Why circular reasoning leads to weak theories – and how open science can help*

Peter M. Dahlgren

Department of Journalism, Media and Communication (JMG), University of Gothenburg

December 22, 2017

Abstract

Circular reasoning occurs when a conclusion is assumed rather than demonstrated. This circularity in empirical research can lead to communication theories that are built on top of findings that are believed to be true, but are in fact false. The main culprit is a failure to distinguish between exploratory and confirmatory research. If they are not separated, the risk of building weak communication theories that are based on circular reasoning increases. Most importantly, negative (falsified) findings are the most certain findings of science, since they are based on deduction. They should therefore be treated as *more* valuable contributions to the scientific knowledge than positive findings, not less. To amend these problems, open science is suggested with increased transparency that separate exploratory and confirmatory research by means of pre-registration of hypotheses before data collection begins. This will help put more confidence in political communication in particular and communication science in general.

Keywords: questionable research practices (QRP), circular reasoning, publication bias, confirmation bias, hindsight bias, null hypothesis significance testing (NHST), philosophy of science, open science

*<b>Work in progress</b>. Latest version available at: http://peterdahlgren.com/papers/circular.pdf. Suggestions for improvements can be sent to peter.dahlgren@jmg.gu.se.

1 The problem

A famous food researcher posted a now deleted blog post detailing how his lab was able salvage a failed experiment by sifting through data until they found significant findings that could be published. This behavior made other researchers suspicious, who subsequently checked the lab's published articles. They found over 150 inconsistencies in the papers (van der Zee, Anaya, & Brown, 2017), and several papers have up to now been retracted or corrected by the journals. The food researcher is not alone in these research practices, and the problem is more pervasive than only positive findings being reported and published. But what did the researcher actually do? And what can we learn from these errors?

The problem can be boiled down to a failure to separate hypothesis-generating from hypothesis-testing research (i.e., exploratory from confirmatory research). As the names imply, hypothesis-generating research is about creating something new whereas hypothesis-testing research is about finding out if that new thing holds up for scrutiny. Since these two modes of research are also based on different logical inferences, conflating them are not only ethically questionable, but can also lead to circular reasoning that stifles scientific discovery—ironically as the point of these research practices is usually to facilitate scientific discovery.

It is well known that theories are heavily influenced by the particular methods used (e.g., focus groups vs. randomized controlled experiments) and that using the same method over and over again to study a phenomenon makes the method part of the operationalized construct, what Shadish, Cook, & Campbell (2001) calls *mono-method bias.* But the problems I consider in this article are not related to the use of methods themselves, but to the use of *reasoning* in relation to methods. More importantly, the inferences made from data to theory is dependent on the type of logic, and therefore also a question of epistemology. If the incorrect type of logic is applied, the theory becomes weak, and the evidence might not substantiate the theory at all.

The problems of questionable research practices have been discussed thoroughly in medicine (e.g., Ioannidis, 2005), psychology (e.g., John, Loewenstein, & Prelec, 2012), political science (e.g., Franco, Malhotra, & Simonovits, 2015), and economics (e.g., Ioannidis, Stanley, & Doucouliagos, 2017). But these discussions has yet to reach communication science or political communication to the same extent. Notable exceptions in communication science exists, however, with regards to inflated p-values (Vermeulen et al., 2015), the extent of publication bias (Konijn, Schoot, Winter, & Ferguson, 2015), small sample sizes in experimental research (Matthes et al., 2015), researcher degrees of freedom in both structural equation modeling (Seaman & Weber, 2015) and the Implicit Association Test, IAT (Ellithorpe, Ewoldsen, & Velez, 2015).

These articles address important research practices, but one might get the impression that these problems are related to specific methods. However, these problems are related to the underlying epistemology of reasoning and logic. Although it would certainly help with a better understanding of, for example, statistics, it does not suffice since statistics is also dependent on reasoning and logic. For example, researchers always deduce hypotheses *from* theories and draw inferences *to* theories, which also makes philosophy of science necessary of attention. Most importantly, abductive logical inferences are in fact often used although researchers claim to use deductive or inductive logical inferences, which I will demonstrate in this article. Abduction is about observing some interesting fact and coming up with an explanation, induction is about verifying the explanation, and deduction is about explicating the explanation.

When ambiguity arise, for example, humans tend to interpret information in favor of their initial belief, a phenomenon known as confirmation bias (Nickerson, 1998). This phenomenon also manifest itself in what information we select as well as remember. At the same time, communication researchers primarily seek to verify hypotheses, rather than falsify them (Matthes et al., 2015). However, this is completely backwards on epistemological grounds since falsification is always stronger than verification (Lakatos, 1999; Meehl, 1990; Popper, 1992). This inadvertently leads to a willingness to confirm weak theories and, in turn, questionable research practices that leads to even weaker theories that can account for many results but seldom predict new observations in advance. However, where there are no predictions, there can be no knowledge (de Groot, 1969).

The aim here is to demonstrate how and why theories become weak from questionable research practices such as circular reasoning. I also present how the different types of logical inferences are related to circular reasoning and the use of Null Hypothesis Significance Testing (NHST) in exploratory and confirmatory research, and why a failure to separate these can inadvertently lead to false positive findings that considerably weaken the confidence we can put in the theories.

At the end I propose several solutions to amend these problems, primarily by a move toward a more open science with increased transparency: separating exploratory and confirmatory research, pre-registration of a priori hypotheses, and peer review before data collection. Many of these suggestions are far from new (see de Groot, 1969), but are nonetheless useful reminders to put appropriate confidence in the literature relative to the actual evidence.

2 Circular reasoning is reasoning in circles

Communication science is not only about describing and classifying phenomena of interest, but also about understanding and explaining the causes and effects of human communication, such as how political messages affect people's motivation to vote. Explanation thus serve an important role, and where there are explanations there is also risk of reasoning in circles.

Circular reasoning—*petitio principii* or *circulus in probando*—as first described by Aristotle in his *Prior Analytics*, occurs when a premise in an argument refers back to itself in order to establish a conclusion. Circular arguments do not give any new information but simply restate information in a different manner, or is dependent on itself in some way. In other words, empirical data brings nothing to the table. If a critic successfully identifies a circular argument, then it is often completely devastating for the argument. Spotting circularity is thus crucial in empirical research where conclusions are supposed to be based on evidence rather than assumptions.

The argument "God exists because God exists" is obviously circular and brings no evidence to the table, although the argument itself is tautological and therefore logically valid. As such, circular reasoning is often deductive (i.e., necessarily certain), and gives no new information about the world not already present in the premises. These circular arguments are thus utterly useless in empirical research.

The God argument serve as an illustration, but most empirical research is not carried out in this way by simple restatements. If any step in the research process is dependent and justified on a previous step, circular reasoning can occur, making it trickier to spot circular reasoning since the circle becomes larger (e.g., A is dependent on B, B is dependent on C, and C is dependent on A).

There are several types of circular reasoning, most notably circular explanations and circular analysis (Hahn, 2011), and some types are indeed part of feasible reasoning that are tentative, as much empirical research is.

2.1 Circular explanations

Why does opium make us sleepy? Because it has a dormative virtue. This classical example is a circular restatement since it "restate the phenomenon in question in different words and pretend to have offered an explanation" (Gigerenzer, 2011, p. 737), which can "both create a theoretical void and cover it up" (2011, p. 738).

Both "sleepy" and "dormative virtue" are logically equivalent, but all restatements are not necessarily fallacious reasoning. Researchers sometimes postulate latent factors that are dependent and justified by their measurable indicators, and the latent factors are thus said to *cause* the changes in the indicators. But the empirical evidence (i.e., a change in the indicator) alone cannot infer the existence of the latent factor without also assuming it. The latent factor is something that is metaphysically postulated, and the evidence thus act as a self-dependent justification (Hahn, 2011). In other words, we infer the latent factor by its indicators, and the indicators are caused by the latent factor, which is a circular explanation—but at least its circularity is probabilistically feasible reasoning (Hahn, 2011; Walton, 2008). It could be something else that caused the change in the indicator, but the more alternative explanations are ruled out, the more confident we can be in our belief that the latent factor actually caused the change.[1]

But if the latent factor does not exist, then the latent factor obviously cannot cause a change in the indicator, and would therefore constitute circular explanation. Whether self-dependent justifications are truly circular is thus an empirical question (Hahn, 2011), and the use of falsificationist hypothesis-testing (Popper, 1992) becomes extremely important.

