Methodological Differences Between Psychological Fields and its Impact on Questionable Research Practices

Julian DiGiovanni

A THESIS SUBMITTED TO

THE FACULTY OF GRADUATE STUDIES

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF MASTER OF ARTS

GRADUATE PROGRAM IN PSYCHOLOGY; YORK UNIVERSITY TORONTO, ONTARIO

SEPTEMBER 2019

Abstract

A recent development in research fields, including psychology, is that several studies have called into question the replicability of findings that were thought to be well-established. This phenomenon, termed the “replication crisis” in psychology, is gaining acceptance as a legitimate concern. This paper explores the quality of research from three prominent psychology journals: The Journal of Experimental Psychology: General, the Journal of Personality and Social Psychology and the Journal of Abnormal Psychology, across the years 1995, 2005 and 2015. The quality of research was determined through creating individual p-distributions, similar to the methods of Masicampo & Lalande (2012). This paper uncovered that there was evidence regarding the use of questionable research practices (QRPs) since 1995. Overall, the quality of each journal's research appeared to be increasing as the years progressed

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Methodological Differences Between Psychological Fields and its Impact on Questionable Research Practices

A recent development in research fields, including psychology, is that several studies have called into question the replicability of findings that were thought to be well-established. This phenomenon, termed the "replication crisis" in psychology, is gaining acceptance as a legitimate concern. Scientists are beginning to look into psychological research and social science research as a whole. Although some differentiate the many branches of psychology (i.e. social psychology, experimental psychology, clinical psychology etc.), many do not. When the separate subdisciplines of psychology are looked at as a whole, the methodological differences and nuances of each subdiscipline disappear. This allows then, for the practices and methodologies of select psychological fields to cast a shadow on the rest of psychology. Take the 2010 article written by social psychologist Amy Cuddy regarding "power posing" and how body language can generate feelings of confidence that would otherwise be absent. This article generated a great deal of scrutiny after Cuddy was accused of engaging in questionable research practices (QRPs) (Dominus, 2017). More recently, prominent food researcher and ex-professor at Cornell, Brian Wansink, resigned due to the discovery that he had engaged in questionable research practices. Wansink wrote a blog post detailing how he had an "interesting" dataset that he "knew", if looked at the right way, could make for interesting research papers. What Wansink actually was doing was using problematic statistical techniques that raised the probability that he would achieve statistically significant results, that may not necessarily be real effects (Bartlett, 2017). Events like this not only cast doubt on the particular subfield, but psychology as a whole. The reality is, however, that psychology is comprised of many different subdisciplines and fields that each have their own histories, subcultures and practices that can differ significantly from the others. Although some areas may not be "up to par" regarding research practices, it does not

speak to psychology as a whole. This paper attempts to uncover the quality of research and methodological differences between three areas of psychology; personality and social psychology, abnormal psychology and experimental psychology. These methodological differences will serve as an explanation to why there may be differences in the quality of each fields' research.

The merging of statistics and social science research is a relatively modern event. In the present day, it is the expectation that psychological research will use statistical analysis to support its claims. There are still some niche researchers who debate whether psychological research should even be quantified, let alone use statistics (Tafreshi, Slaney, & Neufeld, 2016). That being said, across the vast majority of the discipline, it is required that any type of research project employ the use of statistics. However, it has been demonstrated numerous times that psychology researchers may not be fully competent statisticians. For instance, Gigerenzer (2004; 2018) demonstrated widespread misunderstanding regarding the meaning of *p*-values. In terms of the "replication crisis" psychology is undergoing, many papers have pointed to the poor statistical literacy researchers possess (Gigerenzer, 2004; Karen & Lewis, 2014), which may be interpreted as a core cause. Similar to a virus requiring perfect conditions to thrive and transform into a full-blown epidemic, the issue of replicability needed the correct conditions to transform into a "crisis". Arguably some fields have provided ideal conditions in which questionable research practices (QRPs), statistical illiteracy, pressure to publish, strict adherence to methodologies and resistance to change enabled this problem to morph into a crisis. It should be emphasised that some domains of science have allowed for this, while others remain affected to a lesser degree. For example, replication studies are valued equally as high compared to novel studies in the natural sciences (Madden, Easley, & Dunn, 1995). Psychology as a whole has been

identified as being affected by this crisis, but again it should be noted that psychology should not be looked at as a whole, but rather as a grouping of individual fields and disciplines.

Although some research suggests that fields outside the social sciences are being affected by this crisis (Begley & Ellis, 2012; Fanelli, 2009; Ioannidis, 2005), most of the discussion has centered around the social sciences, specifically social psychology (Dominus, 2017; Stanley, Carter, & Doucouliagos, 2018). Psychology may have the perfect environment to foster such a problem. Competition for career advancement emphasizes the need for publication; as many refer to it, "publish or perish". This need for publication interacts with the fact that psychology tends to publish only significant results, 90% of psychological studies reported only significant results. Although other disciplines share this problem, psychology is the discipline with the highest proportion of significant results in its literature (Fanelli, 2012). The need for publication has put the .05 significance level on a pedestal, making statistical significance the primary goal of most research. When the aim is to achieve a *p*-value less than .05, shortcuts can be taken on the road to achieving this goal, as demonstrated by Simonsohn, Nelson, and Simmons (2011). A number of tactics can be employed to aid in the quest for statistical significance. These tactics are often referred to as data dredging, data fishing, data butchering, but most commonly, p-hacking. As will be further discussed in this paper, engaging in p-hacking essentially raises the true alpha level of one's experiment, drastically increasing the odds of a false positive (type I error).

This paper aims to dissect and demonstrate the current status of psychological research by observing the individual methodologies of three distinct fields: abnormal psychology, personality and social psychology and experimental psychology. First, the state of the research of each field must be assessed. Through a method termed *p*-curving, created by Simonsohn, Nelson

& Simmons (2014), possible instances of the use of QRPs may be identified. *P*-curving is the process of plotting the frequencies of statistically significant p-values from published studies and then observing the resulting distribution (Simonsohn, Nelson & Simmons, 2014). The shape of the resulting distributions can help identify areas where there are more p-values in the .04-.05 range than one would expect, given an alpha level of .05. Simonsohn, Nelson & Simonsohn (2014) demonstrated through numerous simulations that even with inadequate power and small effect size, there should be a significantly higher proportion of p-values in the 0 - .02 range, and, as the power and effect size increase, this effect becomes even more pronounced with fewer p-values near to .05. Where there are more than expected frequencies of p-values in the .04-.05 range, a question about the cause of the anomalous distribution arises.

It is hypothesised that differences between the three fields will inevitably exist. These differences will be explained through the different methodologies and cultures that each field possesses. To determine these cultures, research papers were taken and analysed to assess 1. How statistics were used in each field 2. The proportion of studies in each field tagged as replications and 3. The change through time, from 1995, 2005 and 2015. Understanding the culture and methodologies of each field can give one insight into the reasons that one field may foster or discourage the use of QRPs.

Understanding the Replication Crisis

Psychology's statistical background

To understand this topic fully one must first understand the background and sequence of events that had to take place to put psychology in the predicament it finds itself today. This "story" will begin with the marriage of statistics and social science research. The use of statistics can arguably be considered one of the core causes for this crisis. Whether it be the misuse of

statistics or the general lack of knowledge regarding the use of statistics, which will be further explained in this paper. Statistics were not always used to support empirical claims; so how did psychology go from not needing statistical tests to statistical analysis becoming a requirement of psychological research? Gigerenzer (2018) provided a brief summary of what psychology looked like pre-statistical analysis. He states that the typical psychology article would, in detail, report only a single case with the publisher being the one who would be tested. This is accurate in the sense that Wilhelm Wundt commonly stated that any technician could run an experiment, but it took an expert to be able to be tested (Rieber & Robinson, 2001). In fact, Gigerenzer goes on to state "you would have never caught Piaget calculating a *t* test" (Gigerenzer, 2018, p. 4) and explained that the evaluation of whether two means differed was based on judgements, taking many factors into account, not statistical tests; this was the basis of all the classical discoveries in psychology.

The use of statistics for inference began in the 1920s when Sir Ronald Fisher began using statistical inference while working at Rothamsted Experimental Station where he analysed data from crop growth. Gigerenzer (2004) described the hypothesis testing that Fisher proposed be used for inference by summarising it using three rules:

"*Fisher's null hypothesis testing*:

1. Set up a statistical null hypothesis. The null need not be a nil hypothesis (i.e., zero difference).
2. Report the exact level of significance (e.g., p=0.051 or p=0.049). Do not use a conventional 5% level, and do not talk about accepting or rejecting hypotheses.
3. Use this procedure only if you know very little about the problem at hand." (Gigerenzer, 2004, p. 4).

Fisher's proposal was not without its critics however, Egon Pearson commonly criticized Fisher for having no measure of statistical power in his tests and stated that Fisher's test interpreted significance too subjectively and thought that significance should be more of a decision and not a belief (Gigerenzer, 2018). This meaning that Pearson believed that significance should be a yes or no decision based off of specific criteria, whereas Fisher was commonly cited as believing significance was open to interpretation and should not have a clear criteria. Jerzy Neyman and Egon Pearson soon rebutted with what they thought would be a superior tool for inference. Gigerenzer (2004) also described this in three rules:

" *Neyman-Pearson decision theory:*

1. Set up two statistical hypotheses, H1 and H2, and decide about α, β, and sample size before the experiment, based on subjective cost-benefit considerations. These define a rejection region for each hypothesis.
2. If the data falls into the rejection region of H1, accept H2; otherwise accept H1. Note that accepting a hypothesis does not mean that you believe in it, but only that you act as if it were true.
3. The usefulness of the procedure is limited among others to situations where you have a disjunction of hypotheses (e.g., either μ1 = 8 or μ2 = 10 is true) and where you can make meaningful cost-benefit trade-offs for choosing alpha and beta." (Gigerenzer, 2004, p. 5).

Using statistics for inference soon spread through the United States allowing psychologists to get a hold of these ideas and eventually incorporated them into their research. These two competing beliefs became combined into a hybrid theory that we refer to today as null hypothesis significance testing (NHST). Psychologists took rule one from Fisher, rule two from

Neyman and Pearson and collectively disregarded the third rule that explains the limitations of each test. Gigerenzer (2004) refers to NHST as the “null ritual”. He commonly asserts that the behaviours of researchers are incredibly ritualistic in the sense that researchers blindly follow the same steps and conventions over and over and rarely deviate from this “ritual”. NHST today consists of always testing a research hypothesis against a null hypothesis of no effect and to always use 5% as a convention for rejecting the null. The issue is that NHST arguably is not the best way to conduct research, rather there is a larger “toolbox” of statistical methods and tests that psychologists can use that would be more optimal and effective (Gigerenzer, 2004).

There are more issues with NHST than just not being the most optimal tool for the analysis of psychological research. There are also many misconceptions that most students, researchers and professors hold regarding NHST and statistics. There are plenty statistical misconceptions that go unchallenged and are ultimately passed down to students from their mentors and professors. Take for instance the confusion that surrounds p-values and what their use is. Gigerenzer (2004) distributed a quiz consisting of six questions pertaining to *p*-values. He gave this quiz to three groups of people: undergrad psychology students who had taken at least one statistics course, psychology professors who don’t teach statistics and psychology professors who do teach statistics. All six statements were false and, surprisingly, 100% of undergrad students answered true to at least one of these questions, with the mean number of statements answered true to, being 2.5. Perhaps even more shocking is that 90% of psychology professors answered true to at least one of these questions (M = 2.0) 80% of statistics professors answers true to at least one question (M = 1.9).

