistical inference*. London, UK: Cambridge University Press.
- Hacking, I. (2001). *Probability and inductive logic*. New York, NY: Cambridge University Press.
- Hájek, A. (2008). A philosopher's guide to probability. In G. Bammer & M. Smithson (Eds.), *Uncertainty and risk: Multidisciplinary perspectives* (pp. 91–104). London, UK: Earthscan.
- Hall, B. F., Wagge, J. R., Pfuhl, G., Stieger, S., Vergauwe, E., IJzerman, H., ... Stephen, I. (2018, October 15). Accelerated CREP - RRR: Turri, Buckwalter, & Blouw (2015). PsyArXiv. <https://doi.org/10.31234/osf.io/zeux9>
- Hand, D. J. (1994). Deconstructing statistical questions. *Journal of the Royal Statistical Society*, 157(3), 317-356.
- Hardwicke, T. E., Serghiou, S., Janiaud, P., Danchev, V., Crüwell, S., Goodman, S., & Ioannidis, J. P. A. (2019, June 13). Calibrating the scientific ecosystem through meta-research. MetaArXiv. <https://doi.org/10.1146/annurev-statistics-031219-041104>
- Harrison, B. A., & Mayo-Wilson, E. (2014). Trial registration: Understanding and preventing reporting bias in social work research. *Research on Social Work Practice*, 24(3), 372-376.
- 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.
- Heathers, J. (2017, September 28). Why we find and expose bad science. *Medium*. Retrieved from: <https://medium.com/@jamesheathers/why-we-find-and-expose-bad-science-e47387a0e333>
- Heesen, R. (2018). Why the reward structure of science makes reproducibility problems inevitable. *The Journal of Philosophy*, 12, 661-674.
- Hendricks, V. F., & Vestergaard, M. (2018). *Reality lost: Markets of attention, misinformation, and manipulation*. Cham, CH: Springer Open.
- Henkel, R. E. (1976). *Tests of significance*. Newbury Park, CA: Sage Publications.
- Herrera-Perez, D., Haslam, A., Crain, T., Gill, J., Livingston, C., Kaestner, V., ... Prasad, V. (2019). Meta-research: A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *eLife*, 8, e45183. DOI: 10.7554/eLife.45183

- Hodge, D. R., & Lacasse, J. R. (2011). Ranking disciplinary journals with the google scholar H-Index: A new tool for constructing cases for tenure, promotion, and other professional decisions. *Journal of Social Work Education*, 47(3), 579-596.
- Hodge, D. R., Yu, M., & Kim, A. (2019). Assessing the quality and prestige of disciplinary social work journals: A national study of faculty perceptions. *Research on Social Work Practice*, 1-9. <https://doi.org/10.1177%2F1049731519890402>
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J., (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bull Rev*, DOI: 10.3758/s13423-013-0572-3
- Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1), 1-6.
- Howard, M. H., & Garland, E. L. (2015). Social work research: 2044. *Journal of the Society for Social Work and Research*, 6(2), 173-200.
- Howson, C., & Urbach P. (1993). *Scientific reasoning: The Bayesian approach*. Chicago and La Salle, IL: Open Court Publishing.
- Hung, H. M., O'Neill, R. T., Bauer, P., & Köhne, K. (1997). The behavior of the p-value when the alternative hypothesis is true. *Biometrics*, 53(1), 11-22.
- Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124.
- Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspective on Psychological Science*, 7(6), 645-654.
- Ioannidis, J. P. A. (2013). Implausible results in human nutrition research. *BMJ*, 347, f6698.
- Ioannidis, J. P. A. (2014). Publication and other reporting biases in cognitive sciences: Detection, prevalence and prevention. *Trends in Cognitive Science*, 18(5), 235-241.
- Ioannidis, J. P. A. (2016). Evidence based medicine has been hijacked: A report to David Sackett. *Journal of Clinical Epidemiology*, 73, 82-86.
- Ioannidis, J. P. A. (2018). The challenge of reforming nutrition epidemiologic research. *JAMA*, 320(10), 969-970.
- Ioannidis, J. P. A. (2019). Calibrating the scientific ecosystem through meta-research. *Annual Review of Statistics and Its Application*. Retrieved from: <https://osf.io/krb58/>.
- Ioannidis, J. P. A. (2019). Options for publishing research without any p-values. *European Heart Journal*, 40(31), 2555-2556.