2.2 Circular analysis

Circular analysis occur when the selection of data is dependent on, or justified by, the data itself. It is therefore a two-step process. For instance, a researcher may choose a theory that the researcher believes will explain a phenomenon. The theory then advises what data to select and hence analyze. The researcher then (1) use the theory to select what to analyze, and (2) analyze the data with the help of the theory. But this is circular since the data necessarily corroborates the theory. Such research project is deductive since it cannot account for any new information not already inherent in the selection criteria, although disguised as an inductive (or abductive) approach that uncovers new information from the

[1] I.e., we can update our posterior belief probabilistically in accordance with Bayes' theorem.

data. In other words, the analysis will only reiterate what the researcher selected in the first place.

Researchers may also “preselect data points on the basis of their ability to maximize some statistical criterion and then use those same data points to conduct, in full, the statistical analysis in question” (Hahn, 2011, p. 177), which cannot anything but support the hypothesis. However, not all preselections needs to be circular as long as they are independent of the results:

> selective analysis is a powerful tool and is perfectly justified whenever the results are statistically independent of the selection criterion under the null hypothesis (Kriegeskorte et al. as quoted in Hahn, 2011, p. 178).

If the sampling criteria is dependent on the data, however, the research is necessary tautological since one cannot “find” something that was put there in the first place. This problem is most pronounced in research areas where careful selection of observations is the most feasible solution, since the researcher usually samples instances that are most likely to yield fruitful results (see Nickerson, 1998).

3 How logical inferences relate to circular reasoning

3.1 Why deduction is stronger than both induction and abduction

Deduction, induction, and abduction are three modes of logical inferences used to draw conclusions, which also reflect decreasing degree of certainty. Deduction starts out with known premises and draws a necessary conclusion. That means that there is no uncertainty, and the conclusion follows with logical necessity. Induction starts out with particulars, and generalizes beyond the premises toward a conclusion, and the conclusion is therefore always uncertain to some degree. Abduction, on the other hand, starts out with some interesting fact that has been observed, from which explanations can be suggested.[2]

Abduction is often used in exploratory research, where the point is to generate theories, while deduction and induction is often used in confirmatory research, where the point is to test theories. In short, “abduction creates, deduction explicates, and induction verifies” (Yu, 2006, p. 62). When these modes are conflated, however, the resulting conclusions can be circular.

Deduction might be best known in the empirical sciences under the hypothetical-deductive method, that begins with theory from which hypotheses are derived and tested with the goal to verify or falsify the theory. Alternative explanations are tested, one by one, to rule them out, and therefore explicate the theory’s boundary conditions. The more and stronger falsification attempts the theory can withstand, the more confidence we should put in the theory as true, at least until a better theory comes along. If the hypothesis is verified, the reasoning is based on *modus ponens* (Table 1), and the conclusion is therefore logically valid. If the hypothesis is falsified, the reasoning is based on *modus tollens*, and that means

[2] Deduction and induction is often described in method text books as reasoning from general-to-specific and specific-to-general, respectively. But this is wrong. Both deduction and induction can be used either as general-to-specific and specific-to-general.

that falsification constitutes a logical proof that is actually more severe than modus ponens (Popper, 1992). Both modus ponens and modus tollens is logically valid inferences (i.e., deductions), which means that if the premises are true, the conclusion must be true. Both can make a substantial contribution to a theory, but *only* if one have a strong theory (that makes predictions) to begin with. However, modus ponens is never actually fully certain in empirical research, given the problem of induction (Vickers, 2016). That means that conclusions from modus tollens is always stronger than modus ponens in empirical research. Stated in another way, modus tollens *destroys*.

Table 1: Four forms of logical inferences relevant to empirical research, based on the material implication (the pattern *if P, then Q*), where P is the antecedent and Q is the consequent. A valid conclusion is also deductive.

Name	**Logical form**	**Conclusion**	**Should be used in**
Modus ponens	If P, then Q P Therefore, Q	Valid	Confirmatory research
Modus tollens	If P, then Q Not Q Therefore, not P	Valid	Confirmatory research
Affirming the consequent	If P, then Q Q Therefore, P	Invalid	Exploratory research
Denying the antecedent	If P, then Q Not P Therefore, not Q	Invalid	Exploratory research

Researchers nonetheless often strive to verify their results by affirming the consequent, which is a logically invalid deductive inference. In other words, researchers are not taking advantage of neither the strength of modus ponens and verification, nor the strength of modus tollens and falsification. Although, affirming the consequent is by no means fallacious, but instead similar to abduction, or *inference to the best explanation* as it is also called sometimes (Douven, 2016):

> The surprising fact, C, is observed
> But if A were true, C would be a matter of course
> Hence, there is reason to suspect that A is true

Even though this is a perfectly sensible inference, it is nonetheless not a (deductive) test of a theory like modus tollens or modus ponens, which means that the conclusion is substantially less certain. It is therefore ethically questionable research practice to find something, abductively, and then rewrite the article as if it were explicitly tested for (de Groot, 1969; Kerr, 1998; Wagenmakers, Wetzels, Borsboom, Maas, & Kievit, 2012). When exploratory analyses and confirmatory analyses are not separated, the scientific literature can become meaningless and instead reflect researchers ability to spin a persuasive narrative that fit the data. Abductive inferences are only as certain as the number of alternative explanations that have effectively been ruled out, and abduction is therefore in dire need of modus tollens.

When researchers begin with a vague idea of either hypotheses or theories, they can derive and test several hypotheses in order to verify at least one of them. The hypothesis that lend a significant result is then chosen and selectively reported in the final article. This approach is also abductive. An unfortunate consequence is that the more attempts to verify the theory, the *less* confidence we should put in the theory as true, since more attempts will ultimately yield at least *some* support for the theory. Publication bias therefore work in the precise opposite way as intended. Instead of only negative findings being put in the file drawer, the rate of positive findings are greatly exaggerated. Most importantly, abduction, induction and deduction have their own strengths and weaknesses and should be considered *together* over the course of time to give a comprehensive picture of the phenomenon in question (Douven, 2016).

Qualitative research is not exempt from the problems outlined here, since the problem is not due to methods, but of reasoning:

> The errors of reasoning and irresponsible conclusions from the quantitative case originate from acts, which in principle are performed exactly the same in qualitative research. In that case, one attempts to "let the material speak" as well, by ordering the findings *ad hoc* and by systematizing and associating them by means of perspectives that were obtained *ad hoc*; likewise, one often uses the findings to draw conclusions that purport to generalize to other situations. (de Groot, 1956, p. 8)

However, the difference between quantitative research and qualitative research is that these errors are more often demonstrable in quantitative research (de Groot, 1956).

3.2 Why the logical distance makes it harder to infer conclusions

Null Hypothesis Significance Testing (NHST) is based on experimental agronomy research (Yu, 2006). The logical distance between data and theory is much shorter when studying which fertilizer will best grow a corn field, compared to the distance between data and theory when studying how, for example, political messages affect people's voting behavior. Consequently, corroboration of a theory becomes harder as the logical distance between data and theory increase (Meehl, 1990). Researchers therefore have to assume many auxiliary hypotheses in order for the result to be valid (Lakatos, 1999).

Hypotheses are therefore conjoined with auxiliary hypotheses that is either known or unknown, sometimes spelled out but often not, which means that we can never be entirely certain that a logically valid inference (e.g., modus ponens or modus tollens) is valid in practice. A null result could be caused by low statistical power, improper survey measurement or whatever; and the same argument applies to positive findings. For instance, a recent review of political communication in the high-choice media environment reminded us that "not all trends are global trends, nor do they affect all countries to the same extent" (Van Aelst et al., 2017, p. 19). In other words, all hypotheses may occasionally be true in some context, at some moment in time, viewed from a certain perspective. This highlights the need to delineate the boundary conditions when the theory is false, which should always be spelled out explicitly before data collection (Lakatos, 1999).

4 Causes of circular reasoning

4.1 Conflating exploratory and confirmatory analyses — and how pre-registration can help

The primary culprit to circular reasoning and weak theories is a failure to distinguish between exploratory and confirmatory analysis. Exploratory and confirmatory analysis corresponds to the context of discovery and context of justification. Where basically anything goes in the context of discovery, where fantasies can provide insightful material for the sciences; the context of justification requires more rigorous methodology and reasoning that subjects data to severe tests, most notably the possibility of the results being false (Popper, 1992). When the exploratory phase is cast into a confirmatory language, circular reasoning easily occurs and provides weak justification for theories (Simmons, Nelson, & Simonsohn, 2011). A theory must therefore be able to make predictions *in advance*. If predictions are made after the results are known, it is not hypothesis-testing, but hypothesis-generating, which is based on abduction rather than induction or deduction. And where there are no predictions, there can be no knowledge (de Groot, 1969).