Many of the questions that participants agreed were true were related to p-values being markers of effect size, which is not the case at all. P-values don’t tell you anything about a

hypothesis, they only tell you the probability of obtaining the data you have (or more extreme data) if the null hypothesis is true (no effect). Students are inheriting these misconceptions from their instructors, so the question is where did their instructors acquire them? It turns out these illusions can be pinpointed to the first textbooks introducing psychologists to null hypothesis testing more than 60 years ago. Guilford's Fundamental Statistics in Psychology and Education, first published in 1942, was probably the most widely read textbook in the 1940s and 1950s. Guilford suggested that hypothesis testing would reveal the probability that the null hypothesis is true. "If the result comes out one way, the hypothesis is probably correct, if it comes out another way, the hypothesis is probably wrong." (Guilford, 1940, p. 156). It didn't stop there; this continued through time with early authors promoting the illusion that the level of significance would specify the probability of hypothesis including: Anastasi (1958, p. 11), Ferguson (1959, p. 133), Lindquist (1940, p. 14), Miller and Buckhout (1973, p. 523), Nunnally (1975, pp. 194–196) (Keren & Lewis, 2014). Psychologists have taken concepts they do not fully understand and have incorporated into them into rigid "ritualistic" methods that use these concepts in ways they were never intended.

The aim of inferential statistics was to assert with some degree of probability that two populations may differ from each other. As noted earlier, Fisher commonly stated that using a rigid cutoff was unwise and that exact probabilities should be reported. Fisher's ideas have been continually misinterpreted until now in the present day, a cutoff of .05 is virtually always used in psychology. On the contrary, there are times in the past where Fisher was quoted stating that a results that are twice the standard deviation could be accepted as significant. However, Fisher was usually quick to ascertain that the cutoff should always be based off of the odds that one thought were acceptable, for example if 1 in 20 was not stringent enough for the question at

hand, then the cutoff of 1 in 50 could be used (Cowells & Davis, 1982). Cowells and Davis (1982) argued that the answer as to why 5% has become the norm for phenomena to be considered significant, the question has to be thought in terms of how we interpret probability. This means that researchers must believe an occurrence that happens 5% of the time must be a rare occurrence. However, as will be explained further, researchers can often find ways that make this 5% cutoff much higher without readers and researchers even knowing.

Better practices of statistics

There are camps of researchers that believe NHST is in fact a horrible way to conduct research. In terms of *p*-values, there is a strict and narrow way in which they can be interpreted. As stated above, the only meaning a *p*-value will convey, is the probability in which one will collect the data at hand (or more extreme) given that the null hypothesis is true (no effect). Kline (2004) goes into great detail outlining common fallacies that surround *p*-values, stating beliefs such as *p*-values are markers of effect size, rejecting the null hypothesis confirms the alternative hypothesis and *failing* to reject the null hypothesis indicates that two populations are equivalent and the effect size is zero. These beliefs are all false, *p*-values do not hint in the slightest way about any size of effect; a smaller *p*-value does not imply a large effect and a larger *p*-value does not imply a small effect. Also, in terms of rejecting and failing to reject the null hypothesis, a *p*-value below .05 does not confirm nor refute a hypothesis, a *p*-value only speaks in terms of the data at hand. This is why replication is incredibly important, in terms of NHST, for a hypothesis should be supported through many replications and many datasets, a single *p*-value conveys almost no meaning! (Cumming, 2014).

Cumming (2014) visually represented how meaningless a *p*-value can be. It makes logical sense that if an experiment returns a specific *p*-value, then a replication of that same

experiment should result in the same, or similar *p*-value, but this is not the case. Cumming exhibited just how unpredictable *p*-values can be; through simulating 25 experiments with the same populations, Cumming demonstrated that a *p* can take on almost any value. The two populations consisted of 32 participants each and had a real effect with an effect size od $d = .5$ (which is considered a medium sized effect). The simulations however, produced *p*-values ranging from .586 to <.001. Cumming went on to argue that solely relying on *p*-values for interpretation is unwise and makes an argument that confidence intervals (CIs) should be used instead. The argument that CIs should be used in favor over *p*-values is not a new one (Cohen, 1990; Cohen, 1994; Lakens, 2013) and intuitively, the argument makes sense; talk about research in terms of effect size and confidence intervals instead of *p*-values because the former actually conveys valuable meaning.

For some time now individuals have advocated for better statistical practices. Cohen (1990) outlined what he had learned regarding the proper way to conduct research in terms of research design and use of statistics. He advocates that "less is more", that research designs with a ridiculous number of independent and dependent variables will in fact increase ones false positive rate well above the self proclaimed .05. He also explained that moving forward less emphasis should be placed on *p*-values and the attention should shift to CIs and effect sizes. But what Cohen does extraordinarily well, is explain just how odd NHST really is. He explains using an example that we can never confirm a hypothesis through the use of NHST, instead we can only conclude that we have failed to find any data that can refute our hypothesis, two very different things. This paper was written in 1990 and definitely was not the first of its kind (advocating that statistics are carried out poorly and that CIs and effect sizes should *always* be reported) and not the last (see Lakens, 2013). If this is so, why in the present day are we still

faced with the same problems? Why have we not shifted completely from *p*-values and speak only in terms of CIs and effect size? Why are we still interpreting the results from NHST incorrectly?

The Perils of Power

A topic akin to the idea that statistics are done poorly is the idea that statistical power seems to be completely disregarded in psychological research (Cohen, 1992; Rossi, 1990) and many other fields as well (Reinhart, 2015). The idea of the importance of statistical power arose in the 1960's when Jacob Cohen calculated the power of all of the studies published in *The Journal of Abnormal and Social Psychology* for the year 1960. What he found was that the average power of each study was 0.48 (Cohen, 1962), basically, the odds of finding an effect that actually existed were similar to that of a coin flip. Sedlmeier and Gigerenzer (1989) set out to determine whether if the preceding published literature regarding the importance of power had any effect on researchers judgement but found that practically 30 years since Cohen's study, the average power had actually decreased. However, it was speculated that the decrease in power was due to researchers employing more complex research designs that increased the number of comparisons making the need for certain pairwise adjustments necessary which, in turn, decreased the overall power. Reinhart (2015) does a satisfactory job at detailing why researchers still to this day seem to disregard statistical power. He stated that researchers probably still follow old rule of thumbs and believe that if an effect does not appear in a sample of, lets say, 50 participants, then the effect is probably not relevant. However, psychology has been known to study topics that are small in effect size, and to reliably detect these effects with the recommended 80% power, one would need upwards of 400 participants.

Although being adequately powered is crucial when attempting to uncover an effect, inadequate power can also work in the opposite direction. Underpowered studies are the culprit of an effect known as a type M error (M for magnitude) or truth inflation. To briefly explain how truth inflation occurs, suppose a researcher is testing a medication to see if it can alleviate the symptoms from the common cold. However, the researcher uses only a sample size of $n = 10$. There is a chance that the medication being studied has no effect but the researcher could in fact have gotten "lucky" and selected the 10 participants who miraculously get relieved of their symptoms the next day, just by chance. Had this researcher selected a sample size of 10,000, they would have been able to see that the medication really has no effect on reducing these symptoms. Truth inflation often occurs in fields where there is competition to publish, making the most exciting results the most attractive ones, giving them president to be published. This problem is not a problem only psychology problem, other fields see this too, Ioannidis (2008) claims that majority of the most cited medical research are the product of truth inflation. Could it be that if journals and the academic publication system did not look to publish the most "exciting" and novel results and instead focused on admirable practices and methodologies, these problems would cease to be an issue?

The perils of publication

Producing an exciting and novel study will increase the chances of publication. Publishing such a study may in turn increase the citations of said study Increased citations will then result in a greater impact factor for the journal that publishes this study thereby supporting the "eliteness" and "prestige" of the journal. A journals impact factor is measured by counting the number of citations from a particular journal from a given year. Only citations of articles published within the two preceding years of the year in question are counted. This number is

then divided by the total number of articles published in the two preceding years. All of the major journals are then ranked based off of the this calculation. This idea of tracking the number of citations to see which journals are home to the "best" research seemed like a good idea until now, in the present day, many advocate that the system of impact factor can be, and is, easily abused. Moustafa (2015) has argued that journal impact factor has now become a tool to promote or unfairly criticize individuals, journals, universities etc. Moustafa goes on to explain how some editors can be "elitist" and are overly selective in evaluating submissions with the journal's impact factor in mind. One can deduce the importance this places on finding interesting and statistically significant results, ultimately adding pressure to researchers, fueling the publish or perish ideology. Others have endorsed the viewpoint that using this system is flawed. Hossenfelder (2017) comments that the role of science should be to test hypotheses and then either keep, revise or discard the hypotheses. If the incentive is to solely create research that will be cited numerous times, this adds an aura of negativity around returning non-significant results. This emphasis on significant results opens the door to a number of techniques that allow one to artificially achieve statistical significance. These methods can be summed up by the umbrella term; p-hacking and HARKing (hypothesizing after the results are known). Whether engaging in these practices unknowingly or not, the pressure to publish novel and interesting results may have fueled the use of these tactics, which is essentially the core problem of this "crisis".

Questionable research practices

To begin, one must first understand what constitutes questionable research practices and the effect is has on ensuing research. Although the term questionable research practices sounds menacing, many of the practices and methods that are considered questionable can quite possibly be engaged in without one knowing so. Through using an alpha level of .05, it is inferred that for

psychological research, the false positive rate would be 5%, just as the alpha level suggests. However, Simmons, Nelson and Simonsohn (2011) demonstrated that when engaging in certain practices, the alpha level dramatically increases. Simmons et al., illustrated this by statistically supporting hypotheses that are impossible, one of them being participants were a year and a half younger after listening to "When I'm Sixty-Four" by The Beatles. How is it possible that seemingly objective statistics can be used to support such obviously false hypotheses? As Simmons et al., describe it, these hypotheses become statistically supported through increased "researcher degrees of freedom", or in other word, how they chose to analyse the data.

To support this claim, Simmons et al., conducted 15,000 simulations using data that *should* show a non-significant effect; employed the use of questionable practices and then used the amount of false-positives to gauge what the actual false positive rate was. The techniques they studied were: flexibility in analyzing two dependant variables, adding more participants and then re-testing if the initial test is non-significant, freedom in controlling for independent variables or interactions and testing and combining covariates. What they found was that when using these techniques, on average they about doubled the expected false-positive rate. The most alarming discovery was that when these techniques are used in conjunction with one another they can inflate the false-positive rate to an astounding 61%. That suggests that a researcher will have a higher probability of finding an effect when none exists then to truly reject the null hypothesis.

These are only a few of the many "researcher degrees of freedom" that exist. Most of the inflation regarding false-positive rates occur when a researcher tests data multiple times which increases the familywise error rate. But why are multiple comparisons such an issue? Researchers have agreed to a false-positive rate of .05 or 1 in 20. This means that for every 20

statistical tests performed, on average there will be one that is the result of a false-positive. Now, imagine researchers are testing a phenomenon that in reality has no effect and imagine that they are controlling for a number of variables and interactions that in total, equal 20 statistical tests. Imagine these individual tests as marbles in a jar, all of the blue marbles (19) reflect the true non-significant results, but there is one red marble that represents the false positive outcome. The researcher can reach into the jar and look at their initial hypothesis and voila, the marble is blue and the test is non-significant. But what happens as the researcher looks to uncover the rest of the combinations of variables and interactions? There are now only 19 marbles in the jar, meaning the new false-positive rate is now 1 in 19 and not the predetermined 1 in 20. This continues for each marble the researcher takes from the jar, that is why multiple comparison corrections are so important.

Schooling has taught that when in situations such as post-hoc testing after a statistically significant ANOVA, one is carrying out multiple comparisons and must adjust using an appropriate method. However, multiple comparisons are not always obvious. In a situation where one is testing multiple combinations of dependent variables and covariates or testing and then re-testing, they are increasing the amount of comparisons and the increased false-positive rate reflects that. The reality is that statistics are more sensitive to these researcher degrees of freedom than researchers would like to believe. Even something as simple as removing an outlier can have an effect on the statistical significance of a dataset. The main issue is, however, that researchers are not encouraged to be transparent about their analyses. Currently, there is no way to know with certainty what specific techniques a researcher may have employed. For example, a researcher could have tested multiple combinations of dependent variables but only report the dependent measures that were statistically significant.