- Ioannidis, J. P. A., Fanelli, D., Dunne, D. D., & Goodman, S. N. (2015). Meta-research: Evaluation and improvement of research methods and practices. *PLoS Biology*, 13(10), e1002264.
- Jureidini, J. N., McHenry, L. B., & Mansfield, P. R. (2008). Clinical trials and drug promotion: Selective reporting of study 329. *International Journal of Risk and Safety Medicine*, 20, 73-81.
- Kahneman, D. (2012, September 26). A proposal to deal with questions about priming effects. Retrieved from: https://www.nature.com/news/polopoly_fs/7.6716.1349271308!/suppinfoFile/Kahneman%20Letter.pdf
- Kanyongo, G. Y., Brook, G. P., Kyei-Blankson, L., & Gocmen, G. (2007). Reliability and statistical power: How measurement fallibility affects power and required sample sizes for several parametric and nonparametric statistics. *Journal of Modern Applied Statistical Methods*, 6(1), 81-90.
- Kass, R. E. (2011). Statistical inference: The big picture. *Statistical Science*, 26(1), 1-9.
- Keller, M. B., Ryan, N. D., Strober, M., Klein, R. G., Kutcher, S. P., Birmaher, B., ... McCafferty, J. P. (2001). Efficacy of paroxetine in the treatment of adolescent major depression: A randomized, controlled trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(7), 762-772.
- Kelly, C. D. (2006). Replicating empirical research in behavioral ecology. *The Quarterly Review of Biology*, 81(3), 221-236.
- Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, 2(3), 196-217.
- Kirk, S. A., & Fischer, J. (1976). Do social workers understand research? *Journal of Social Work Education*, 12(1), 63-70.
- Kirk, S. A., & Reid, W. J. (2002). *Science and social work: A critical appraisal*. New York, NY: Columbia University Press.
- Kraemer, H. C., & Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage Publications.
- Kruschke, J. K. (2011). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. London, UK: Academic Press.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago, IL: The University of Chicago Press.

- Kuhn, T. S. (1970). Logic of discovery or psychology of research? In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 1-23). London, UK: Cambridge University Press.
- Kuhn, T. S. (1977). *The essential tension: Selected Studies in scientific tradition and change*. Chicago, IL: University of Chicago Press.
- La Caze, A. (2016). Frequentism. In A. Hájek & C. Hitchcock (Eds.), *The Oxford Handbook of Probability and Philosophy*. DOI: 10.1093/oxfordhb/9780199607617.013.17
- Lacasse, J. R. (2008). *Patient-level demographic and clinical variables associated with sustained antipsychotic polypharmacy: A naturalistic study of state hospital inpatients*. Retrieved from: http://purl.flvc.org/fsu/fd/FSU_migr_etd-7178
- Lacasse, J. R. (2011). Reanalyzing a randomized controlled trial of combination antidepressant treatment with mirtazapine: Confidence intervals suggest substantial uncertainty. *Ethical Human Psychology and Psychiatry*, 13(2), 149-154. doi:10.1891/1559-4343.13.2.149
- Lacasse, J. R., & Leo, J. (2010). Ghostwriting at elite academic medical centers in the United States. *PLoS Medicine*, 7(2), e1000230. <https://doi.org/10.1371/journal.pmed.1000230>
- Lakatos, I. (1968). Criticism and the methodology of scientific research programmes. *Proceedings of the Aristotelian Society*, 69, 149-186.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4(863). <https://doi.org/10.3389/fpsyg.2013.00863>
- Lakens, D. (2017). Professors are not elderly: Evaluating the evidential value of two social priming effects through p-curve analyses. PsyArXiv. <https://psyarxiv.com/3m5y9/>
- Lakens, D. (2017, May 11). How a power analysis implicitly reveals the smallest effect size you care about. *The 20% Statistician* [Blog]. Retrieved from: <http://daniellakens.blogspot.com/2017/05/how-power-analysis-implicitly-reveals.html>
- Lakens, D., Adolphi, F. G., Albers, C. J., Anvari, F., Apps, M. A. J., Argamon, S. E., ... Zwaan, R. A. (2018). Justify your alpha: A response to “Redefine statistical significance”. *Nature Human Behavior*, 2, 168-171.
- Lazic, S. E. (2018). Four simple ways to increase power without increasing sample size. *Laboratory Animals*, 52(6), 621-629. doi: 10.1177/0023677218767478
- LeCroy, C. W. (1985). Methodological issues in the evaluation of social work practice. *Social Service Review*, 59(3), 345-357.