Hypothesizing after the results are known, or *HARKing*, occurs when the researcher presents a post hoc hypothesis as if it were in fact an a priori hypothesis (Kerr, 1998). This is problematic because an hypothesis can seldom be tested on the same data that was used to generate the hypothesis, since that would be to "find" something that was actually put there in the first place (Hahn, 2011). Data can only provide evidence for a hypothesis if the hypothesis has been subjected to a test that has the possibility of being true or false, and is severe enough to expose flaws if they are present. If an hypothesis is made to support the result post hoc, the result is in fact abductive—not deductive or inductive—and therefore at best suggestive of what future studies should look in to. Presenting a post hoc hypothesis as an a priori hypothesis is therefore often circular analysis.

HARKing is unfortunately a common practice among researchers (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017), and is a serious problem because it easily leads to post hoc justification of weak theories due to researcher's confirmation bias (Nickerson, 1998), hindsight bias (Roese & Vohs, 2012) or motivated reasoning (Kunda, 1990). An infinite number of hypotheses can fit the data by abduction, and the more the researcher is HARKing, the greater the probability of finding results that are statistically significant way more than the conventional 5% of the time (Ioannidis, 2005; Kerr, 1998; Seaman & Weber, 2015; Simmons et al., 2011).

Lakatos (1999) argued that in a progressive research program, theory is ahead of data. But in a degenerating research program, theory is trying to keep up with all new data. This is perhaps most noticeable in "big data" and Twitter research, where the easy access to data has lead to thousands of empirical research articles (Williams, Terras, & Warwick, 2013), but also a substantial lack of theorizing. For example, a review of election forecasting with Twitter data found that prediction models are exclusively derived *after* the election results are known (Gayo-Avello, 2013). However precise those models may be, as long as they are not tested prior to elections we should not put any confidence in predictions that are tailored to fit the data. They are abductive inferences, and therefore at best suggestive.

When it comes to humans and the social science, everything is correlated with everything to some degree, a phenomenon known as the *crud factor* (Meehl, 1990). This makes it even more important to create theories that predict, especially in a time where data (and therefore spurious correlations due to the crud factor) are overwhelming. Theories that predict therefore becomes *more* important with the increase of big data, not less. As statistician's many times pointed out, it is "Far better [with] an approximate answer to the *right* question, which is often vague, than an *exact* answer to the wrong question, which can always be made precise" (Tukey, 1962, p. 13). Although big data may have arrived, big insights have not.

Pre-registration clearly separate confirmatory from exploratory analyses (Wagenmakers et al., 2012). In a pre-registered study, hypotheses, power analyses and exclusion criteria (among other things) are explicitly stated prior to any data collection (see Veer & Giner-Sorolla, 2016). The pre-registration is then made available in a public repository.[3] Pre-registration force researchers to think about theory and the study rationale and statistical tests *before* the study is carried out, and also increase transparency.

Pre-registration originated in medicine. When pharmacologists pre-registered their studies and therefore had to state their hypotheses before data collection, statistical significant findings dropped from 57% to 8%, and the rate worsened over time (Kaplan & Irvin, 2015). Another study found that only 11% of promising preclinical medical studies could be confirmed in later studies (Begley & Ellis, 2012). Psychology and political science is no better. Of one hundred classical psychology studies, about half replicated (Open Science Collaboration, 2015). By comparing experimental political survey questionnaires with the published article, 80% of the studies did not report all experimental conditions and outcomes (Franco et al., 2015). This indicates that researchers are not necessarily testing a priori hypotheses, but rather explore the data in order to find out which hypotheses and results to selectively include in the final article. However, that is abduction and should be treated (and written) as such.

One common objection is that pre-registration limits creativity and stifles scientific discovery. But the point is not to enforce hypothesis-testing at the expense of hypothesis-generation, but to draw a sharp line between them. This line prevents HARKing as well as limit the researcher degrees of freedom (Simmons et al., 2011; Wagenmakers et al., 2012). In other words, to limit the possibility of recasting an exploratory analysis into a confirmatory language.

4.2 P-hacking and the approaching significance

P-hacking refers to the common practice in Null Hypothesis Significance Testing (NHST) to repeat statistical analyses (with variations) with the purpose of getting a $p < .05$. It has

[3] There are several internet repositories for pre-registrations, such as osf.io and aspredicted.org, and some of them can be set to become public *after* the article is published while still keeping the original submission date, making it easy to keep confidentiality during the process.

also been called *multiple comparisons problem*, *garden of forking paths* (Gelman & Loken, 2013), and sometimes joined together under the heading of *researcher degrees of freedom* (Simmons et al., 2011). P-hacking is not necessarily an accusation of scientific misconduct, but a failure to appropriately plan the analysis.

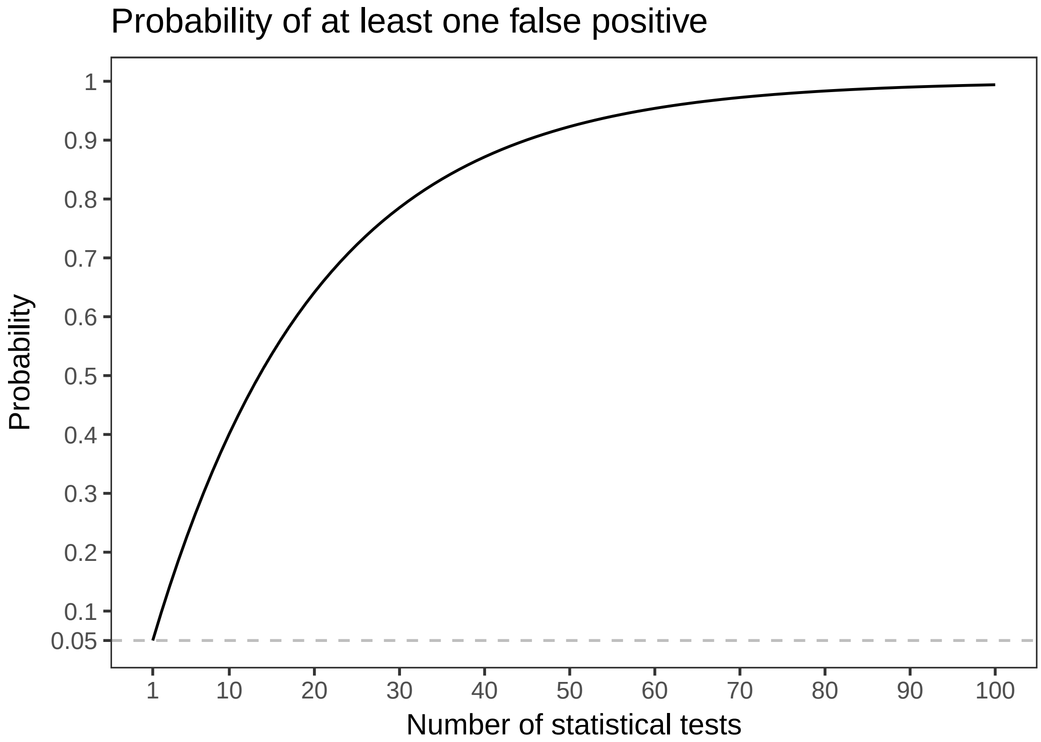

Figure 1: The more statistical tests you perform, the easier it gets to obtain a low p-value. The graph shows the probability of at least one false positive finding (type I error), at the typical .05 significance level, as the number of statistical tests increases (de Groot, 1956). After 14 tests, for example, the result is in actuality significant at the .51 level—not the .05 level—which is equivalent to tossing a coin to determine whether a discovery was "found".

The logic of NHST is to compute the probability of the observed data (D), given that the null hypothesis (H_0) is true, formally $P(D|H_0)$. This results in the infamous p-value. NHST and the p-value can thus only inform us whether the null hypothesis should be rejected, it can never say that the null hypothesis or alternative hypothesis should be confirmed. Only Bayesian statistics can give the probability of an hypothesis, given the data, formally $P(H_1|D)$.

Treating p-values in an abductive way can give impossible results. For example, Simmons et al. (2011) demonstrated the problem by adding and removing different covariates to their statistical model until they found a $p < .05$ that music affects a person's age (an impossible result). The reason for this is the flexibility the researcher has in the analysis, where testing different covariates and removing outliers ad hoc, for example, can dramatically change the p-value, especially when the sample size is small.

P-values in exploratory analyses are more or less meaningless, because the requirement of independence and random sampling has often never been met when results are selectively reported and therefore abductively inferred. Thus, the p-value will eventually drop below .05 as the number of statistical tests increases (Figure 1), and this is a function of how the p-value works, and not a sign of a discovery. This is, unfortunately, seldom understood by researchers (Gigerenzer, 2004). Blurring the line between exploratory and confirmatory analyses increase the probability of interpreting findings as true when they are in fact false, or even meaningless, which results in lots of studies that are actually analyzing statistical artifacts such as sampling noise rather than true effects (Ioannidis, 2005).