Now, going back to the example of the experimenter drawing marbles from a jar, imagine this experimenter did not think of a hypothesis before hand, and on his 17th draw from the jar, the marble was finally red. This outcome would have had a 1 in 4 chance of being the result of a false positive, a 25% chance! This experimenter can then formulate a hypothesis off of this significant result, write a paper based off of this result, excluding the fact that he ever conducted other tests - that there were only three other marbles in the jar. This is essentially what HARking is and why it can be such a problem. By definition HARKing is presenting a post-hoc hypothesis as an a priori hypothesis (Kerr, 1998). Kerr (1998) goes on to confirm that a main issue with HARKing is the fact that results may be the product of false-positives and calls for increased replication a solution this this problem.

Although these practices exist, how prevalent are they in psychology? Psychology has already been identified as a field where engagement in QRPs are unusually high (John, Loewenstein, & Prelec, 2012). John et al., (2012) used incentives to promote truth telling in regards to engaging in QRPs. Participants were asked if they themselves had engaged in specific questionable practices and to provide an estimate of how prevalent they believed the engagement of QRPs were in the field of psychology. Based on participants' admissions and how prevalent they believed QRPs were in the field, an estimation was able to be calculated of the true prevalence of engagement in QRPs. Surprisingly, John et al., found that 78% of psychological researchers admitted to at some point having failed to report all dependent measures. Further 72% admitted to collecting more data and re-testing once the initial results were not significant and finally, 67% admitted to selectively reporting significant results and excluding data after looking at the impact of doing so. These are just the numbers for the researchers that have

actually admitted to engaging in these practices, John et al., estimate that the number in reality is probably much greater.

The problem of replication, historically

For some time now there have been questions regarding the reproducibility of scientific literature. These questions have revolved around the soundness of research in terms of research design, methods, and experimenter biases. In the 1980's there was discussion regarding the quality of clinical trials and use of statistics in the medical field (Altman, Moore, Gardner, & Pocock, 1983). Altman et al., commented on the alarming truth that medical research contains analyses that have been carried out without the aid of a statistician. This worry is still expressed today. Gigerenzer (2014) demonstrated that most psychological researchers and professors (including statistics professors in the field of psychology) still have difficulty regarding the understanding of *p*-values and what they represent. Many researchers agreed that a *p*-value can be used to express the probability of one's hypothesis to be true, when in actuality all a *p*-value represents is the probability of obtaining the data one has obtained (or more extreme) if the null hypothesis were true. Poor statistical practice may have caught up with the respective fields to the point where now, there are groups of researchers whose academic purpose has shifted to exposing the poor statistical practices of researchers (Marcus & Oransky, 2018).

In the early 1990's researchers argued that with advancing technology came the ability to cite every step a researcher takes to get to the final destination of their analysis. Claerbout and Karrenbach (1992) maintained that with the aid of word processing and command scripts, researchers have the ability to reveal all the steps for other researchers to be able to recalculate all figures and tests. Claerbout et al., set a list of milestones in hopes that they may be achieved by the end of the 1990's. This list included goals such as educating researchers on how to include

pushbuttons on electronic articles that demonstrate the steps to recreate every figure and to produce all new documents in an identical format. It is clear that in the present day, this goal has not been met. Barely any researchers are transparent enough with their data to the point that their analyses can be completely reproduced. This does serve to show however, that reproducibility was on researchers' minds well before the 2000's, where one of the major turning points in this crisis lay.

In August of 2005, medical research professor John Ioannidis published an article titled "Why Most Published Research Findings are False". In this article, Ioannidis (2005) made a bold claim that most research findings are the result of false positives. The logic being, in research areas where statistical power is low, and where there is a larger proportion of "false" to "true" hypotheses being tested is very small, then it can be inferred that most of those findings can be attributed to Type I errors (here "false" and "true" do not refer to research results, but instead refer to actually being false or true in real life). To better explain this, imagine a field where the ratio of actually false hypotheses to actually true hypotheses being tested was 1:100. Even with an alpha level of .05, majority of the findings from this field would be actually false due to the large discrepancy of true to false hypotheses being tested. The reality is that researchers will never really know which hypotheses are actually true vs. actually false, but researchers do have control over the power of their studies and psychology has been a notorious offender for producing underpowered studies (Rossi, 1990).

Ioannidis' 2005 article stirred the scientific landscape. Today it has been viewed close to 2 million times. A partnering idea was one by Rosenthal (1979) where he labelled the publishing of only significant results "the file drawer problem". As mentioned earlier, psychology has a problem of publishing mainly significant results, Rosenthal asserted that "journals are filled with

the 5% of the studies that show Type I errors, while researchers' file drawers are filled with the 95% of the studies that show non-significant results" (Rosenthal, 1979, p. 1). Rosenthal discovered a way to determine how many non-significant results would have to exist, to determine whether the literature on a given topic has a high chance of being the result of a Type I error. On topics that are robust and heavily researched, the amount of non-significant results would have to be in the thousands, but for more lightly research topics, there may only have to be non-significant results in the single digits for them to be the result of type I errors (Rosenthal, 1979).

These claims were reinforced in August of 2015 when the Open Science Collaboration (OSC) published a paper estimating the reproducibility of psychological research. OSC (2015) attempted to replicate 100 studies from three prominent psychology journals: *Psychological Science, Journal of Personality and Social Psychology* and *Journal of Experimental Psychology: Learning, Memory and Cognition*. Their analysis suggested that out of these articles, only 47% of them were successful in being replicated. On top of this, effect sizes were also compared for the studies where it could be calculated (99) and discovered that 82 of these studies showed a higher effect size than its replication counterpart. Many mass replication attempts have preceded this study and since taken place, most notable are the many labs projects. Many Labs 1 (2014) looked to replicate 13 of the classic findings in psychology such as anchoring and sex differences in attitudes toward math. What they found was that they were able to successfully replicate 11 of the 13 studies (Klein et al., 2014). However, Many Labs 2 (2018) attempted at replicating a total of 28 studies and these results were more in line with OSC's findings; only half of the studies were successfully replicated (Klein et al., 2018). Replication is supposed to be the safety net of science, allowing for hypotheses to either be refuted or supported further. That being said, some

psychological journals will go as far as not publishing replication studies. A question must be asked as to why the replication rate of studies are so painfully low in psychology. This leads researchers to turn to the idea that QRPs allow these studies to get published in the first place. This notion has fueled a type of counter-culture that has devoted itself to creating tools that can uncover analyses where QRPs may have been employed.

Tools used uncover questionable practices

The past decade has brought into use, many forensic techniques that aid in determining the quality of research (Brown & Heathers, 2016; Heathers, 2017; Masicampo & Lalande, 2012; Simonsohn, Nelson & Simmons, 2014). The first wave of forensic techniques began by plotting p-values and observing the resulting distributions. P-distributions were first plotted and studied in the early 2000's where the p-distribution was found to resemble an exponential curve; that is more p-values would pile up in the .001-.01 range and gradually fan out as the p-values increased (Berger, 2001; Cumming, 2008). Masicampo and Lalande (2012) then took this idea and put it to practice by plotting the p-values from the 2008 issues of *The Journal of Experimental Psychology: General* (JEPG), *The Journal of Personality and Social Psychology* (JPSP) and *Psychological Science* (PS). Their method of choice was fitting a line of best fit to the plotted p-values and then calculating the residuals to see where, if any, discrepancies lay. What they found was that the resulting distributions did follow the exponential curve that both Berger (2001) and Cumming (2008) described. However, using a chi-square test to assess if any discrepancies exists in the residuals of the p-distribution, they found that in all three journals there was evidence to believe that there were an unusually high frequency of p-values in the .045-.05 range of the distribution. Masicampo and Lalonde (2012) cited that the result of this study may be evidence for publication bias; that is that researchers and reviewers may place an unreasonable

priority on achieving statistical significance making them turn to use of researcher degrees of freedom (Simonsohn et al., 2011) in order to achieve this goal of statistical significance.

Simonsohn, Nelson and Simmons (2014) took this idea a step further by creating the *p*-curve. The ideas of Rosenthal (1979) had already been established in terms of the file drawer problem, but Simonsohn et al., built on this and hypothesised that it is not entire studies that researchers are keeping in their file drawers, it is individual analyses. They thought that in some areas, researchers tested a hypothesis a number of different ways, ignoring all the non-significant results and publishing only the significant ones. Building on top of their 2011 article which already described the problem of computing multiple tests and what effect that has on the false-positive rate, Simonsohn et al., set out a way to determine whether the research on a particular topic or hypothesis were the result of false-positives or actual legitimate findings.

P-curving is the process of plotting the frequencies of statistically significant p-values from published studies and then observing the resulting distribution (Simonsohn, Nelson & Simmons, 2014). The shape of the resulting distributions can help identify areas where there are more p-values in the .04-.05 range than one would expect, given an alpha level of .05. Nelson & Simonsohn (2011) demonstrated through numerous simulations that even with inadequate power and small effect size, there should be a higher proportion of p-values in the .001 - .02 range, and, as the power and effect size increase, this effect becomes even more pronounced with fewer p-values around the .05 range. Where there is a more than expected number of p-values in the .04-.05 range, a question about the cause of the anomalous distribution arises. One possible explanation is that the researchers engaged in practices that actually raised the true alpha level well above .05. Notice Figure 1, each graph is considered a *p*-curve. When there is no effect, logically, the distribution of p-values should be uniform because there is an absence of an effect.

However, when QRPs are used one can notice the left skew, with the highest proportion of p-values just below .05. Simonsohn, Nelson, and Simmons (2014) explain this phenomenon by researchers who engage in QRPs having little ambition to achieve p-values in the expected range of significance (.01 range) and instead only aim for a p-value just under the cutoff, .05.

There have also been recent attempts at recreating what distributions and datasets may have looked like through the aid of computer programs. GRIM, GRIMMER and SPRITE all set out to achieve this in their own ways. GRIM and GRIMMER both work to to determine if the calculations in a given study are mathematically possible. The tests rely on the fact that given a dataset containing any given number of integers, there are only certain values that the mean of the dataset can be. Brown and Heathers (2016) used the GRIM tests to assess 260 article from *Psychological Science, Journal of Experimental Psychology: General* and *Journal of Personality and Social Psychology.* What they found was 32 of these articles contained at least one mean that was impossible to obtain, given what the researchers reported. On top of this, they found that 16 of the articles contained multiple means that were impossible. SPRITE on the other hand, works to rebuild a possible dataset for a given study based off of the information the study reports (Heathers, 2017). Heathers (2017) used sprite to analyze the statistics Brian Wansink reported in one of his famed studies. Wansink reported that giving special names to foods like carrots would make children more inclined to eat them. However, when SPRITE was used to create a possible dataset for this study, what they found was that for what Wansink reported to be true, some children would have had to eat in excess of 60 carrots!

Possible Solutions

There is no shortage of proposed solutions for the problems that psychology faces (Antonakis, 2017; Aschwanden, 2015; Cumming, 2014; Nosek, Spies & Motyl, 2012; Wagenmakers et al., 2017). Nosek et al., (2012) has an interesting viewpoint where he speaks on the idea that what is good for scientists may not be good for science, ie., publishing new and novel studies at an impossible rate. They speak to the idea that the accumulation of knowledge about nature is the primary objective of science and how "revealing" something new advances that goal whereas reaffirming does not. This adds to the issue of under-replication stressing the idea that replications are not as important to novel research. The solution they speak of, however, is one of openness. Openness will increase accountability as well as critique the methods used in a particular study. Open data and open methods and tools are a common trend amongst advocated for change in the psychological field. Another being the badge system presented by the Open Science Framework (OSF). OSF will award badges that will appear on researchers articles if the researcher has followed their guidelines. The badges are for: open data, open materials and pre-registered hypothesis. The idea is that the more researchers who partake in this the more common it will be until it becomes common practice to have these badges on your research. There is also a push for editors to only accept articles that have at least one of these badges on them, which in turn would reinforce the transparency that is needed in psychology today.