- LeNoury, J., Nardo, J., Healy, D., Jureidini, J., Raven, M., Tufanaru, C., & Abi-Jaoude, E. (2015). Restoring study 329: Efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ*, 351, doi: <http://dx.doi.org/10.1136/bmj.h4320>
- Lindsay, D. S. (2018). Sharing data and materials in Psychological Science. *Psychological Science*, 28(6), 699-702.
- Lipsey, M. W. (1990). *Design sensitivity: Statistical power for experimental research*. Newbury Park, CA: Sage Publications.
- Ly, A. (2017). *Bayes factors for research workers*. University of Amsterdam. Retrieved from: <http://www.alexander-ly.com/wp-content/uploads/2018/01/Ly-A-2017.-Bayes-factors-for-research-workers.pdf>
- Machery, E. (2012). Power and negative results. *Philosophy of Science*, 79(5), 808-820.
- Machery, E. (2019). The alpha war. *Review of Philosophy and Psychology*, 1-25. <https://doi.org/10.1007/s13164-019-00440-1>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147-163.
- Mayo, D. G. (2003). Could Fisher, Jeffreys and Neyman have agreed? Commentary on J. Berger's Fisher address *Statistical Science*, 18, 19-24.
- Mayo, D. G. (2018). *Statistical inference as severe testing*. New York, NY: Cambridge University Press.
- McCloskey, D. N. (1985). The loss function has been mislaid: The rhetoric of significance tests. *The American Economic Review*, 75(2), 201-205.
- McCloskey, D. N. (1995). The analytical economist: The insignificance of statistical significance. *Scientific American*, 272(4), 32-33.
- McCloskey, D. N., & Ziliak, S. T. (1996). The standard error of regressions. *Journal of Economic Literature*, 34(1), 97-114.
- McElreath, R., & Smaldino, P. E. (2015) Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE* 10(8): e0136088. doi:10.1371/journal.pone.0136088
- McNeeley, S., & Warner, J. J. (2015). Replication in criminology: A necessary practice. *European Journal of Criminology*, 12(5), 581-597. <http://dx.doi.org/10.1177/1477370815578197>

- McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician*, 73(S1), 235-245.
- Meehl, P. E. (1976). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, 34(2), 103-115.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46, 806-834.
- Meehl, P. E. (1986). What social scientists don't understand. In D. W. Fiske & R. A. Shweder (Eds.). *Metatheory in social science: Pluralism and subjectivities* (pp. 315-338). Chicago, IL: University of Chicago Press.
- Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66, 195-244.
- Merton, R. K. (1973). *The sociology of science: Theoretical and empirical investigations*. Chicago, IL: The University of Chicago Press.
- Mischel, W. (2009). The toothbrush problem. *APS Observer*, 21. Retrieved from: <https://www.psychologicalscience.org/observer/the-toothbrush-problem>
- Morey, R. D. (2017). When the statistical tail wags the scientific dog. Should we 'redefine' statistical significance? *Medium* [Blog] Retrieved from: <https://medium.com/@richarddmores/when-the-statistical-tailwags-the-scientific-dog-d09a9f1a7c63>
- Morey, R. D. (2018). Redefining statistical significance: The statistical arguments. *Medium* [Blog] Retrieved from: <https://medium.com/@richarddmores/redefining-statistical-significance-the-statistical-arguments-ae9007bc1f91>
- Morey, R. D., & Lakens, D. (2016). Why most of psychology is statistically unfalsifiable. Retrieved from: https://raw.githubusercontent.com/richarddmores/psychology_resolution/master/paper/response.pdf
- Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, 23(1), 103-123.
- Morey, R. D., Romeijn, J-W., & Rouder, J. N. (2016). The philosophy of Bayes factors and the quantification of statistical evidence. *Journal of Mathematical Psychology*, 72, 6-18.
- Morrison, D. E., & Henkel, R. E. (1970). *The significance test controversy*. Oxford, England: Aldine.