Let say we have an experiment with two groups and want to find the differences between group means. In the top graph of Figure 2, we can see the effects of increasing the sample size and repeatedly doing a t-test. We can see that the p-value is unstable until it permanently drops below the .05 threshold. In the bottom graph, there is no difference between group means. The p-value nonetheless drops below the .05 threshold several times, and that is because smaller samples have larger variance and therefore also larger effect sizes (everything else being equal).

A common belief is that a $p < .05$ means that an effect was found (and indirectly that the statistical power was sufficient to find the effect). But this belief is mistaken. If there is no effect, *all p-values are equally likely to occur* since p-values are then uniformly distributed (Figure 2, bottom graph). This means that a p-value of .001 is just as likely as a p-value of .999, if no effect exists (e.g., Pearson's $r = 0$). But the p-value will randomly fall below .05 about 5% of the time, if we choose the 5% significance level.

The point is that small samples combined with removal of outliers, adding covariates and redoing statistical tests over and over again will guarantee us significant results sooner or later, but such results are meaningless (Simmons et al., 2011). A literature that primarily deals with small samples, publication bias, and HARKing, is therefore likely abundant with inflated effects that in practicality does not exist (Ioannidis, 2005). This have two important implications. First, if p-values are often meaningless in exploratory analyses, and exploratory analyses is often disguised as a confirmatory analysis, it follows that the literature contains a great deal of false positives that we should not take seriously. Second, if it is the case that researchers predominantly try to verify theories, rather than falsify them, we should expect the literature to contain p-values just below the .05 threshold and more hypotheses confirmed than falsified. In fact, that seems to be the case (Matthes et al., 2015; Vermeulen et al., 2015). The false positive rate in communication science is estimated between 8% and 35% (Vermeulen et al., 2015). Misreported p-values is also still frequent in communication science (Figure 3) and these errors "appear to be driven by researchers' motivations to demonstrate significant relationships" (Vermeulen et al., 2015, p. 270).

4.3 Publication bias, low power, and meaningless replications

By emphasizing that each submitted manuscript contributes positively to our knowledge, there is a publication bias against null results. Focusing on new and interesting results will undoubtedly lead to false discoveries since the incentives are to disproportionately verify theories instead of falsifying them. In other words, saying something new is more important than saying something true. This is partly based on the view that a significant result "proves something" whereas as non-significant result "doesn't prove anything" or is perhaps "inconclusive", which is expressed in researchers willingness to find significant results (Matthes et

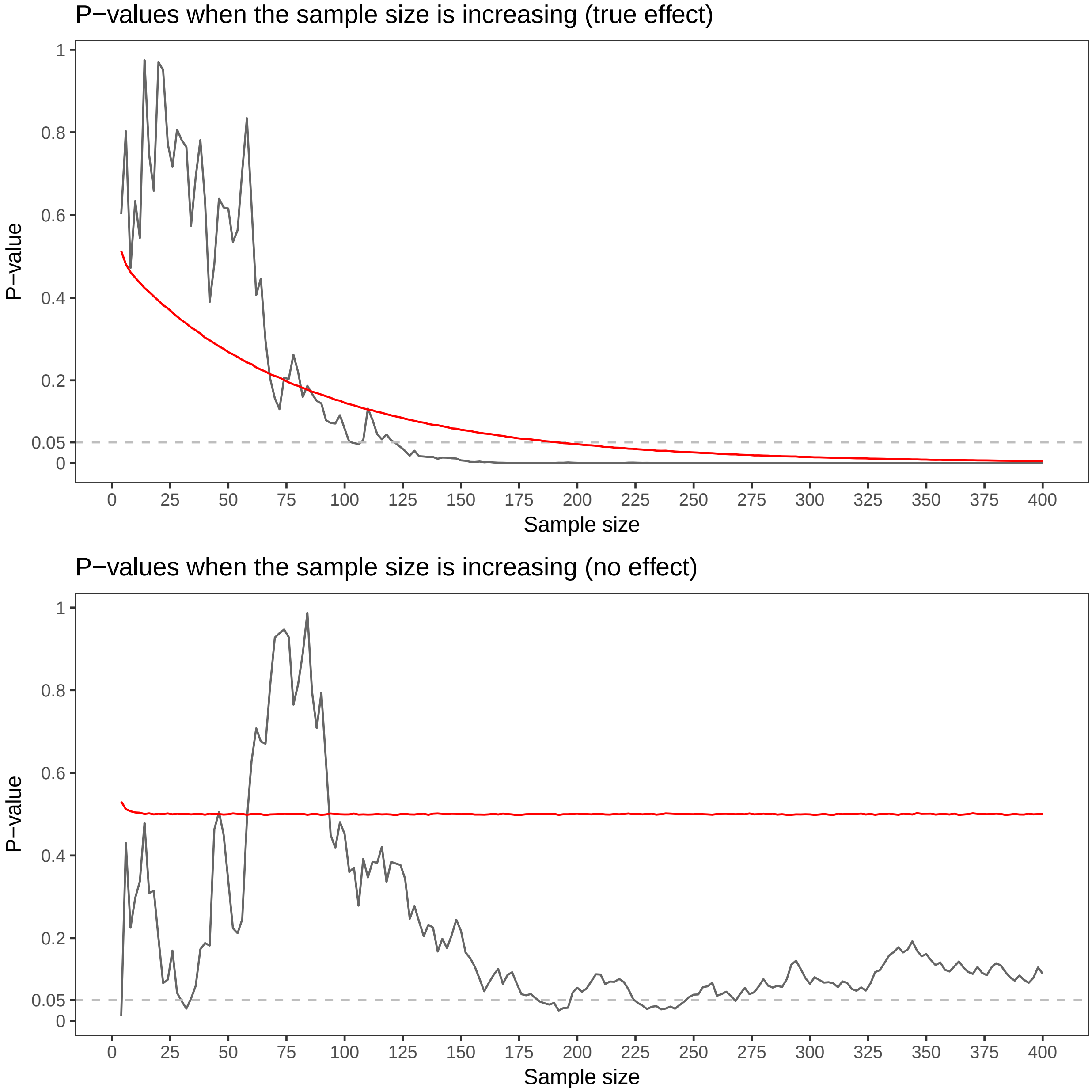

Figure 2: P-values cannot be trusted at low sample sizes. Top graph shows fluctuation of p-values at different sample sizes and redoing a t-test when there is a true difference between the means (Cohen's d = .4). The p-value fluctuates heavily and cannot really be trusted until about n = 200, in this particular case. Bottom graph also shows fluctuation of p-values, but without differences between group means (d = 0), and so the p-values drop below the significance threshold at about the expected 5% of the time. In both graphs, the smooth solid line shows the average p-values based on 100,000 Monte Carlo simulations at each sample size. Reproducible R script in appendix.