Method

As explained above, past research has paved the way toward the ability to identify areas where QRPs may be utilized in research areas. In the present study, methods similar to Masicampo and Lalonde (2012) were used with every article from *The Journal of Abnormal*

Psychology (JAbP), *The Journal of Personality and Social Psychology* (JPSP) and *The Journal of Experimental Psychology: General* (JEP:G) from the years 1995, 2005 and 2015. The *p*-values of every article were plotted and the resulting distributions were explored. To extract all of the *p*-values from their respective articles, a program called Statcheck, created by Dutch researchers Michèle Nuijten and Sacha Epskamp. Statcheck takes the tests statistics reported in research articles and recalculates the *p*-values. It does so by extracting the degrees of freedom while identifying the type of test that is being conducted (*t, F, X^2* etc…). The *p*-value is then recalculated for each statistic and reported in a chart. The chart was then brought into Rstudio, where the *p*-values could be plotted in a histogram, and the resulting distributions could be analyzed. The resulting graphs were truncated, with the x-axis' beginning at .01, all groups had the majority of *p*-values in the 000 - .001 range, making the .01 - .06 range less visible.

In addition to collecting the *p*-values from every article, information was also taken from each article regarding the individual practices they employed. Using the TidyText package in R, relevant data was extracted from each article. The information that was taken from each subset of articles was: the types of statistical tests used, reporting of confidence intervals and effect sizes and the controlling of covariates. This information was extracted because it was hypothesized that exceptional statistical practices (i.e. the use of non-parametric tests, reporting of effect sizes, use of confidence intervals, etc…) could be used as an explanation if there were differences between the graphs of *p*-values. The basis of this analysis was extracting word counts of target words. Each article was put through the program in the R software and had a number of words counted within it. The words counted were: statistical tests (i.e. ANOVA, ANCOVA, chi-square, Kruskal-Wallis one -way ANOVA, Mann-Whitney *U*, MANOVA, MANCOVA, regression, *t*-test and Wilcoxon signed-rank test) and the word "covariate" for the reason that past research

(e.g., Simonsohn et al., 2014) has shown that the unprincipled use of covariates can lead to elevated Type 1 error rates. Also, Scholars Portal (a database for research articles) has put tags on articles regarding their content since the year 2006. These pages of tags and descriptions of each article were taken and words regarding replication (replicated, replication, reproduce, imitate and confirm) were counted.

In total, 684 articles were examined; 251 from JAbP, 113 from JEP:G and 320 from JPSP. For a more complete breakdown of journal amounts, refer to Table 1; For the breakdown of the number of *p*-values taken across journals, through years, see Table 2. In addition to looking at the content of each journal, the histories of the three fields were also taken into account. This will allow for a full understanding of the methodological differences of each journal, which may be used as an explanation for each journal's respective practices.

Results

Journal of Abnormal Psychology

1995

All 76 articles from the *Journal of Abnormal Psychology* had their *p*-values extracted and graphed, see Figure 2 for the resulting distribution. As expected there was a higher proportion of *p*-values in the .00 - .01 range, but there were bumps at the .03 - .035 and .04 - .045 ranges.

Regression was the was the most widely mentioned test in JAbP 1995, with the *t*-test and ANOVA being second and third. Surprisingly, there were a few mentions of the Mann-Whitney *U* test.

2005

All 76 articles from the *Journal of Abnormal Psychology* had their *p*-values extracted and graphed, see Figure 3 for the resulting distribution. As expected there is a large proportion of *p*-

values in the 0 - .01. There was a less dramatic bump at the .04 - .045 range for this graph, but still one present nonetheless.

The majority of the statistical tests mentioned were ANOVA and regression while t-tests, MANOVA and chi-square were the clear minority. For a full breakdown of the statistical tests mentioned in the literature, refer to Figure 4.

2015

All 99 articles from the *Journal of Abnormal Psychology* had their *p*-values extracted and graphed, see Figure 5 for the resulting distribution. Again, there were a greater number of *p*-values in the 0 - .01 range. There were bumps at the .03 - .04 range as well as a drop off at .05.

The proportions of statistical tests changed quite a bit from 2005 to 2015; regression became the most used while ANOVA and t-test had similar proportions as the second and third most widely used tests. There was also usage of non-parametric tests such as the chi-square and the Mann-Whitney U test.

Journal of Experimental Psychology: General

1995

All 22 articles from JEP:G had their *p*-values extracted and graphed. As expected there was a higher proportion of *p*-values in the .00 - .01 range. There was a notable number of *p*-values in the .04 - .045 range, accounting for 10.7 percent of all *p*-values reported.

More than half the mentions regarding statistical tests were regression in JEP:G 1995 with ANOVA as a clear second. There were slight mentions of the chi-square test as well as the t-test. Unique to the year 1995 were mentions of Fisher's exact test.

2005

All 41 articles from the *Journal of Experimental Psychology: General* had their *p*-values extracted and graphed, see Figure 9 for the resulting distribution. Again, there were the greatest number of *p*-values in the 0 - .01 range. There was a dramatic decrease in the number of *p*-values in the .04 - .045 range from 1995 to 2005, but still a pronounced bump at the .045 - .05 range.

The vast majority of tests conducted in JEP:G 2005 were t-tests with ANOVA being a clear second. There was minimal use of regression and non-parametric tests like the chi-square. See Figure 10 for a complete breakdown.

2015

All 50 articles from the *Journal of Experimental Psychology: General* had their *p*-values extracted and graphed, see Figure 11 for the resulting distribution. Again, the greatest number of *p*-values were in the 0 - .01 range. The number of *p*-values seemed to be decreasing as the *p*-value increased. However, there was a large drop off between .05 and .055, which could potentially be the result of publication bias.

As for the mentions of statistical tests, ANOVA was the most widely mentioned test in 2015 for JEP:G, with regression being a clear second. Interestingly, the Mann-Whitney U test was mentioned a sizeable amount, being the third most mentioned test for the literature in 2015.

Journal of Personality and Social Psychology

1995

All 68 articles from the *Journal of Personality and Social Psychology* had their *p*-values extracted and graphed, see Figure 13 for the resulting distribution. Again, the greatest number of *p*-values were in the 0 - .01 range. There was a large number of *p*-values in the .03 - .035 range, also there was a bump at the .04 -.055 range.

As for the mentions of statistical tests, regression was the most mentioned test in the journal for 1995. This was followed by ANOVA and chi-square. Notably, MANOVA accounted for a noteable number of mentions in the literature.

2005

All 133 articles from the *Journal of Personality and Social Psychology* had their *p*-values extracted and graphed (see Figure 15 for the resulting distribution). Again, the greatest number of *p*-values were in the 0 - .01 range. The distribution looked to be uniform from .025 - .05 with a slight drop off at .055.

As for the mentions of statistical tests, regression was the most widely mentioned test in 2005 for JPSP, with ANOVA being a close second. Interestingly, ANCOVA was mentioned, a new addition that was not present in 1995.

2015

All 119 articles from the *Journal of Personality and Social Psychology* had their *p*-values extracted and graphed (see Figure 17 for the resulting distribution). Again, the greatest number of *p*-values were in the 0 - .01 range. The distribution looked similar to 2005's, with a more pronounced drop off at .05, which again could possibly be the result of publication bias.

As for the mentions of statistical tests, regression was the most widely mentioned test in 2015 for JPSP, with ANOVA being a clear second and the t test being a close third.

Mentions of Covariates, Confidence Intervals, Effect Size and Replication

As mentioned above, each journal was assessed for the use of covariates and replications. Regarding the mentions of covariates, it was clear that in the three years selected for analysis, JEP:G employed very little use of covariates. JAbP demonstrated an upward trend in the use of covariates with the number of mentions increasing as the years progressed. For JPSP however,

the peak of mentions were in 2005 with practically five times the number of mentions compared to 1995. See Figure 19 for a graphical breakdown of the number of mentions in each journal and year. There are plenty of claims in the literature that stress a move toward reporting confidence intervals and effect sizes (Cumming & Finch, 2005), so the amount of articles reported both confidence intervals and effect sizes in each journal across the selected years were counted. See Figures 20 and 21 for a breakdown.

Differing Cultures

Part of the main hypothesis was each psychological subdiscipline has its own unique culture, that in turn affects the practices of the subdiscipline, which would then influence and could be used as an explanation regarding each subdiscipline's plot of *p*-values. It is clear that the three subdisciplines selected for analysis in this current study (Personality and Social, Experimental and Abnormal Psychology) do indeed display differences between one another. The differences that were uncovered in the present study mainly had to do with the types of statistical analysis conducted in each subdiscipline as well as markers of ideal practices (i.e., use of confidence intervals and effect sizes). This is an area of importance, mainly because many argue that in some situations research topics are often chosen to fit the statistical methodologies and practices of the said subculture (Flis & Van Eck, 2018).

Flis and Van Eck (2018) took ideas from Lee Cronbach (1957, 1975) and applied them to modern day psychology. Cronbach (1957, 1975) described the state that he viewed psychology to be in; split between two distinct groups: correlational psychologists and experimental psychologists. Flis and Van Eck then took these groups and attempted to see if Cronbach's descriptions of the psychological field held steady over a 49 year span. Surprisingly, it seemed to. Though there was evidence for this divide between correlational and experimental

psychology, fields like neuro, experimental and animal psychology were much more coherent while other fields like social, personality and clinical psychology had a very loose alliance.

What kind of cultures then, are each group in the present study taking part in? For starters, experimental psychology seems to be above the rest in terms of statistical practices. JEP:G demonstrated having the highest proportion of articles employing the use of confidence intervals as well as effect sizes; over half of the articles in the journal employed their use. What is interesting, however, is attempting to fit these groups into the framework that Cronbach (1957, 1975) proposed. An argument can be made that Experimental Psychology of course fits in the experimental category; most of their research is the manipulation of variables and this is further evidenced through the statistical tests that have been most used in the literature (*t*-tests and ANOVAs). The other side of the argument though can be that Personality and Social Psychology as well as Abnormal Psychology can both be grouped into the correlational category. This is evidenced through the lack of experimental manipulation in the literature, the use of mainly regression and the difficult to quantify subject matter of the groups.

Similar to Flis and Van Eck (2018) the abstracts and titles for every article published in 2015 for the three journals were taken and analysed for the type of language they used. While JPSP and JAbP used language such as “perceived”, “emotion”, “depression”, “anxiety”, JEP:G used language such as “control”, “effects”, “differences”. The subject matter in JPSP and JAbP such as emotion and anxiety are incredibly difficult to quantify, manipulate and control for, which is why it is argued here that the two groups are more likely to be categorized in the correlation psychology grouping. Topics such as these require further research in hopes to gain a higher understanding of the field of psychology.

Discussion

As previously stated, there was already considerable speculation regarding the state of current and recent psychological research. There can be multiple points of analysis here: Has each individual journal improved its practices over time? Do some journals have fewer instances of questionable *p*-values in the .04 - .05 range? Using these journals as a sample, does it seem like psychology as a whole is trending toward favourable practices? In short, yes, it seems as though each journal has improved their individual practices over time, whether it be the more representative graph of *p*-values that Masicampo and Lalonde (2012) described, or simply the use of a wider range of statistical tests and employment of confidence intervals and effect sizes.

Starting with the graphed *p*-values for JAbP, there was a modest trend toward fewer values in the .04 - .05 range. In 1995 and 2005 there were clear abnormalities in the graphs; the proportion of values in the .03 - .035 range was remarkably high, and comparable to the proportion of *p*-values in the .01 range. In 2015, however, this bump was non-existent compared to past years. However, the area of interest is the .04 - .05 range, this is the area that suspected users of QRPs would strive to obtain, making an abnormal bump in the distribution or a sharp drop off. For JAbP, from 1995 to 2005 there was a substantial difference in the proportion of p-values in the .04 - .05; in 1995, 5 percent of all p-values were in this range. However, by 2005 this value had decreased to only 3 percent. In 2015 there was only a marginal difference, 2.8 percent of all p-values fell into the .04 - .05 range.

In terms of the distributions for each selected year in JAbP, there was a clear improvement from 1995 to 2005, however in 2015 there were comparable amounts of *p*-values in the .04 -.05 range but there was also a bump at the .03 - .04 range, this does not follow the smooth distribution of *p* that Masicampo and Lalonde describe to be ideal. This has the

possibility of being interpreted as a problem based off of past literature (Masicampo et al., 2014; Simonsohn et al., 2014) whether it be the use of QRPs, poor statistical practices or another issue that the research world is not yet aware of.