- Moshontz, H., Campbell, L., Ebersole, C. R., IJzerman, H., Urry, H. L., Forscher, P. S., ... Chartier, C. R. (2018). The Psychological Science Accelerator: Advancing psychology through a distributed collaborative network. *Advances in Methods and Practices in Psychological Science*, 1(4), 501-515.
- Motulsky, H. (2016). *Essential biostatistics: A nonmathematical approach*. New York, NY: Oxford University Press.
- Motyl, M., Demos, A. P., Carsel, T. S., Hanson, B. E., Melton, Z. J., Mueller, A. B., ... Skitka, L. J. (2017). The state of social and personality science: Rotten to the core, not so bad, getting better, or getting worse? *Journal of Personality and Social Psychology*, 113(1), 34-58.
- Mudge, J. F., Baker, L. F., Edge, C. B., & Houlahan, J. E. (2012). Setting an optimal α that minimizes errors in null hypothesis significance tests. *PLoS One*, 7(2), e32734. Retrieved from: <https://doi.org/10.1371/journal.pone.0032734>
- Mullen, E. J., & Dumpson, J. R. (1972). *Evaluation of social intervention*. San Francisco, CA: Jossey-Bass Publishers.
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., ... Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behavior*, 1(1), Retrieved from: <http://dx.doi.org/10.1038/s41562-016-0021>
- Murphy, K. R., & Aguinis, H. (2019). HARKing: How badly can cherry-picking and question trolling produce bias in published results? *Journal of Business and Psychology*, 34(1), 1-17.
- Murphy, K. R., & Myers, B. (2004). *Statistical power analysis A simple and general model for traditional and modern hypothesis tests*. Mahwah, NJ: Taylor & Francis.
- Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behavior*, 3(3), 221-229.
- Naimi, A. I., & Whitcomb, B. W. (2020). Can confidence intervals be interpreted? *American Journal of Epidemiology*, 1-8. Retrieved from: <https://doi.org/10.1093/aje/kwaa004>
- National Academies of Sciences, Engineering, and Medicine. (2016). *Statistical challenges in assessing and fostering the reproducibility of scientific results: Summary of a workshop*. Washington, DC: The National Academies Press. Retrieved from: <https://doi.org/10.17226/21915>
- National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/25303>

- National Association of Social Workers (NASW). (2017). NASW Code of Ethics (Rev ed.). Retrieved from: <https://www.socialworkers.org/About/Ethics/Code-of-Ethics/Code-of-Ethics-English>
- Neyman, J. (1937). Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 236, 333-380.
- Neyman, J. (1957). "Inductive behavior" as a basic concept of philosophy of science. *Revue de l'Institut International de Statistique / Review of the International Statistical Institute*, 25, 7-22.
- Neyman, J., & Pearson, E. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231, 289-337.
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241-301.
- Nosek, B. A., & Errington, T. (2019). What is replication. *MetaArXiv*. Retrieved from: <https://osf.io/preprints/metaarxiv/u4g6t>
- Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *PNAS*, 115(11), 2600-2606.
- Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7(6), 615-631.
- Nugent, W. R. (1987). Use and evaluation of theories. *Social Work Research & Abstracts*, 23(1), 14-19.
- Nuijten, M. (2018). *Research on research: A meta-scientific study of problems and solutions in psychological science* s.l.: Gildeprint.
- O'Connor, C. & Weatherall, J. O. (2020). Why false claims about COVID-19 refuse to die: Tracking the information zombie apocalypse. *Nautalis* (Online Magazine). Retrieved from: nautil.us/issue/84/outbreak/why-false-claims-about-covid-19-refuse-to-die
- Oakes, M. (1986). *Statistical inference: A commentary for the social and behavioural sciences*. New York: John Wiley.
- Okpych, N. J., & Y, J. L-H. (2014). A historical analysis of evidence-based practice in social work: The unfinished journey toward an empirically grounded profession. *Social Service Review*, 88(1), 3-58.

- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943-951.
- Orme, J. G. (1987). Statistical power as a methodological issue in the evaluation of social work practice. *Social Service Review*, 61(1), 165-167.
- Orme, J. G., & Combs-Orme, T. D. (1986). Statistical power and Type II error in social work research. *Social Work Research & Abstracts*, 22(3), 3-10.
- Orme, J. G., & Tolman, R. M. (1986). The statistical power of a decade of social work education research. *Social Service Review*, 60(4), 619-632.
- Osborne, J. W. (2008). Sweating the small stuff in educational psychology: How effect size and power reporting failed to change from 1969-1999. *Journal of Experimental Educational Psychology*, 28(2), 151-160.
- Parsons, S. (2018, February 5). Ignoring measurement reliability is a real-life horror story. Medium [Blog]. Retrieved from: www.medium.com/@Sam_D_Parsons/ignoring-measurement-reliability-is-a-real-life-horror-story-b98a2517db26
- Pederson, D. B., & Hendricks, V. F. (2014). Science bubbles. *Philosophy and Technology*, 27, 503-518.
- Penick, J. E., & Brewer, J. K. (1972). The power of statistical tests in science teaching research. *Journal of Research in Science Teaching*, 9(4), 377-381.
- Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology*. DOI: [10.3389/fpsyg.2015.00223](https://doi.org/10.3389/fpsyg.2015.00223)
- Poole, C. (1987). Beyond the confidence interval. *American Journal of Public Health*, 77(2), 195-199.
- Popper, K. R. (1959/1968). *The logic of scientific discovery*. New York, NY: Harper & Row Publishers.
- Popper, K. R. (1963). *Conjectures and refutations: The growth of scientific knowledge*. London, UK: Routledge & Kegan Paul plc.
- Popper, K. R. (1970). Normal science and its dangers. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 51-58). London, UK: Cambridge University Press.
- Popper, K. R. (1979). *Objective knowledge*. New York: Oxford University Press.
- Pridemore, W. A., Makel, M. C., & Plucker, J. A. (2018). Replication in criminology and the social sciences. *Annual Review of Criminology*, 1, 19-38.

- Quintana, D. S., & Eriksen, D. R.. (2017, April 8). Bayesian alternatives for common null-hypothesis significance tests in psychiatry: A non-technical guide using JASP. <https://doi.org/10.1186/s12888-018-1761-4>
- Rasnick, B. (2017, July 31). What a nerdy debate about p-values shows about science — and how to fix it. *Vox*. Retrieved from: <https://www.vox.com/science-and-health/2017/7/31/16021654/p-values-statistical-significance-redefine-0005>
- Reichenbach, H. (1938). *Experience and prediction: An analysis of the foundations and the structure of knowledge*. Chicago, IL: The University of Chicago Press.
- Reichenbach, H. (1951). *The rise of scientific philosophy*. Berkeley, CA, US: University of California Press.
- Renkewitz, F., & Keiner, M. (2019). How to detect publication bias in psychological research: A comparative evaluation of six statistical methods. *Zeitschrift für Psychologie*, 227(4), 261-279. <http://dx.doi.org/10.1027/2151-2604/a000386>
- Richmond, M. (1917). *Social diagnosis*. New York: Russell Sage Foundation.
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86, 638-641.
- Ross, P. T., & Bibler Zaidi, N. L. (2019) Limited by our limitations. *Perspectives on Medical Education*, 8(4), 261-264.
- Rothman, K. J. (1978). A show of confidence. *The New England Journal of Medicine*, 299(24), 1362-1363.
- Rothman, K. J. (1998). Writing for Epidemiology. *Epidemiology*, 9(3), 333-337.
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E-J. (2016). Is there a free lunch in inference? *Topics in Cognitive Science*, 8, 520-547.
- Royall, R. (1997). *Statistical evidence: A likelihood paradigm*. Boca Raton, FL: Chapman & Hall/CRC.
- Rubin, M. (2019). What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*. <https://doi.org/10.1007/s11229-019-02433-0>
- Sabeti, P. (2018, January 21). For better science, call off the revolutionaries. *The Boston Globe*. <https://www.bostonglobe.com/ideas/2018/01/21/for-better-science-call-off-revolutionaries/8FFEmBAPCDW3IWYJwKF31L/story.html>