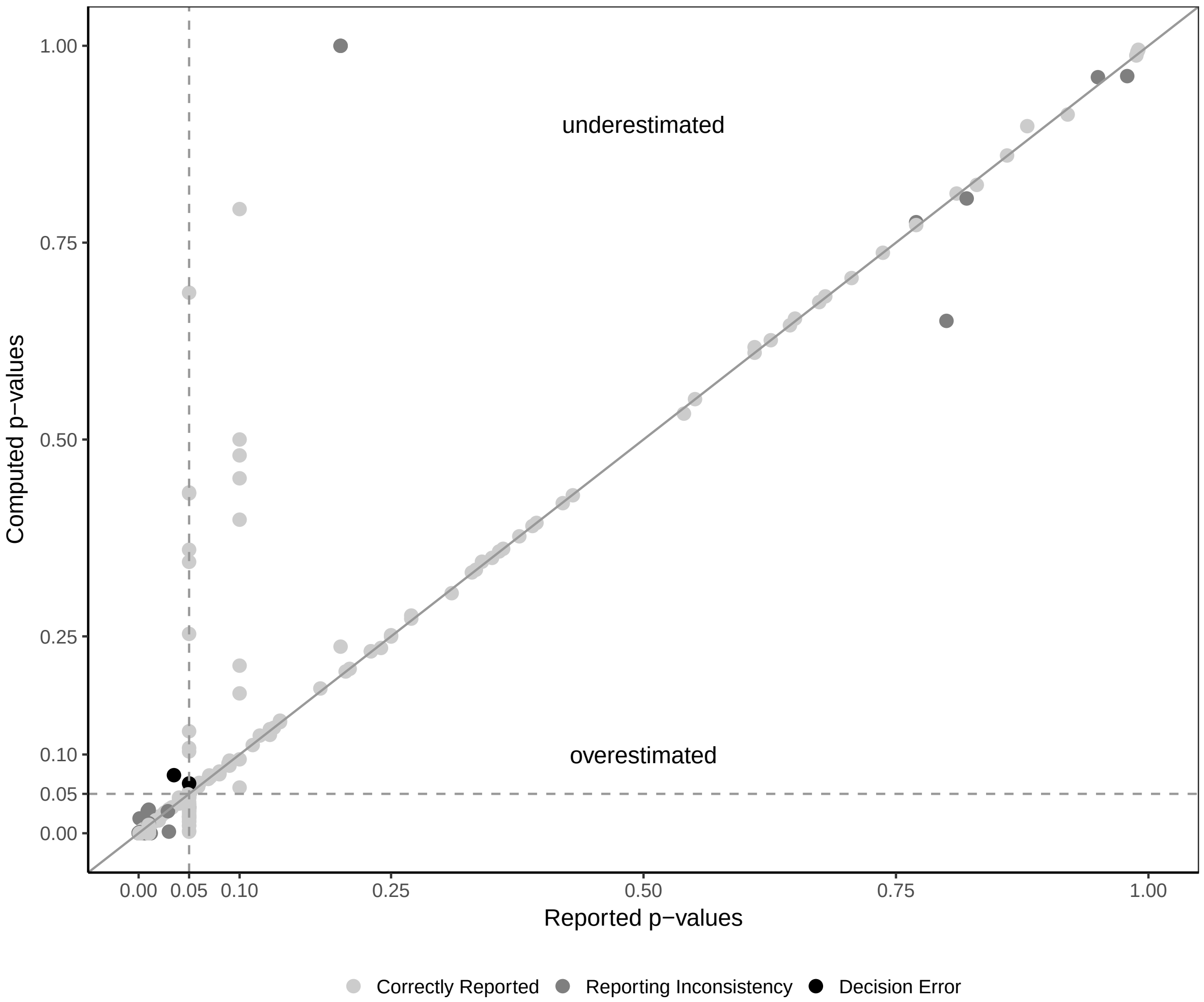

Figure 3: About one tenth of the sampled p-values were erroneously reported. Extracted statistics (n = 393) from three years of empirical articles in Journal of Communication, 2015 issue 1 to 2017 issue 4 (n = 119) and articles in Political Communication, 2015 issue 1 to 2017 issue 3 (n = 83). A total of 9% (n = 35) of the identified p-values were erroneously reported, according to statcheck, an R package for automatic analysis. *Correctly Reported* refers to computed p-values congruent with reported p-values, *Reporting Inconsistency* refers to computed p-values not congruent with reported p-values. *Decision Error* refers to reported p-value being significant whereas computed p-value is not, or vice versa. Method described in appendix.

al., 2015; Meehl, 1990; Vermeulen et al., 2015; Wagenmakers et al., 2012), as well as the number of verified hypotheses that are larger than the number of falsified hypotheses in experimental communication research, a ratio that is increasing over time (Matthes et al., 2015).

But this is precisely backwards, and wreak havoc on falsificationist ideals of hypothesis-testing (Lakatos, 1999; Meehl, 1990; Popper, 1992). If a null result proves nothing, then there is no need to do a study at all, since only positive outcomes are counted—and all theories are thus necessarily corroborated given enough time. Given that deductive inference is necessary for knowledge, though not sufficient, we implicitly say that the weakest logical inference is the best since it lead to positive findings ("discoveries"). That is ironic given that the typical α and β is set at .05 and .20 (if they are even set at all). In plain English, a false negative (type II error) is four times more permissible than a false positive (type I error), which is the precise opposite of what researchers should do if they were actually looking for discoveries. This might be an instance of what Gigerenzer (2004) calls *mindless statistics* of the null ritual.

Since falsificationist hypothesis-testing relies on modus tollens, it is therefore the *only* way of actually (dis)proving results. Researchers should therefore strive for statistical *non-significant* results—the precise opposite of current bias toward significant results. The revised edition of The European Code of Conduct for Research Integrity now specifically states that authors and publishers should "consider negative results to be as valid as positive findings for publication and dissemination" (Allea, 2017, p. 5). This is a good step forward.

Every introductory statistics textbook warns against spurious correlations, but in practice they are encouraged. If we publish positive results, found abductively, but not subsequent deductive refutation attempts, we simultaneously say that spurious correlations are acceptable as long as they are novel, interesting and perhaps even cast into a confirmatory language with hypotheses rather than research questions. But this devalues knowledge and truth. Just as the democratic system needs checks and balances, so does science. Verification must be balanced with falsification, hypothesis-generating research must be balanced with hypothesis-testing research, and abduction must be balanced with induction and deduction.

What might be beneficial for an individual researcher might not be beneficial for the advancement of knowledge, especially not in a academic system that promotes quantity over quality. For instance, researchers increasingly describe their findings in positive words, such as amazing, astonishing, groundbreaking, novel, unique, unprecedented, and so on (Vinkers, Tijdink, & Otte, 2015)—at the same time as pre-registration have made significant findings drop considerably (Begley & Ellis, 2012; Franco et al., 2015; Kaplan & Irvin, 2015; Open Science Collaboration, 2015).

There is a fallacy of composition, i.e., what is true of the individual parts is not necessarily true of the whole (Hamblin, 1970), since what is considered a good contribution for each individual researcher or journal, is not necessarily a good contribution to neither knowledge nor the collective of scientists as a whole. The paradoxical consequence is that each individual researcher can thrive climbing the academic career ladder at the same time as scientific knowledge is declining. It is like observing that the Titanic may be leaking in water, but at least it is not leaking at *my* end.

If falsification is so important, attention also have to be given to (the lack of) statistical power. Statistical power refers to the "probability that [the statistical test] will lead to the rejection of the null hypothesis, i.e. the probability that it will result in the conclusion that the phenomenon exists" (Cohen, 1988, p. 4). Weak theorizing and low statistical power can put up straw-men hypotheses for testing, which can then easily be falsified. Two examples can illustrate this point.

A recent pre-registered experiment did not find an effect of a previous failed attempt on media priming (Kobayashi, Miura, & Inamasu, 2017). One could argue that this have shown that priming does not exist by the use of falsification and modus tollens, at least in this particular case. The replication used 104 participants and the authors said they were able to "maximize the statistical power" from 94 in the original study (see pre-registration at osf.io/bgk29). However, the replication needed 404 participants in order to find an effect with *at least* the same size as the original experiment.[4] But even if the original experiment was a fluke, and thereby did find an effect (which it didn't), the replication experiment cannot tell us anything because the sample size is too small to successfully reject the null hypothesis even if there was an effect of the same size or smaller. This is an unfortunate situation since many social priming effects have not been able to replicate (Open Science Collaboration, 2015), and whether this is also the case for media priming is thus an highly important question for political communication.

Another example. A recent experiment on motivated reasoning had the same problem. The original experiment had over 1,000 participants. A later independent replication had 55 participants, and found no effect. A replication by the original authors with over 1,000 participants once again found the effect (Kahan & Peters, 2017).

The two replications had too few participants and therefore too low statistical power. Such replications are meaningless and "a tremendous amount of taxpayer money goes down the drain in research that pseudotests theories" (Meehl, 1990, p. 230).

The median sample size in communication science is n = 154 (Vermeulen et al., 2015), which is at least higher than in social psychology's n = 95.[5] The average effect size in social psychology is r = .21 (Richard, Bond, & Stokes-Zoota, 2003). For that size, a total of 246 participants is needed in order to find an effect with the same size or larger.[6] This means that half of the studies in communication science typically have lower power than needed in order to find an average sized effect, if we grant some assumptions and that social psychology is somewhat equivalent to communication science. Once again, small samples is highly problematic if discoveries are valued, but very convenient when HARKing.

[4] By extracting the descriptive statistics from the original experiment, reported in Kobayashi et al. (2017), we can calculate Cohen's d =

$$\frac{M_1 - M_2}{\sqrt{\frac{SD_1^2+SD_2^2}{2}}} = \frac{2.46 - 2.33}{\sqrt{\frac{0.39^2+0.53^2}{2}}} \approx 0.28$$

from which an a priori power analysis reveals that 404 participants are needed assuming .80 power, .05 two-tailed alpha, and independent t-test. R code for power analysis: `ceiling(power.t.test(d=0.28, sig.level=0.05, power=0.80, alternative="two.sided")$n) * 2`.

[5] The use of median, rather than the mean, is because the large variability of sample sizes in communication science, where some studies have thousands and thousands of participants.

[6] Pearson's r = .21 is equivalent to Cohen's d = .36. R code for power analysis given some assumptions: `ceiling(power.t.test(d=0.36, sig.level=0.05, power=0.80, alternative="two.sided")$n) * 2`.

5 How circular reasoning make theories weaker

Thus far I have hopefully shown the problems of circular reasoning, HARKing, low statistical power, and the backwards reasoning of publication bias given the strength of falsificationist hypothesis-testing.