Regression was the most widely mentioned test in the journal across the three years, and the use of covariates rose as the years progressed. Simonsohn et al., (2011) made it clear that using researcher degrees of freedom while creating a regression model can have a very significant effect on the false positive rate. This does not imply that using regression and controlling for covariates mean that QRPs were employed. However, based on these findings, further, more in-depth analysis must be conducted to see how the tests were used.

For JEP:G it was somewhat of a different story. JEP:G had the least desirable distribution for 1995; there was a remarkably high proportion of *p*-values in the .04 - .05 range. However, post 1995, the proportions of *p*-values were comparable with that of JAbP (see Figure 22). Through the three years selected, JEP:G saw the greatest improvement; there was a dramatic decrease in the number of *p*-values in the .04 - .05 range and by 2015, the distribution of *p*-values was satisfactory. JEP:G trended from regression being the most mentioned test in the literature in 1995, to ANOVA in 2015, and its use of covariates was almost non-existent. An argument can be made that when using ANOVA, the decision of what constitutes a "family" of tests and when the family-wise error rate needs to be corrected is much more clear then when using regression.

Finally, for JPSP, the field that has seen the most scrutiny regarding this "crisis", showed somewhat of an evolution through the three selected years. Similar to JAbP and JEP:G, the year 1995 displayed an alarming number of *p*-values in the .04 - .05 range along with a large number in the .03 - .035 range. However, there was improvement nonetheless, by 2015 the distribution

still did not represent what Masicampo and Lalonde describe as the optimal distribution of *p*, but was still more desirable compared to the preceding years. In terms of the proportion of *p*-values, there were only slight decreases in the proportion and from 2005 onward, had the highest proportion of *p*-values in the .04 - .05 range compared to the other two journals.

There were substantial decreases in the number of *p*-values in the .04 - .05 range for all three journals analysed in this study; why was this? A possible explanation could indeed be that past researchers who traditionally followed Neyman-Pearson theory religiously, could have begun to report exact *p*-values compared to only stating "$p < .05$". Also, new waves of researchers could now be adopting the practice of reporting *p*-values in more specific groupings (i.e. $p < .001$, $p < .01$, $p < .05$). To know this for certain, as stated above, the exact *p*-values must be calculated through the information reported in each article. Another explanation may well be that researchers are trending toward increasingly favourable statistical and research practices.

It was hypothesized that where practices seemed favourable (i.e. there was not an abnormal bump of *p*-values in the .04 - .05 range) that better statistical practices would also be present. Although for JAbP and JPSP, the breakdown of mentioned statistical tests remained relatively the same in that the same groups of tests were constantly mentioned in the literature. For JEP:G, the journal that saw the greatest decrease in *p*-values in the .04 - .05 range, also saw the greatest change in statistical tests mentioned in the literature. 2015 saw a rise to a large proportion of mentions in the literature regarding the Mann-Whitney *U* test. The *U* test is a non-parametric test that does not follow the same assumptions as the *t* test; namely, that the data does not have to be normally distributed to be able to use the *U* test. This can imply then, that researchers in experimental psychology are becoming increasingly conscious regarding

assumptions of statistical tests and are turning to a wider range of tests, compared to the past.

Past literature has emphasised a move toward reporting confidence intervals and effect sizes, stressing the need to exemplify superior statistical practices. What is encouraging is that all three journal exhibited an increase in the number of articles that reported both confidence intervals (CI) and effect sizes (ES). This trend can indicate a move toward superior statistical practices. However, what was surprising was the proportion of articles that reported both CIs and ESs in JEP:G compared to the other two journals. By 2015, a shocking 54% of articles published in JEP:G reported CIs and 42% reported ESs. This is compared to 32% and 38% for JAbP and 41% and 39% for JPSP respectively. This trend follows the decrease in *p*-values in the .04 - .05 range, with the journal with the least amount of *p*-values in the .04 - .05 range also having the highest proportions of articles employing the use of CIs and ESs. This can be used to support the claim that superior statistical practices, in fact may play a part in the reduction of *p*-values in the .04 - .05 range, which can imply a decrease in QRPs.

These claims are in-line with current research (Nuijten, Hartgerink, van Assen, Epskamp & Wicherts, 2016) that has explored the prevalence of QRPs from the past to the present through inferential techniques. Nuijten et al., (2016) examined the prevalence of reporting errors from 1985 - 2013 regarding *p*-values in eight major psychology journals, of which JPSP and JEP:G were included. What they found was that errors that would most likely change the statistical outcome of a test were actually declining as the years progressed. Similarly, in this current study, the amounts of *p*-values in the .04 - .05 range have also declined implying the same conclusion; researchers are employing better statistical practices. Nuijten et al., (2016) also compared JPSP

and JEP:G with one another and found too, that in terms of errors, JEP:G had less compared to JPSP.

Differences were present between the three journals studied; by 2015, JEP:G and JAbP exhibited the most desireable distribution of graphed *p*-values. In terms of the cultural differences between experimental psychology, abnormal psychology and social and personality psychology, experimental psychology displayed superior statistical practices. By 2015 the number of statistical tests mentioned in the literature were more diverse compared to the other journals and proportionally, experimental psychology had the greatest use of CIs and ESs. JAbP also showed use of non-parametric tests like the Mann-Whitney *U* test, demonstrating that researchers were looking beyond the usual statistical toolbox and using methods that were more specific to their needs. Theses findings are all in line with the hypothesis that superior statistical practices will yield distributions closer to the ideal distribution of *p*.

Limitations and future research

Using a program like Statcheck comes with its limitations. The first being that Statcheck is extremely limited in what it is able to read. For Statcheck to work properly, all statistics must be reported in APA format with no deviations whatsoever. Nuijten (2016) provides examples of how researchers slightly deviate from APA formatting and that these small deviations render entire statistics unreadable for Statcheck. Some of these examples are: Failing to put a space between the test stat and df (i.e. t(37) = …), Inserting other statistics between a statistic (i.e. F (3,144) = 5.21, **Mse = 2.1**,...), Including the sample size with the statistic (i.e. [x^2 (1, **N = 226**) = 5.3…) and Combining two results (i.e. t (45) and t (31) were 3.12 and 5.21 respectively…). Statcheck can also only read statistics that are reported within the text of an article, meaning that any statistics reported in chart form are unreadable for the program.

Also, often times certain symbols such as “<”, “>” and “=” will be saved as image files within the text and not as ASCII characters. In this case the articles then have to be fed through an optical character recognition program (OCR) to either convert them to .pdf or .html documents where Statcheck can then read and interpret them. Doing so adds another level where errors can occur, such as the OCR not reading the proper text.

Conclusion

This paper should be used as a starting point in answering the question “are there differences between psychology subfields regarding the use of QRPs and statistical practices?” It was hypothesized that the difference between journals and the corresponding distribution of *p*-values could be partially explained by each journal’s individual “cultures” (i.e. statistical practices). This was somewhat true indicated through the results; JEP:G showed the most sizeable improvement by 2015 and also exhibited the most diverse group of statistical tests mentioned in its literature while employing the most CIs and ESs. Also asked was a question regarding the practices of each subfield through time. It is clear that there are trends toward favourable statistical practices. This is in line with current research (see Nuijten, et al., 2016) which also saw a decreasing trend in poor statistical practices.

Moving forward there are plenty of routes that can be taken to further ameliorate the issue of QRPs in psychological research. Again are calls to turn to CIs and ESs; this should be encouraged to be included in every research article that contain statistical tests. The APA’s task force for statistical inference has been a noted supporter of the use of confidence intervals and have said that confidence intervals are the best strategy for reporting data (Cumming & Finch, 2005). Confidence intervals not only display significance, they also display some information regarding the magnitude of the effect as well as the precision of the measurement. However,

more “radical” are the calls to turn to a completely different statistical framework all together; Bayesian inference. It has been argued that by turning to Bayesian inference, as opposed to NHST, many of the problems that face the field in the present day would be solved (Wagenmakers, et al., 2018). There are two main arguments for the use of Bayesian inference. Firstly, it answers the questions that researchers are actually asking; Bayesian inference tells one the probability of a hypothesis given the data, whereas NHST tells one the probability of the data given the hypothesis, two extremely different answers, to two extremely different questions. Second, is that Bayesian inference has the ability to incorporate past knowledge. Computing Bayesian inference requires one to use past knowledge to form a prior (what they believe to be the probability that the hypothesis in question is true). These priors can be constantly updated and researchers can collaborate with each other to create a more precise estimation. Bayesian methods can be adapted to suit all types of hypothesis testing that a researcher desires, but to implement this in the literature and create a reform is a different story. The problem as to be tackled from the journal editors standpoint, if “different” types of articles will not be published, than they will not be written (Moustafa, 2015).

Appendix A: Figures

Figure 1

Example p-curves for different effect sizes with and without true effects. Reprinted from Simonsohn, Nelson, & Simmons, 2014.