- Schelbe, L., & Thyer, B. A. (2019). Child and Adolescent Social Work Journal editorial policy: Guidelines for authors. *Child and Adolescent Social Work Journal*, 36, 75-80.
- Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, 22(4), 1105-1120.
- Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin*, 105(2), 309-316.
- Sham, P. C., & Purcell, S. M. (2014). Statistical power and significance testing in large-scale genetic studies. *Nature Reviews Genetics*, 15, 335–346.
- Shlonsky, A., D'Andrade, A., & Brookhart, M. A. (2002). JSWE submission suggestions for statistical methods. *Journal of Social Work Education*, 38(1), 5-13.
- Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., ... Nosek, B. A. (2018). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. *Advances in Methods and Practices in Psychological Science*, 1(3), 337-356. <https://doi.org/10.1177/2515245917747646>
- 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. (2012). A 21-word solution. *Dialogue: The Official Newsletter of the Society for Personality and Social Psychology*, 26, 4-7.
- Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2017). Outliers: Evaluating a new p-curve on power poses [66]. *Data Colada (blog)*. Retrieved from: <http://datacolada.org/66>
- 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.
- Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics along. *Psychological Science* 24(10) 1875–1888
- Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*. 143(2), 534-547.
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3, 163084.
- Sober, E. (2008). *Evidence and evolution: The logic behind the science*. Cambridge, UK: Cambridge University Press.

- Soydan, H. (2010). Editorial: A glimpse into contemporary American social work research. *Journal of Social Work, 11*(1), 4-7.
- Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science, 10*(6), 886-899.
- Spellman, B. A., Gilbert, E. A., & Corker, K. S. (2018). Open science: What, why, and how. In J. Wixted & E.-J. Wagenmakers (Eds.), *Stevens' handbook of experimental psychology and cognitive neuroscience* (4th ed.) (pp. 729–776). New York, NY: Wiley.
<https://doi.org/10.1002/9781119170174.epcn519>
- Stang, A., Poole, C., & Kuss, O. (2010). The ongoing tyranny of statistical significance testing in biomedical research. *European Journal of Epidemiology, 25*, 225-230.
- Steenen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702-712.
- Stegenga, J. (2016). Hollow hunt for harms. *Perspectives on Science, 24*(5), 481-504.
- Stegenga, J. (2018). *Medical nihilism*. Oxford, UK: Oxford University Press.
- Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific misconduct and the myth of self-correction in science. *Perspectives on Psychological Science, 7*, 670–688.
- Szucs, D., & Ioannidis, J. P. A. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797.
- Task Force on Social Work Research. (1991). *Building social work knowledge for effective services and policies: A plan for research development*. Austin, TX: Capital Printing Co.
- Tennant, J., Agarwal, R., Baždarić, K., Brassard, D., Crick, T., Dunleavy, D. J., ... Yarkoni, T. (2020a). A tale of two 'opens': intersections between Free and Open Source Software and Open Scholarship. SocArXiv. <https://doi.org/10.31235/osf.io/2kxq8>
- Tennant, J., Francuzik, W., Dunleavy, D. J., Fecher, B., Gonzalez-Marquez, M., & Steiner, T. (2020b). Open Scholarship as a mechanism for the United Nations Sustainable Development Goals. SocArXiv. <https://osf.io/preprints/socarxiv/8yk62>
- Thyer, B. A. (1989). First principles of practice research. *British Journal of Social Work, 19*, 309-323.
- Thyer, B. A. (1991). Guidelines for evaluating outcome studies on social work practice. *Research on Social Work Practice, 1*(1), 76-91.