I now turn to the problems of theories that do not exclude or predict. Predictions is one of the most important aspects of theories, perhaps more so in communication science than other fields due to the use of abundant digital media data that favors extensive empirical work, sometimes at the expense of theoretical work. Theories may be initially embraced and generate substantive empirical work, but later abandoned due to loss of interest rather than being clearly falsified (Meehl, 1990). If hypotheses are written in a way that does not permit clear-cut falsification, even in principle, this problem will persist. In contrast to a predictive theory, weak theories can account for almost all evidence after the results are known, and make little to no new predictions. “In the absence of theory, the temptation is to look first at the data and then see what is significant.” (Gigerenzer, 2004, p. 602). But even as theories do make new predictions, they are sometimes so vague that they are not even falsifiable.

5.1 Meaningless hypotheses that does not exclude

A theory is weak if it is flexible enough to account for more data than it can exclude. With the rejection of a straw-man null hypothesis seen as corroborating evidence, researchers sometimes postulate meaningless hypotheses, such as “the correlation between X and Y will be non-zero”, which eliminates practically nothing. Directional hypotheses are better, such as “group A is larger than group B”, but still only eliminates half of the possible results. Despite the use of statistics, these hypotheses are not quantified, but only qualified.

If the threshold for verification is set too low, almost all hypotheses can be verified (and thereby corroborate the theory). Given publication bias toward verification of positive results, this also creates incentives for weak hypotheses that include more data than they exclude. Since NHST depends on rejecting a null, the alternative hypothesis can be framed in any way to be counted as supporting evidence at $p < .05$. Moreover, if the researcher is HARKing, the data necessarily supports the hypothesis since it does no exclude anything at all by means of circular reasoning.

One suggested improvement to hypotheses is to quantify how large a difference should be if it is to be considered meaningful, such as “group A is at least 10% larger than group B”. The standard objection is that social science does not permit us to be explicit about relative sizes, and we must account for all the nuances and complexity of the social world. However, that is misguided on at least two accounts.

First, just because the world is complex, does not mean our theories need to. Demanding more nuance is an easy way to criticize anything, since everything is always “more complicated”. It may in fact be viewed as antitheoretical as it “blocks the process of abstraction on which theory depends, and it inhibits the creative process that makes theorizing a useful activity” (Healy, 2017, p. 119). A good concept, for example, should aim for exclusion by throwing away the details of particulars, and focus on similarities and abstraction.

Second, if we do not know what is considered a meaningful difference beforehand, how could we determine what a meaningful difference is after we have seen the results? Weak theories can be corroborated by almost any evidence, since they do not exclude enough, lending them little to no explanatory power. Without predictions, it is just a matter of personal preference whether the hypothesis corroborate the theory. Theories that exclude more actually explains more, since its degree of falsifiability increases, while a theory that include more will have a harder time being falsified. Communication science may therefore progress quicker with theories that exclude more, since those theories may be more easily discarded. “We measure, we define, we compute, we analyze. But we do not exclude.” (Platt, 1964, p. 352).

Borrowing from the philosophy of language, *intension* and *extension* can be useful tools in thinking about hypotheses and theories. Intension is how we define a concept, whereas extension is all tangible things the concept applies to. A concept with a large intension has a small extension, and vice versa. The concept “media” has a small intension, but a large extension since television, radio, letters and so on can fit the definition. “The largest social networking site in the world with a blue F as logo”, on the other hand, has a large intension and therefore small extension since very few things fit this definition. This means that the more the definition includes, the more things it necessarily excludes, and is therefore more easily showed to be wrong. Transferring this reasoning to hypotheses, it means that a more detailed hypothesis thus have a larger intension, and thus a smaller extension since fewer results can corroborate the hypothesis. The more detailed the hypothesis, the more it exclude; and the more it excludes, the more powerful it becomes in predicting.

This argument is also applicable to model fitting. A complex statistical model will exclude more data than a loose model flexible enough to fit any data. Therefore, a model that can only be fitted to a very limited data does not need to be an instance of circular analysis—even if this model is fitted abductively—if and only if probable alternative models are thereby effectively ruled out. In other words, when a complex model only fits a particular data, it is likely that the model actually is a true model that underlies the data (Hahn, 2011). However, if the model is flexible enough to account for a large amount of data, then the false positive rate is still exceptionally high and circular reasoning due to HARKing can occur.

Another suggested improvement to hypotheses, and perhaps more straightforward, is to use conditional statements, such as “group A is larger than group B, *but only among youngsters*”, which for each additional conditional decrease the probability of obtaining a corroborating result, and thus increase explanatory power. A surprising result, as reflected by a low p-value, can then be held with greater certainty. Another suggestion is to predict the *form* of a relationship, such a curvilinear (Meehl, 1967). Regardless of solution, the need for stronger theories that predict and exclude are nonetheless warranted.

5.2 Not letting theories compete

A field with a large number of theories is a sign of a healthy field. Even though all theories are underdetermined by data (i.e., data may fit many theories), there is also a problem that occurs when data are contradicted by two theories, but theories are selected after the results are known. Take the example of media effects. One idea is that the frequent media coverage make certain attitudes more salient. Another idea is that the frequent media coverage leads to familiarity and boredom, thus weakening the salience on attitudes (i.e., *habituation hypothesis*).

If a researcher is HARKing, and afterwards select the theory that can best explain the data, circular reasoning occurs. Since the theories are not explicitly tested *against* each other, all theories will necessarily be supported since data always supports one of them. However, prediction will suffer because the specific condition of *when* one theory is more relevant than the other does not become known, with the unfortunate consequence that *both* theories becomes unfalsifiable in conjunction due to circular reasoning (since the conclusion is not independent of the analysis). Even though it may seem obvious that one cannot pick and choose neither hypotheses nor theories after the results are known (at least not in hypothesis-testing research), this way of working is encouraged, for example in the book *The Compleat Academic*, cited almost 300 times according to Google Scholar:

> There are two possible articles you can write: (a) the article you planned to write when you designed your study or (b) the article that makes the most sense now that you have seen the results. They are rarely the same, and the correct answer is (b). (Bem, 2012, p. 186)

This is a seriously dubious suggestion and conflates hypothesis-testing with hypothesis-generating research. If you set out to test a theory, you should write the article you planned to write and report failed results (since a failure, based on modus tollens, is the most certain findings of all), as well as results found during exploration. But you can *never* treat results found during exploration as a result that originated from an a priori hypothesis, since that would require a time machine. They are based on different logics, which means that we should only be so confident in the result as the logic permits. The type of inference must therefore be clearly stated (such as “we found an effect” or “we tested a specific hypothesis”).

Platt (1964) suggested that it was precisely the lack of competing theories that was the reason for the lack of clearly falsifiable hypotheses, and asked what would *dis*prove your hypothesis. Since people are prone to dichotomous thinking and fallible at probabilistic thinking, a frequent use of rival hypotheses could reinvigorate both empirical and theoretical work, since two hypotheses (theories) tested simultaneously will undoubtedly cast shadow on the (weak) theories if the data can be made to support both or none of them (Platt, 1964). A new theory should account for the old anomalies, old observations, and most importantly, predict new observations that can be tested in future studies (Lakatos, 1999). Theories should therefore pass *harder* hurdles with time, not easier (Meehl, 1967).

Saving a theory from falsification by adding ad hoc hypotheses as a protective belt around the hard core of the theory may be permissible conjectures as the ad hoc hypotheses may actually turn out to be true in the long run. Popper (1992) may have cautioned against it, but Lakatos (1999) contended that researcher cannot have an “instant rationality” after each hypothesis test, but instead have to evaluate theories over time. But as each ad hoc hypothesis is added, a new falsification attempt should begin. There must be a balance between abductive and inductive/deductive inferences. If abduction is piled on top of another abduction, there is no need for science in the first place since anything can be made true (Simmons et al., 2011).

There is always a call for more empirical research, but that time is not very well spent if the conditions when a theory is expected to be false is not clearly defined, statistical tests underpowered and conclusions HARKed. Perhaps time would be better spent by composing better theories rather than more empirical work that try to test hypotheses that cannot be shown to be false (neither empirically nor theoretically), and where reviews ends with the frustrating remark that half of the studies show P, while the other half show *non-P*, followed by the conclusion that more empirical work is needed. However, it is paradoxical that verified studies are not separated from falsified studies in reviews (Meehl, 1967). They are not equivalent and cannot be given equal weight, since falsification *destroys*. More empirical work will therefore never solve problems that originate from questionable research practices, circular reasoning, or a motivation to disproportionally publish novel discoveries rather than to justify them by means of falsification attempts.

6 Concluding remarks

Circular reasoning in empirical sciences occurs when a researcher assumes the very conclusion that is sought to be demonstrated in the first place. Circular reasoning is often devastating for a conclusion in an empirical article, but need not always be. For example, latent factors may be postulated by their measurable indicators, even though it is (probabilistically) circular.