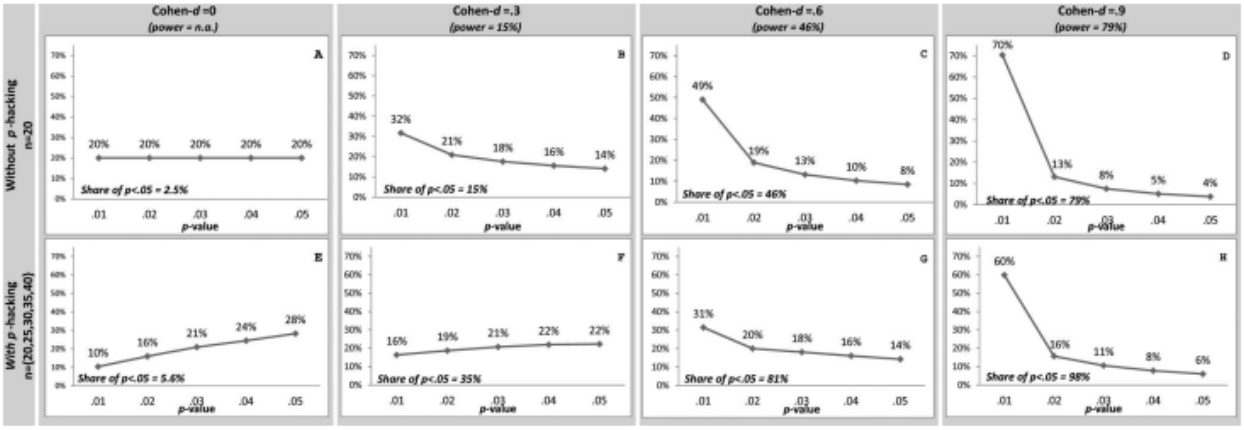


Figure 2

Distribution of p-values from JAbP, 1995

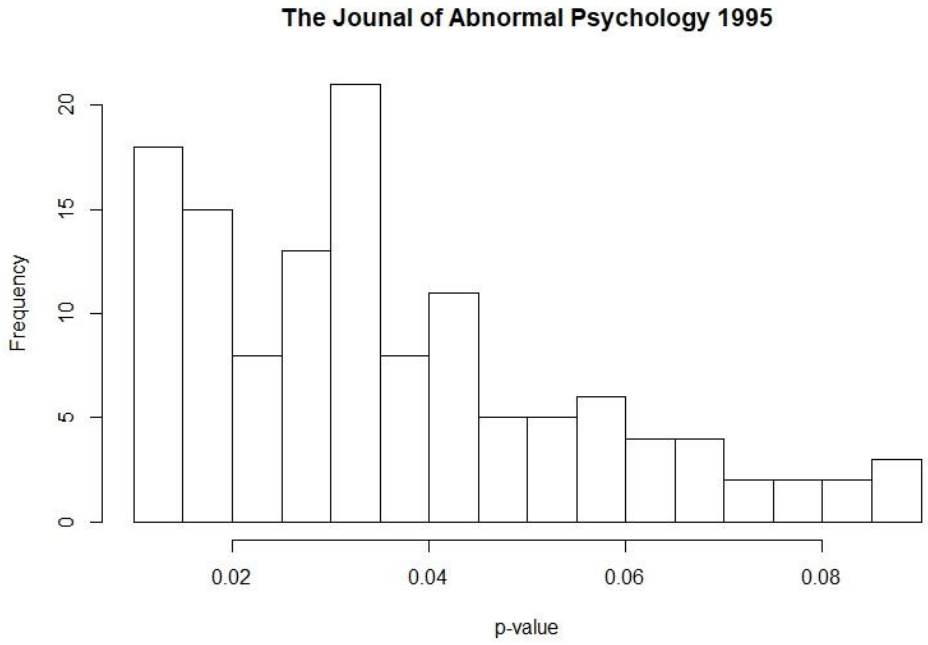


Figure 3

Proportion of statistical tests used in JAbP 1995

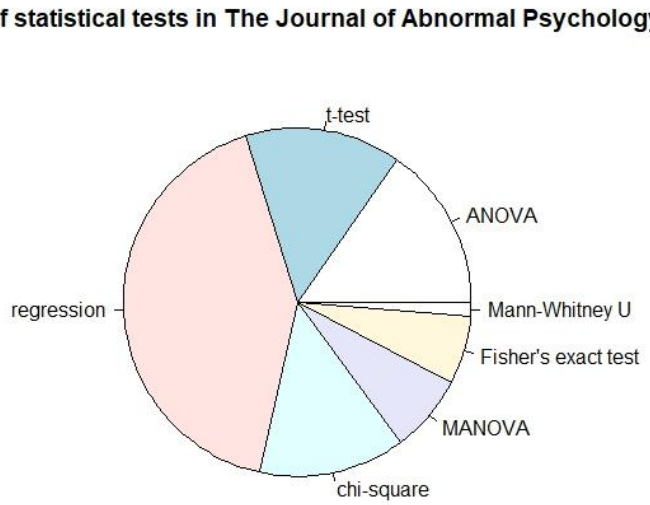


Figure 4

Distribution of p-values from JAbP, 2005

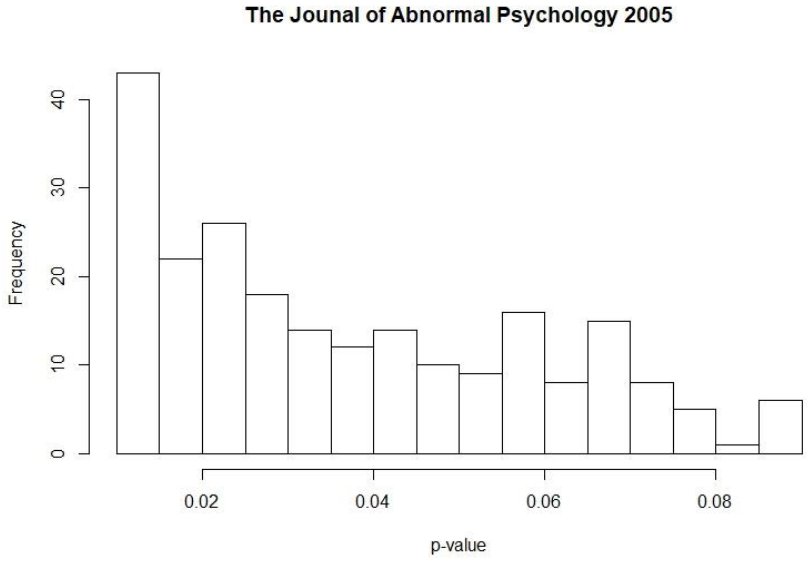


Figure 5

Proportion of statistical tests used in JAbP 2005

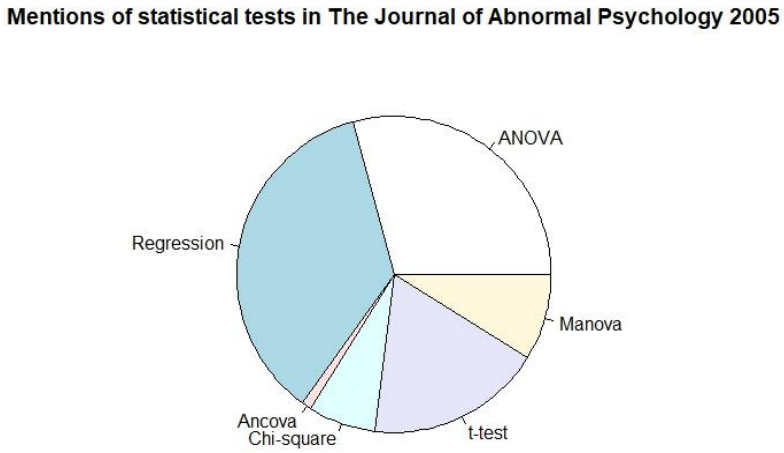


Figure 6

Distribution of p-values from JAbP, 2015

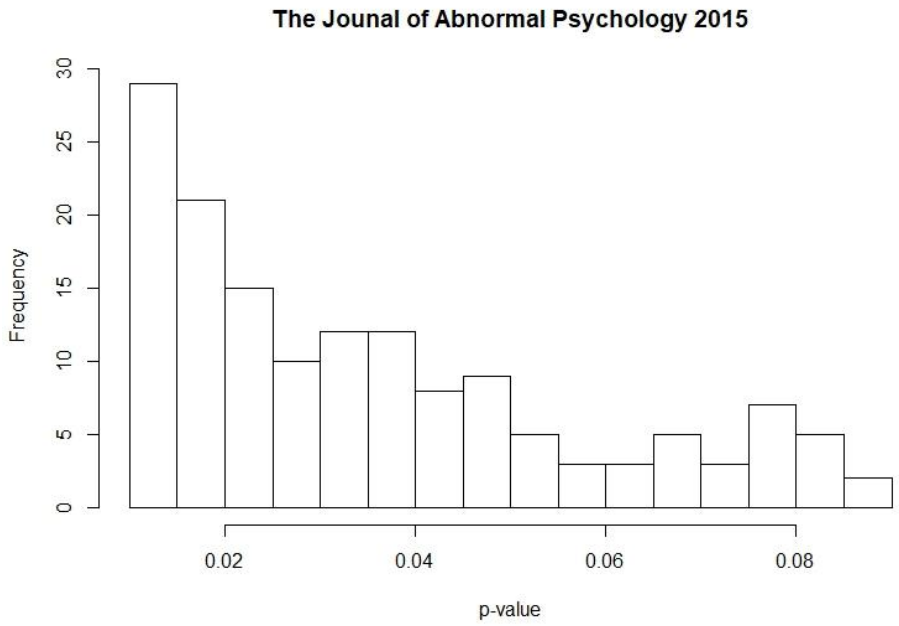


Figure 7

Mentions of statistical tests in JAbP 2015

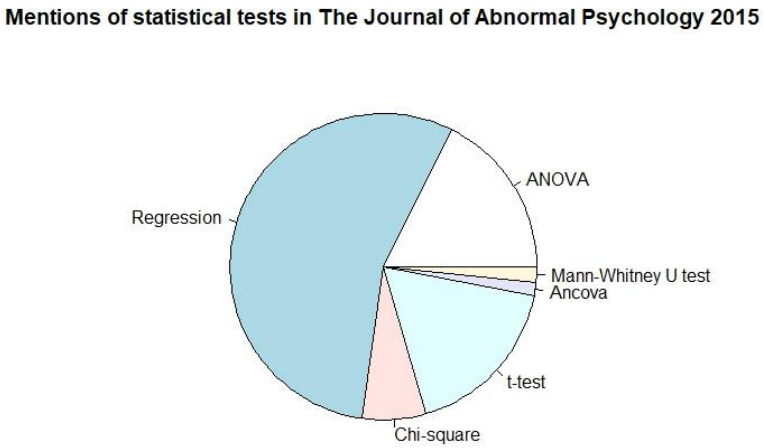


Figure 8

Distribution of p-values from JEP:G, 1995

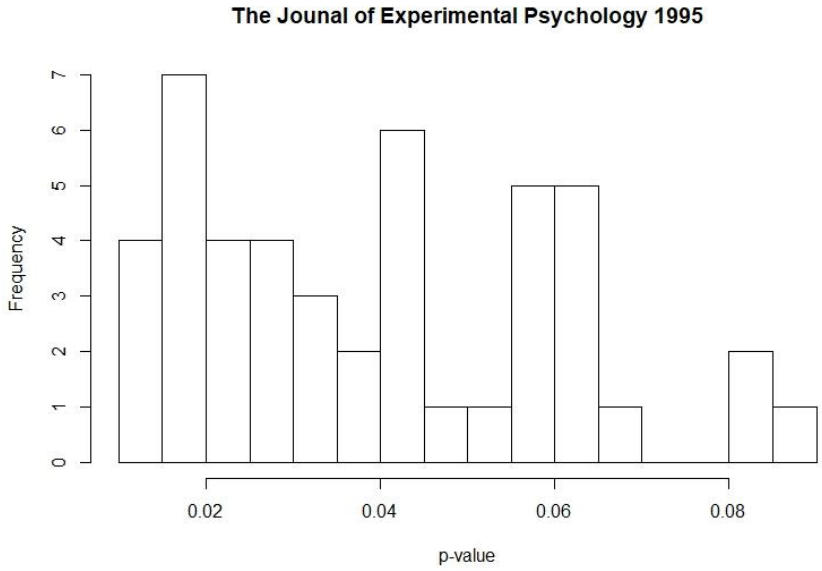


Figure 9

Mentions of statistical tests in JEP:G 1995

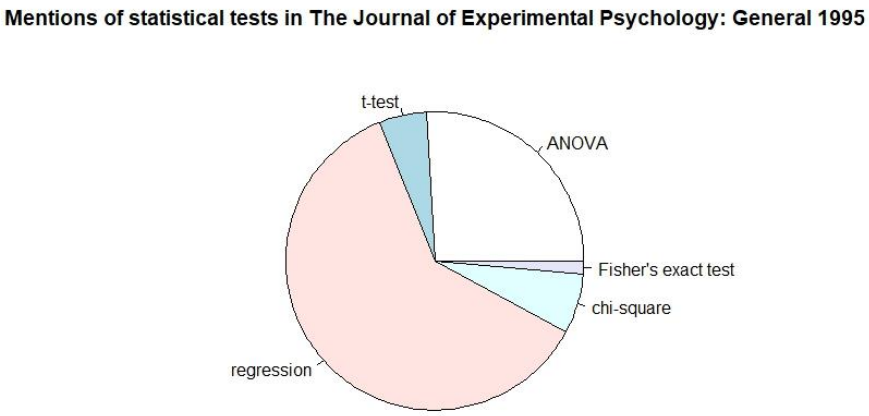


Figure 10

Distribution of p-values from JEP:G 2005

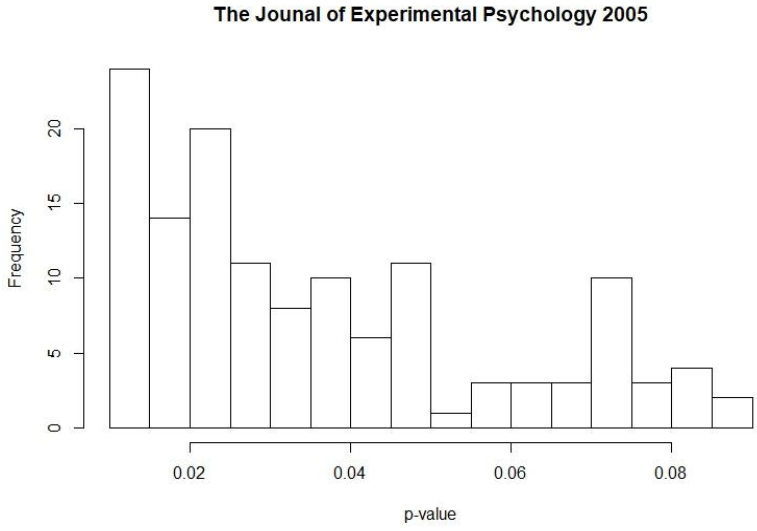


Figure 11