- Thyer, B. A. (2015a). A bibliography of randomized controlled trials in social work (1949-2013): Solvitur ambulando. *Research on Social Work Practice*, 25(7), 753-759.
- Thyer, B. A. (2015b). Celebrating our 25th anniversary: Research on Social Work Practice is becoming bigger and better. *Research on Social Work Practice*, 25(1), 5-9.
- Thyer, B. A. (2017). Suggested reporting guidelines to improve health-related social work research. *Health & Social Work*, 42(4), 1-4.
- Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social Psychology*, 37(1), 1-2, DOI: 10.1080/01973533.2015.1012991
- Tripodi, T. (1974). *Uses and abuses of social research in social work*. New York, NY: Columbia University Press.
- 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 randomized controlled trials published in medical journals? A Cochrane review. 1(60), <https://doi.org/10.1186/2046-4053-1-60>
- Turner, R. H., Matthews, A. M., Linardatos, E., Tell, R. A., & Rosenthal, R. (2008). Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine*, 358, 252-260.
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76, 285-192.
- van der Zee, T., Anaya, J., & Brown, N. J. L. (2017). Statistical heartburn: An attempt to digest four pizza publication from the Cornell Food and Brand Lab. *PeerJ Preprints*, <https://doi.org/10.7287/peerj.preprints.2748v1>.
- Vazire, S. (2017). Quality uncertainty erodes trust in science. *Collabra: Psychology*, 3(1), DOI: <https://doi.org/10.1525/collabra.74>
- Vazire, S. (2017, October 24). Criticizing a scientist's work isn't bullying. It's science. *Slate*. Retrieved from: <https://slate.com/technology/2017/10/criticizing-a-scientists-work-isnt-bullying.html>
- Wagenmakers, E-J. (2004). A practical solution to the pervasive problem of p-values. *Psychonomic Bulletin & Review*, 14(5), 779-804.
- Wagenmakers, E-J., Lee, M., Lodeqtckx, T., & Iverson, G. J. (2008). Bayesian versus frequentist inference. In H. Hoijtink, I. Klugkist, & P. A. Boelen (Eds.), *Statistics for social and behavioral sciences. Bayesian evaluation of informative hypotheses* (pp. 181-207). New York, NY, US: Springer Science + Business Media.

- Wagenmakers, E-J., Verhagen, J., Ly, A., Bakker, M., Lee, M. D., Matzke, D., Rouder, J. N., & Morey, R. D. (2015). A power fallacy. *Behavior Research Methods*, 47(4), 913-917.
- 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.
- Wagenmakers, E-J., Wetzels, R., Borsboom, D., Kievit, R. A., & van der Maas, H. L. J. (2011). Yes, psychologists must change the way they analyze their data: Clarifications for Bem, Utts, and Johnson (2011). PsyArXiv, Retrieved from: <https://psyarxiv.com/tvarg>
- 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(6), 632-638.
- Wansink, B. (2016, November 21). The grad student who never said “no”. *Healthier & Happier* [Blog]. Retrieved from: <https://web.archive.org/web/20170312041524/http://www.brianwansink.com/phd-advice/the-grad-student-who-never-said-no>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129-133.
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “ $p < 0.05$ ”. *The American Statistician*, 73(suppl), 1-19. Retrieved from: <https://doi.org/10.1080/00031305.2019.1583913>
- Weissgerber, T. L., Garcia-Valencia, O., Garovic, V. D., Milic, N. M., & Winham, S. J. (2018). Why we need to report more than ‘Data were analyzed by t-tests or ANOVA’. *eLife*, 21(7). <https://doi.org/10.7554/eLife.36163>
- Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E-J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3), 291-298.
- White, P., & Gorard, S. (2017). Against inferential statistics: How and why current statistics teaching gets it wrong. *Statistics Education Research Journal*, 16(1), 55-65.
- Yaffe, J. (2019). From the editor — Do we have a replication crisis in social work research? *Journal of Social Work Education*, 55(1), 1-4.
- Young, N. S., Ioannidis, J. P. A., & Al-Ubaydli, O. (2008). Why current publication practices may distort science. *PLoS Medicine*, 5(10), e201.
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. Ann Arbor, MI: University of Michigan.

Zollman, K. J. S. (2019). The scientific ponzi scheme. Retrieved from: <http://philsci-archive.pitt.edu/16264/>

BIOGRAPHICAL SKETCH

Daniel Joseph Dunleavy was born and raised in Broward County, Florida. After graduating from Cardinal Gibbons High School, he attended Florida State University, where he received his Bachelor of Arts in Sociology and Psychology in 2011. He subsequently pursued his Master of Social Work degree from Florida State University, graduating in 2013. Daniel worked as a substance abuse counselor in residential care prior to entering the Ph.D. program. Starting in the Fall of 2020, Daniel will be a Postdoctoral Associate at the Center for Translational Behavioral Science, at the Florida State University College of Medicine.