The problem of circularity depends first and foremost on reasoning, logic, and therefore epistemology—not methods. The problems of circularity can lead to weak theories that can account for more findings than they exclude, which in turn does not facilitate predictions. A theory must exclude more than it can encompass, and also make predictions *in advance*. If predictions are made after the results are known, it is not hypothesis-testing but hypothesis-generating research, which is based on abduction rather than induction or deduction. Because where there are no predictions, there can be no knowledge (de Groot, 1969).

The main culprit is when confirmatory analyses is not separated from exploratory analyses. Confirmatory analysis is based on deduction and induction, while exploratory analysis is based on abduction, where the two former have a different role and poses a more severe test of a theory than the latter.

Exploration is necessary and should neither be abandoned nor replaced by confirmatory analysis. It should, however, be clearly separated from its confirmatory counterpart. If they are not treated as such, we may put more confidence in the result than is warranted by the evidence, and instead rely on meaningless circular reasoning.

The problems discussed here are basically ethical in nature and primarily refers to misrepresentation of the scientific literature. The solutions are therefore trivial—such as spelling out whether an effect was found during exploration or explicitly tested for—but nonetheless something that is both often violated (see Agnoli et al., 2017) and encouraged (e.g., Bem, 2012). This is understandable, given the academic reward system valuing publications with significant findings. As previous proponents of open science have argued, though, "The critical barriers to change are not technical or financial; they are social" (Nosek & Bar-Anan, 2012, p. 217).

Most importantly, null results are highly informative and should not be treated as failures. They are the heart of falsificationist hypothesis-testing and modus tollens, the strongest proof science can give us, and they are opportunities to further explicate theories that predicts and subject theories to refutation. It is only when a theory has survived several refutation attempts the theory should be trusted, and *never* when repeated attempts to verify the theory has been made. Even though researchers may be rewarded for discoveries, it is more important that researchers systematically try to falsify hypotheses, otherwise there is no grounds for developing strong theories on a systematic basis.

There are two types of suggestions for improvements with to regards research practices, and indirectly theories. They include pre-registration and clearly separating exploratory and confirmatory analyses (Wagenmakers et al., 2012), and encourage a "replicate and extend" mindset. Taking cognitive science into account during data analysis to understand human biases in interpreting data would also be highly beneficial (Ennser-Jedenastik & Meyer, 2017; Greenland, 2017).

Improvements to the theoretical development include valuing null results as important contributions, and letting theories compete with each other rather than testing one theory at a time. Most importantly, explicate *when* and under *which conditions* the results would be false to make the theory exclude more than it include. This also has the benefit of a "predict and explain" mindset that values predictions in advance (e.g., Lakatos, 1999). This mindset suggests not only future research areas for other researchers, but also specific predictions that can be tested, and is more in line with communication being dependent on, for example, individual characteristics or the type of country and its political system and media system. Because where there are no predictions, there can be no knowledge (de Groot, 1969).

7 References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among italian research psychologists. *PLOS ONE*, *12*(3), e0172792. doi:10.1371/journal.pone.0172792

Allea. (2017). *The European Code of Conduct for Research Integrity*. All European Academies (ALLEA). Retrieved from http://www.allea.org/

Begley, C. G., & Ellis, L. M. (2012). Drug development: Raise standards for preclinical cancer research. *Nature*, *483*(7391), 531–533. doi:10.1038/483531a

Bem, D. (2012). Writing the Empirical Journal Article. In J. M. Darley, M. P. Zanna, & H. L. Roediger III (Eds.), *The Compleat Academic: A Career Guide*. Washington, DC: American Psychological Association.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). Hillsdale, N.J: Routledge.

de Groot, A. D. (1956). The Meaning of "Significance" for Different Types of Research. *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden*. Retrieved from http://www.ejwagenmakers.com/inpress/DeGroot1956_TA.pdf

de Groot, A. D. (1969). *Methodology: Foundations of inference and research in the behavioral sciences*. The Hague: Mouton.

Douven, I. (2016). Abduction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*

(Winter 2016.). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/win2016/entries/abduction/

Ellithorpe, M. E., Ewoldsen, D. R., & Velez, J. A. (2015). Preparation and Analyses of Implicit Attitude Measures: Challenges, Pitfalls, and Recommendations. *Communication Methods and Measures*, *9*(4), 233–252. doi:10.1080/19312458.2015.1096330

Ennser-Jedenastik, L., & Meyer, T. M. (2017). The Impact of Party Cues on Manual Coding of Political Texts. *Political Science Research and Methods*, 1–9. doi:10.1017/psrm.2017.29

Franco, A., Malhotra, N., & Simonovits, G. (2015). Underreporting in Political Science Survey Experiments: Comparing Questionnaires to Published Results. *Political Analysis*, *23*(2), 306–312. doi:10.1093/pan/mpv006

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, *31*(6), 649–679. doi:10.1177/0894439313493979

Gelman, A., & Loken, E. (2013, November). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved from http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, *33*(5), 587–606. doi:10.1016/j.socec.2004.09.033

Gigerenzer, G. (2011). Surrogates for Theory. *APS Observer*, *22*(2).

Greenland, S. (2017). The need for cognitive science in methodology. *American Journal of Epidemiology*. doi:10.1093/aje/kwx259

Hahn, U. (2011). The Problem of Circularity in Evidence, Argument, and Explanation. *Perspectives on Psychological Science*, *6*(2), 172–182. doi:10.1177/1745691611400240

Hamblin, C. L. (1970). *Fallacies*. London: Methuen.

Healy, K. (2017). Fuck Nuance. *Sociological Theory*, *35*(2), 118–127. doi:10.1177/0735275117709046

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

Ioannidis, J. P. A., Stanley, T. D., & Doucouliagos, H. (2017). The Power of Bias in Economics Research. *The Economic Journal*, *127*(605), F236–F265. doi:10.1111/ecoj.12461

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. doi:10.1177/0956797611430953

Kahan, D. M., & Peters, E. (2017). *Rumors of the "Non-Replication" of the "Motivated Numeracy Effect" are Greatly Exaggerated* (SSRN Scholarly Paper No. ID 3026941). Rochester, NY: Social Science Research Network. Retrieved from https://papers.ssrn.

com/abstract=3026941

Kaplan, R. M., & Irvin, V. L. (2015). Likelihood of Null Effects of Large NHLBI Clinical Trials Has Increased over Time. *PLOS ONE*, *10*(8), e0132382. doi:10.1371/journal.pone.0132382

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217. doi:10.1207/s15327957pspr0203_4

Kobayashi, T., Miura, A., & Inamasu, K. (2017). Media Priming Effect: A Preregistered Replication Experiment. *Journal of Experimental Political Science*, 1–14. doi:10.1017/XPS.2017.8

Konijn, E. A., Schoot, R. van de, Winter, S. D., & Ferguson, C. J. (2015). Possible Solution to Publication Bias Through Bayesian Statistics, Including Proper Null Hypothesis Testing. *Communication Methods and Measures*, *9*(4), 280–302. doi:10.1080/19312458.2015.1096332

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*(3), 480–498. doi:10.1037/0033-2909.108.3.480

Lakatos, I. (1999). *The methodology of scientific research programmes*. Cambridge: Cambridge University Press.

Matthes, J., Marquart, F., Naderer, B., Arendt, F., Schmuck, D., & Adam, K. (2015). Questionable Research Practices in Experimental Communication Research: A Systematic Analysis From 1980 to 2013. *Communication Methods and Measures*, *9*(4), 193–207. doi:10.1080/19312458.2015.1096334

Meehl, P. E. (1967). Theory-Testing in Psychology and Physics: A Methodological Paradox. *Philosophy of Science*, *34*(2), 103–115. doi:10.2307/186099

Meehl, P. E. (1990). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*, *66*(1), 195–244. doi:10.2466/pr0.1990.66.1.195

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, *2*(2), 175–220. doi:10.1037/1089-2680.2.2.175

Nosek, B. A., & Bar-Anan, Y. (2012). Scientific Utopia: I. Opening Scientific Communication. *Psychological Inquiry*, *23*(3), 217–243. doi:10.1080/1047840X.2012.692215

Nuijten, M. B., Hartgerink, C. H. J., Assen, M. A. L. M. van, Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985–2013). *Behavior Research Methods*, *48*(4), 1205–1226. doi:10.3758/s13428-015-0664-2

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

Platt, J. R. (1964). Strong Inference. *Science*, *146*(3642), 347–353.

Popper, K. (1992). *The logic of scientific discovery*. London: Routledge.