Proportion of statistical tests used in JEP:G 2005

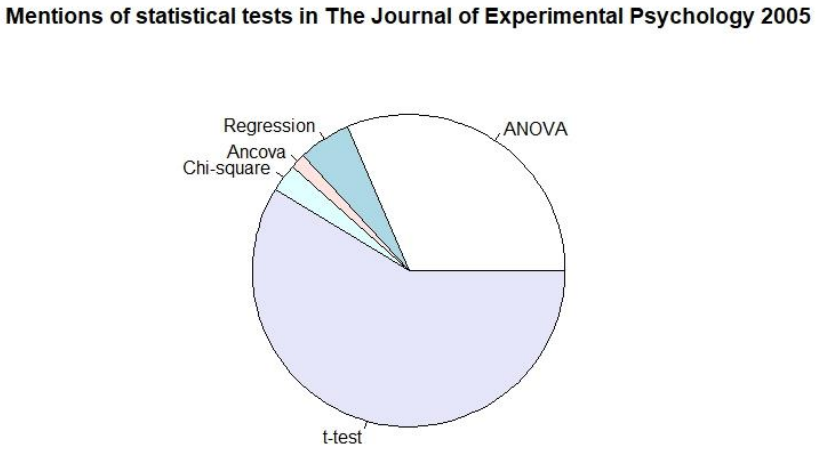


Figure 12

Distribution of p-values from JEP:G 2015

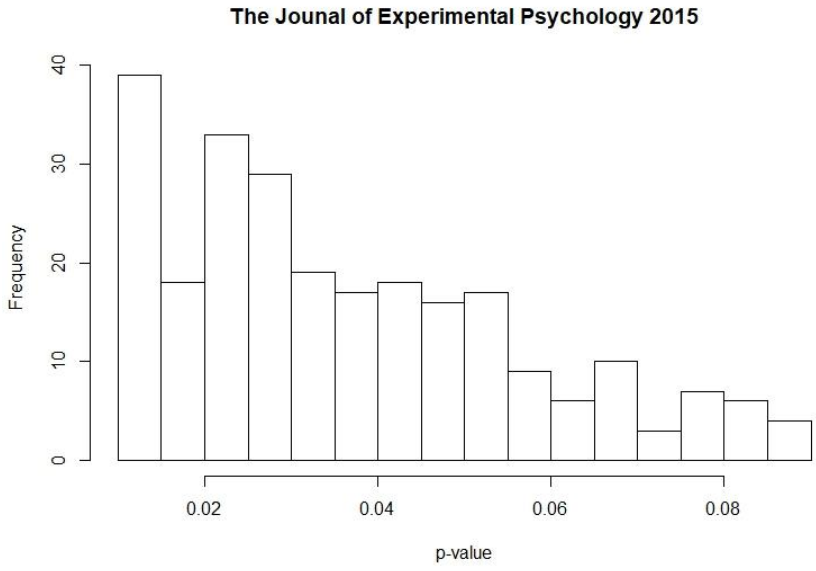


Figure 13

Mentions of statistical tests in JEP:G 2015

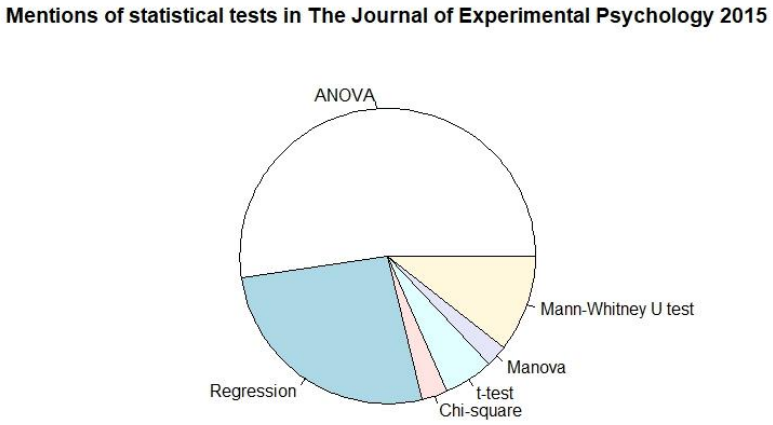


Figure 14

Distribution of p-values from JPSP 1995

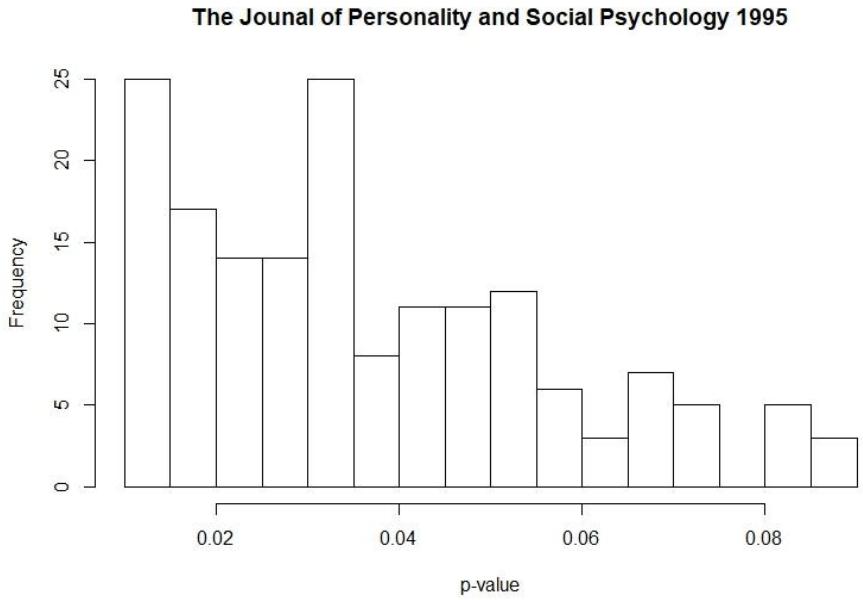


Figure 15

Mentions of statistical tests in JPSP 1995

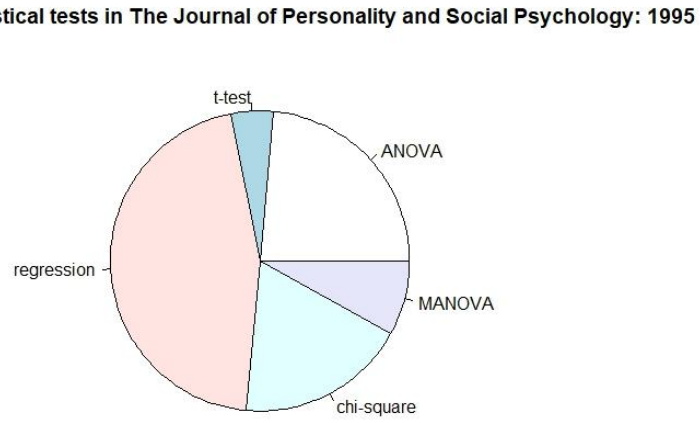


Figure 16

Distribution of p-values from JPSP 2005

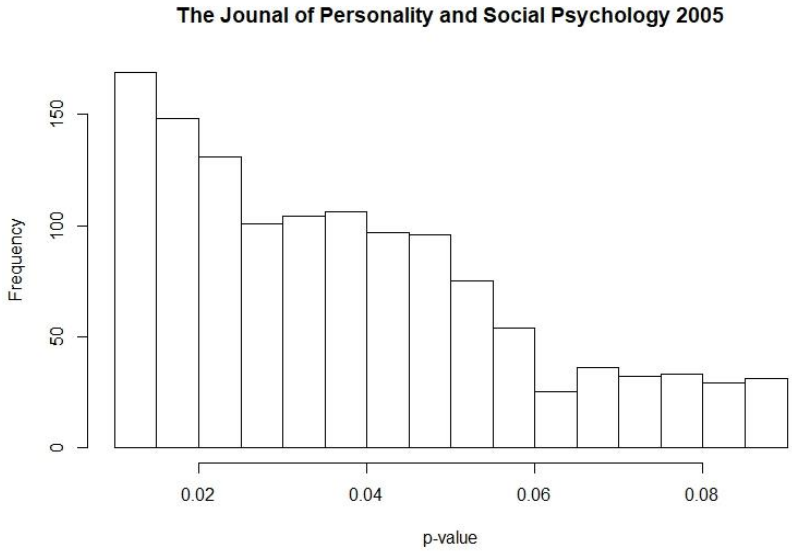


Figure 17

Mentions of statistical tests from JPSP 2005

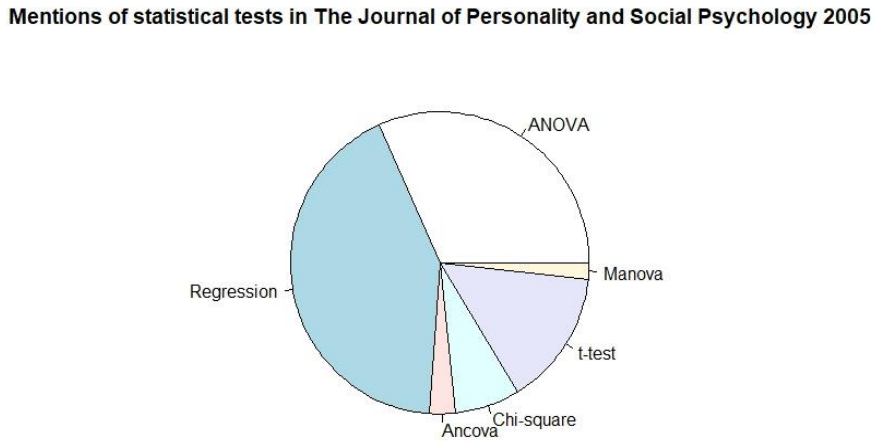


Figure 18

Distribution of p-values for JPSP 2015

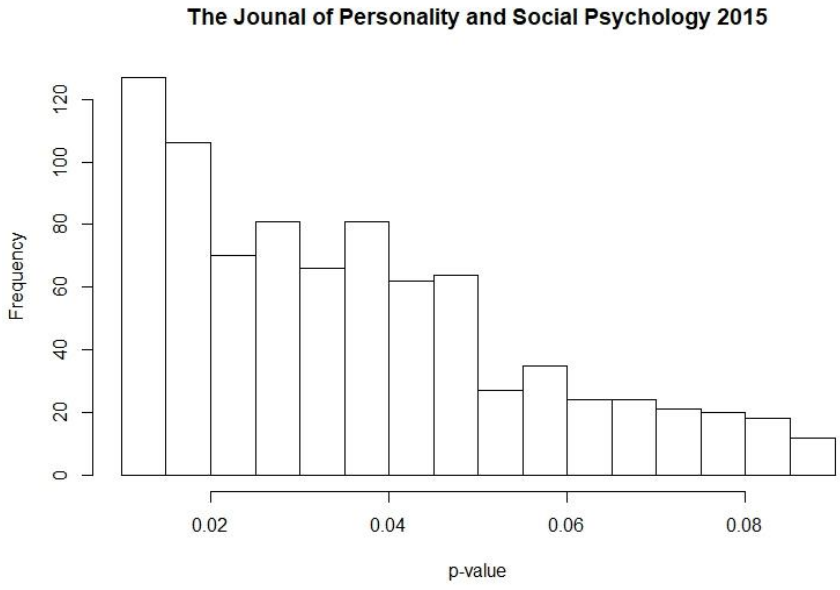


Figure 19

Mentions of statistical tests in JPSP 2015

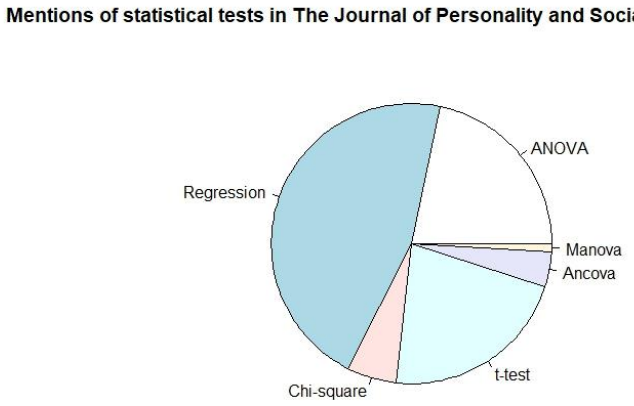


Figure 20

Graphical representation of the use of the word covariate between the three journals

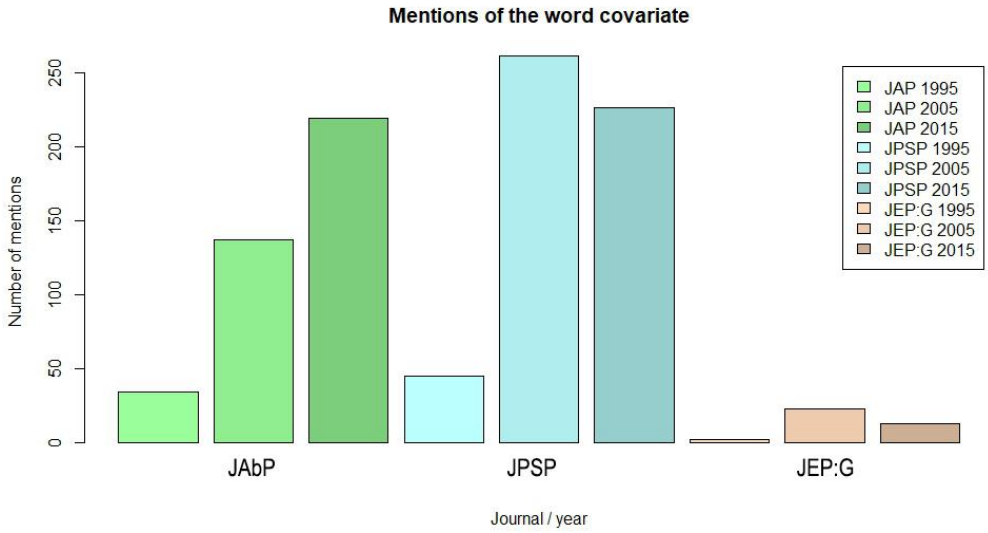


Figure 21

A graphical representation of the number of articles from each journal that reported effect sizes

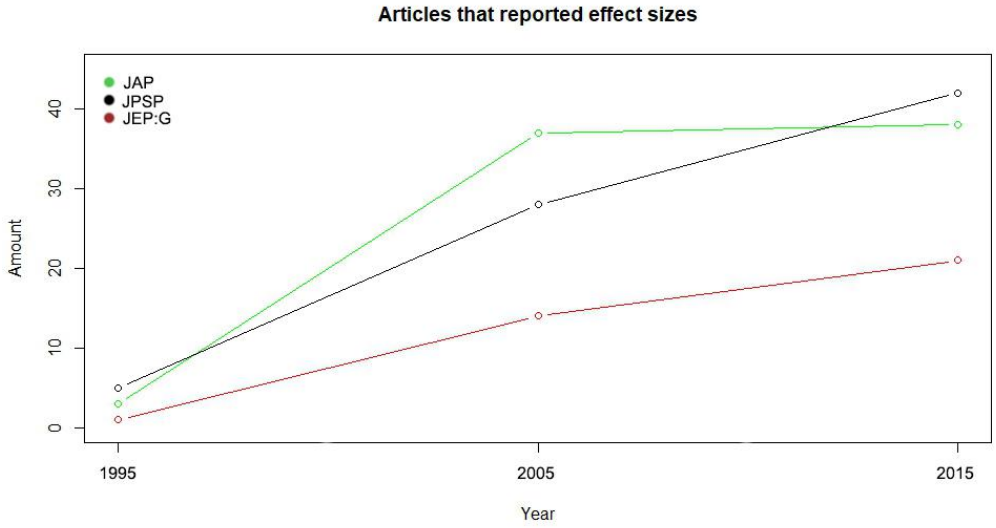
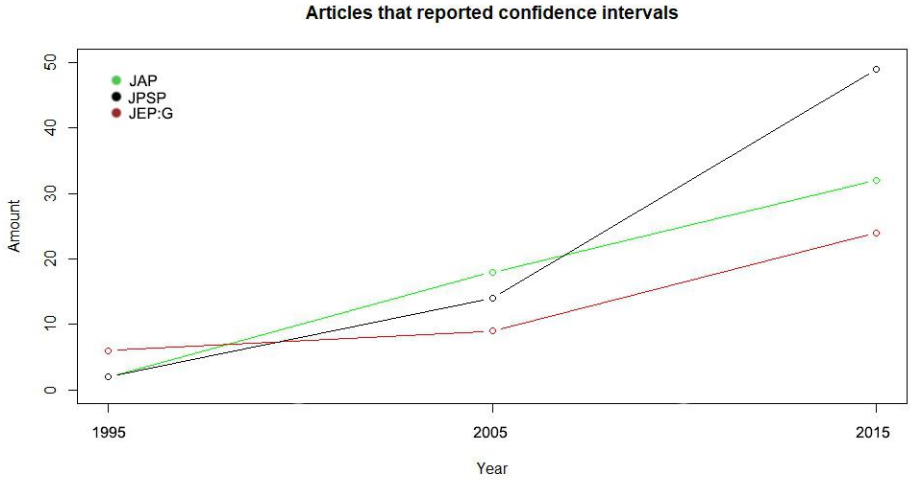


Figure 22

A graphical representation of the amounts of articles from each journal that reported confidence intervals



Appendix B: Tables

Table 1

Breakdown of number of journals per year.

Journal / Year	1995	2005	2015	Total
JAbP	76	76	99	251
JEP:G	22	41	50	113
JPSP	68	133	119	320

Table 2

Number of p-values from each journal and year

Journal	**1995**	**2005**	**2015**	**Total**
JAbP	462	781	599	1,842
JEP:G	178	517	1,143	1,838
JPSP	529	3,767	3,244	7,540
Total	**1,169**	**5,065**	**4,986**	**11,220**

References

Altman, D.G., Gore, S.M., Gardener, M.J., & Pocock, S.J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal,* 286, 1489-1893.

Antonakis, J. (2017). On doing better science: From thrill of discovery to policy implications. *The Leadership Quarterly*, 28(1), 5-21.

Aschwanden, C. (2015, Aug 19). Science isn't broken: It's just a hell of a lot harder than we give it credit for. *FiveThirtyEight* (website). https://fivethirtyeight.com/features/science-isnt-broken/

Bartlett, T. (2018, Sep 20). As Cornell finds him guilty of academic misconduct, food researcher will retire. Retrieved from https://www.chronicle.com/article/As-Cornell-Finds-Him-Guilty-of/244584

Begley, C.G., & Ellis, L.M. (2012). Raise standards for preclinical cancer research. *Nature*, 483, 531-533.

Berger, V.W., (2001). The *p*-value interval as an inferential tool. *Journal of the Royal Statistical Society: Series D*, 50(1), 79-85.

Bown, N.J.L., & Heathers, J.A.J., (2016). The GRIM test: A simple technique detects numerous anomalies in the reporting of results in psychology. *Social Psychological and Personality Science*, 8(4), 363-369.

Claerbout, J.F., & Karrenbach, M. (1992). Electronic documents give reproducible research a new meaning. *SEG Technical Program Expanded Abstracts*, 601-604.

Cohen, J., (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology*, 65, 145-153.

Cohen, J., (1990). Things I have learned (so far). *American Psychologist*, 45, 1304-1312.

Cohen, J., (1992). A power primer. *Psychological Bulletin*, 70(6), 426-443.

Cohen, J., (1994). The earth is round ($p < .05$). *American Psychologist,* 49(12), 997-1003.

Cowells, M., & Davis, C. (1982). On the origins of the .05 level of statistical significance. *American Psychologist,* 37(5), 553-558.

Cronbach, L. J. (1957). The two disciplines of scientific psychology. American Psychologist, 12, 671–684. http://dx.doi.org/10.1037/h0043943

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30, 116–127. http://dx.doi.org/10.1037/h0076829

Cumming, G. (2008). Replication and *p* intervals: *p*-values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science,* 4(3), 286-300.

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

Cumming, G., & Finch, S. (2005). Inference by eye: Confidence intervals and how to read pictures of data. *American Psychologist,* 60(2), 170-180.

Dominus, S. (2017, Oct 18). When the revolution came for Amy Cuddy. *The New York Times,* Retrieved from https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-a my-cuddy.html

Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PLOS,* 4(5).

Fanelli, D. (2012). Bad news: Science is getting too positive. *AQMeNtion,* 9.

Flis, I., & van Eck, N.J. (2017). Framing Psychology as a Discipline (1950-1999): A Large-Scale Term Co-Occurrence Analysis of Scientific Literature in Psychology. *History of psychology,* 21(4), 334-362.

Gigerenzer, G. (2004). Mindless statistics. *Journal of Socio-Economics,* 33(5), 587-606.

Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advanced Methods and Practices in Psychological Science,* 1(2), 198-218.

Guilford, J. P. (1942). Fundamental statistics in psychology and education. New York, NY, US: McGraw-Hill.

Heathers, J. (2017, Mar 3). Introducing SPRITE: (and the case of the carthorse child). *Hackernoon,* (blog). https://hackernoon.com/introducing-sprite-and-the-case-of-the-carthorse-ch ild-58683c2bfeb

Hossenfelder, S. (2017, Dec 12). Research perversions are spreading. You will not like the proposed solution. *Backreaction* (blog) http://backreaction.blogspot.com/2017/12/researc h-perversions-are-spreadingyou.html

Ioannidis, J. (2005). Why most published results are false. *PLoS Medicine, 2*.

Ioannidis, J. (2008). Why most discovered true associations are inflated. *Epidemiology*, 20(4), 629.

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

Keren, G. & Lewis, C. (2014). A Handbook for Data Analysis in the Behavioral Sciences: Volume 1: Methodological Issues Volume 2: Statistical Issues. Psychology Press.

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

Kline, R. B. (2004). Beyond significance testing: Reforming data analysis methods in behavioural research. Washington, DC: American Psychological Association.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A “many labs” replication project. *Social Psychology, 45*(3), 142-152.

Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams, R. B., Jr., Alper, S., … Nosek, B. A. (2018, November 19). Many Labs 2: Investigating Variation in Replicability Across Sample and Setting.

Lakens, D. (2013, Nov 26). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for *t*-tests and ANOVAs. *Frontiers of Psychology*.

Madden, C. S., Easley, R. W., & Dunn, M. G. (1995). How journal editors view replication research. *Journal of Advertising, 24*(4), 77-87.

Marcus, A,. & Oranksy, I. (2018, Dec 26). More science than you think is retracted. Even more should be. *Washington Post,* Retrieved from https://www.washingtonpost.com/opinions/more-science-than-you-think-is-retracted-even-more-should-be/2018/12/26/dc14fa98-0950-11e9-a3f0-71c95106d96a_story.html?noredirect=on

Masicampo, E. J. & Lalande, D. R. (2012). Peculiar prevalence of p-values just below .05. *Quarterly Journal of Experimental Psychology*, 65, 2271-2279.

Moustafa, K. (2015). The disaster of impact factor. *Science and Engineering Ethics, 21*, 139-142.

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*, 615-631.

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

Reinhart, A. (2015). Statistics done wrong: The woefully complete guide. San Francisco, CA, No Starch Press.

Rieber, R. W., & Robinson, D. K. (Eds.). (2001). *PATH in psychology. Wilhelm Wundt in history: The making of a scientific psychology.* New York, NY, US: Kluwer Academic/Plenum Publishers.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin, 86,* 638-641.

Rossi, J. S., (1990). Statistical power of psychological research: What have we gained in 20 years?. *Journal of Consulting and Clinical Psychology*, 58(5), 646-656.

Sedlmeier, P., & Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? Psychological Bulletin, 105(2), 309-316.

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,* 1359-1366.

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

Stanley, T. D., Carter, E. C., & Doucouliagos, H. (2018). What meta-analyses reveal about the replicability of psychological research. *Psychological Bulletin, 144*(12), 1325-1346.

Tafreshi, D. Slaney, K. L. & Neufeld, S. D. (2016). Quantification in psychology: Critical analysis of an unreflective practice. *Journal of Theoretical and Philosophical Psychology*, 36(4), 233-249.

Wagenmakers et al. (2017). Bayeisan Inference for Psychology 1: Theoretical advantages and practical ramifications. *Psychonomics Bulletin & Review, 24*, 35-57.