Richard, F. D., Bond, C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology*, *7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

Roese, N. J., & Vohs, K. D. (2012). Hindsight Bias. *Perspectives on Psychological Science*, *7*(5), 411–426. doi:10.1177/1745691612454303

Seaman, C. S., & Weber, R. (2015). Undisclosed Flexibility in Computing and Reporting Structural Equation Models in Communication Science. *Communication Methods and Measures*, *9*(4), 208–232. doi:10.1080/19312458.2015.1096329

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2001). *Experimental and quasi-experimental designs for generalized causal inference* (2nd ed.). Boston: Houghton Mifflin.

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. doi:10.1177/0956797611417632

Tukey, J. W. (1962). The Future of Data Analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67. doi:10.1214/aoms/1177704711

Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., Vreese, C. de, Matthes, J., … Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the International Communication Association*, *41*(1), 3–27. doi:10.1080/23808985.2017.1288551

van der Zee, T., Anaya, J., & Brown, N. J. L. (2017). Statistical heartburn: An attempt to digest four pizza publications from the Cornell Food and Brand Lab. *BMC Nutrition*, *3*, 54. doi:10.1186/s40795-017-0167-x

Veer, A. E. van 't, & Giner-Sorolla, R. (2016). Pre-registration in social psychology—A discussion and suggested template. *Journal of Experimental Social Psychology*, *67*, 2–12. doi:10.1016/j.jesp.2016.03.004

Vermeulen, I., Beukeboom, C. J., Batenburg, A., Avramiea, A., Stoyanov, D., Velde, B. van de, & Oegema, D. (2015). Blinded by the Light: How a Focus on Statistical "Significance" May Cause p-Value Misreporting and an Excess of p-Values Just Below .05 in Communication Science. *Communication Methods and Measures*, *9*(4), 253–279. doi:10.1080/19312458.2015.1096333

Vickers, J. (2016). The Problem of Induction. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Spring 2016.). Metaphysics Research Lab, Stanford University. Retrieved from https://plato.stanford.edu/archives/spr2016/entries/induction-problem/

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: Retrospective analysis. *BMJ*, *351*, h6467. doi:10.1136/bmj.h6467

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., Maas, H. L. J. van der, & Kievit, R. A. (2012). An Agenda for Purely Confirmatory Research. *Perspectives on Psychological*

Science, *7*(6), 632–638. doi:10.1177/1745691612463078

Walton, D. N. (2008). *Informal Logic: A Pragmatic Approach* (2nd ed.). Cambridge; New York: Cambridge University Press.

Williams, S. A., Terras, M. M., & Warwick, C. (2013). What do people study when they study Twitter? Classifying Twitter related academic papers. *Journal of Documentation*, *69*(3), 384–410. doi:10.1108/JD-03-2012-0027

Yu, C. H. (2006). *Philosophical foundations of quantitative research methodology*. Lanham, Md: University Press of America.

8 Appendix

8.1 R code for Figure 1. Probability of false positives

The number of comparisons refers to the family-wise error rate.

```
library(tidyverse)

data.frame(tests=1:100, probability=(1 - (0.95 ^ (1:100)))) %>%
    ggplot(aes(tests, probability)) +
    geom_line() +
    geom_hline(yintercept=0.05, color="gray", linetype=2) +
    scale_x_continuous(breaks=c(1, seq(10, 100, 10))) +
    scale_y_continuous(breaks=c(seq(0, 1, 0.1), 0.05),
                       labels=c(seq(0, 1, 0.1), 0.05)) +
    labs(title="Probability of at least one false positive",
         x="Number of statistical tests",
         y="Probability") +
    theme_bw() +
    theme(legend.position="bottom", panel.grid=element_blank())
```

8.2 R code for Figure 2. P-values and sample size

The sample size is ranging from 2 to 200 in each group, and the mean and standard deviation is 100 and 1 in the first group, and 100.4 and 1 in the second group (Cohen's d = .4). Additionally, 100,000 Monte Carlo simulations with an independent t-test were performed at *each* sample size to show the average p-values expected over time.

```
library(tidyverse)
library(gridExtra)
set.seed(123456)

# Whether or not the t-test should be carried out on a new independent
# sample (TRUE) or an a dependent sample where a pair of observations
# are repeatedly added to the same sample (FALSE).
IndependentSample <- FALSE

# Increasingly add a pair of observations from a population with
# a given mean difference (i.e., add observations, t-test, repeat).
p.increasingsample <- function(n=200, m1=100, sd1=1, m2=100.4, sd2=1) {
  plist <- list() # List of p-values.
  nlist <- list() # List of sample size.
  dlist <- list() # List of Cohen's d.
  population <- data.frame(x=rnorm(n, m1, sd1), y=rnorm(n, m2, sd2))
  for(i in 2:n)
  {
    if(IndependentSample){
      # Independent sample (new sample every t-test).
      t <- t.test(x=rnorm(i, m1, sd1), y=rnorm(i, m2, sd2))
    } else {
      # Dependent sample (add observation to sample every t-test).
      t <- t.test(x=population$x[1:i], y=population$y[1:i])
    }
    plist[i - 1] <- t$p.value
    nlist[i - 1] <- i + i
    dlist[i - 1] <- (2 * t$statistic) / sqrt(t$parameter[[1]]) # t to d.
  }
  return(data.frame(n=unlist(nlist), p=unlist(plist), d=unlist(dlist)))
}

# Simulate statistical power by t-test.
p.power.simulation <- function(n=200, m1=100, sd1=1, m2=100.4, sd2=1,
                                simulations=1000) {
  plist <- list() # List of p-values.
  nlist <- list() # List of sample size.
  for(i in 2:n) {
    pvals <- replicate(simulations, { t.test(x=rnorm(i, m1, sd1),
                                     y=rnorm(i, m2, sd2))$p.value })
    plist[i - 1] <- mean(pvals)
    nlist[i - 1] <- i + i
  }
  return(data.frame(n = unlist(nlist), p = unlist(plist)))
}

# Create increasing sample size (when true effect exist).
df <- p.increasingsample()

# Perform 100,000 simulations (when true effect exist).
system.time(df.sim <- p.power.simulation(simulations=100000))

# Add the line with simulated power (when true effect exist).
gg <- df %>%
  ggplot(aes(n, p)) +
  geom_line(alpha=0.6) +
  geom_hline(yintercept=0.05, color="gray", linetype=2) +
  geom_line(data=df.sim, aes(n, p), color="red", linetype=1) +
  scale_x_continuous(breaks=seq(0, 400, 25)) +
  scale_y_continuous(breaks=c(seq(0, 1, 0.2), 0.05),
                     labels=c(seq(0, 1, 0.2), 0.05)) +
  labs(title="P-values when the sample size is increasing (true effect)",
       x="Sample size",
       y="P-value") +
  theme_bw() +
  theme(legend.position = "bottom", panel.grid = element_blank())

# Create increasing sample size (when no effect exist).
df.nill <- p.increasingsample(n=200, m1=100, sd1=1, m2=100, sd2=1)

# Perform 100,000 simulations (when no effect exist).
system.time(df.nill.sim <- p.power.simulation(n=200, m1=100,
            sd1=1, m2=100, sd2=1, simulations=100000))

# Add the line with simulated power (when no effect exist).
gg.nill <- df.nill %>%
  ggplot(aes(n, p)) +
  geom_line(alpha=0.6) +
  geom_hline(yintercept=0.05, color="gray", linetype=2) +
  geom_line(data=df.nill.sim, aes(n, p), color="red", linetype=1) +
  scale_x_continuous(breaks=seq(0, 400, 25)) +
  scale_y_continuous(breaks=c(seq(0, 1, 0.2), 0.05),
                     labels=c(seq(0, 1, 0.2), 0.05)) +
  labs(title="P-values when the sample size is increasing (no effect)",
       x="Sample size",
       y="P-value") +
  theme_bw() +
  theme(legend.position = "bottom", panel.grid = element_blank())

grid.arrange(gg, gg.nill)
```

8.3 Method for Figure 3. Underestimated p-values

A total of 393 statistics was extracted from three years of empirical research articles (n = 202) in two journals:

- Journal of Communication, 2015 issue 1 to 2017 issue 4 (n = 119)
- Political Communication, 2015 issue 1 to 2017 issue 3 (n = 83)

All articles were manually downloaded as PDF files, and the R package `statcheck` (Nuijten, Hartgerink, Assen, Epskamp, & Wicherts, 2016) were used to automatically estimate the prevalence of misreported p-values.

The package can only analyze statistics written in APA format, and will therefore miss statistics reported in an unconventional way, as well as regression coefficients and the like. The formats that statcheck can analyze are:

- `t(df) = value, p = value`
- `F(df1,df2) = value, p = value`
- `r(df) = value, p = value`
- `[chi]2 (df, N = value) = value, p = value` (N is optional, delta G is also included)
- `Z = value, p = value`