Abstract

In 2005, Ioannidis revitalised a long-lasting debate concerning whether or not null hypothesis significance testing may serve to uphold accessibility related biases, and in effect the file drawer problem as well. Though several authors have argued that significance testing may consequentially impede further progress within the medical and behavioural sciences, it has had little impact on the general consensus, and significance testing is still the most common approach to analyze and interpret data within psychology. In this thesis it is argued that as the philosophical foundation of the argument has never been extensively analysed, reluctance to accept the proposition is understandable, even justified, until sufficient support may decide its validity. In response to these unresolved philosophical and methodological issues, this thesis argues that continuous problem-solving is the driving force of scientific progress, and that problem-solving is contingent upon pluralism. To ensure pluralism, a representative sample of all experimental findings must necessarily be publicly available. However, several well-documented accessibility related biases obstruct theoretical and empirical diversity, and may accordingly inhibit further progress within psychology. In support of Ioannidis' claim, it is argued that null hypothesis significance testing is the underlying cause of the current accessibility related biases. In addition, this methodological dogma aids several conjunctive fallacies, such as *p*-value misinterpretations, cultivation of dichotomous thinking, and the problem of practical significance. Taken as a whole, these premises thoroughly emphasize perhaps the most serious challenge towards further progress within psychology. Lastly, some methodological substitutes are discussed which might liberate the discipline from the current accessibility related biases, encourage further progress, and promote newfound credibility in psychology.

Dedications

First I would like to say thanks to my wife Hanne Horndalen Tveit for enduring a rather absent-minded husband for almost a year or so, while still supporting me all the way through. My two best friends at NTNU, Tor Erik Hellum and Nathalie Sturim, are the reason why the two last years has been superb. Our discussions have been valuable, our friendship invaluable. Professor Hermundur Sigmundsson, thank you for the academic advices and your interest for my research topic throughout the whole process. And for having proofread my thesis on short notice, Alex Chapman, I owe you greatly.

Many people have helped me make sense of different aspects relevant for my thesis: Tone Kvernbekk helped me gain a broader perspective on the philosophy of science, without Robert Biegler I may never have understood how fragile p-values are, and I have greatly benefited from more general discussion on the research topic with Tor Erik Hellum, Heming Strømholt Bremnes, Mathilde Lien, Kristina Mersland, and Dag Jomar Mersland. Thank you!

Finally, to my family, thank you for all the questions you've asked, for all those times you've patiently listened for my response, and for always encouraging me to do my best. You mean the world to me (with no reference to the Toni Braxton song with the same title).

None are held responsible for the views, the arguments, and the potential flaws of this thesis except me. Unfortunately. I wish I could blame someone else.

Trondheim, September 2014

Håvard Tveit

Analytical index

The analytical index comprises a sketch of the main argument, intended to provide coherence and comprehensibility for the readers throughout the thesis. Each chapter is accompanied with a summarizing statement and a brief synopsis.

Chapter 1: On the critical attitude

”The critical attitude is characterized by readiness to change ones convictions”

Questioning dogma, accepting the fallible nature of science, and a readiness to revision, briefly summarizes Karl Popper’s philosophical heritage. These virtues conjure scientists to analyze and challenge influential theories and methodologies in need of critical revision. The goal of this thesis is to employ Popper’s critical attitude towards perhaps the most persistent dogma within psychology today; the null hypothesis significance testing.

Chapter 2: On the nature of scientific progress

”Problem-solving is the driving force of scientific progress,

and problem-solving is contingent upon pluralism”

In response to the commonly stated proposition that null hypothesis significance testing may impede further progress within the behavioural sciences, this thesis aims at analyzing the theoretical foundations of this claim. Throughout the most influential theories within the philosophy of science, continuous problem-solving is presumed the driving force of progress, and problem-solving is furthermore contingent upon pluralism.

Chapter 3: On the prevalence of accessibility related biases

”A representative sample of all experimental findings must be publicly accessible”

As access to a representative sample of all experimental findings is pivotal for pluralism, any accessibility related biases might impede the progressive forces of psychology. By examining the literature, the prevalence of accessibility related biases is overwhelming, potentially obstructing theoretical and empirical diversity, and ultimately restraining scientific progress.

Chapter 4: On the fallacies of significance testing

”Significance testing is the underlying cause of accessibility related biases”

Null hypothesis significance testing is the most common approach to analyze and interpret data in psychology, yet also the most criticised. Ioannidis’ claim that significance testing may be the underlying cause of the current accessibility related biases is perhaps the most severe, but several conjunctive fallacies are found to exist. In particular, that *p*-values are commonly misinterpreted, that it cultivates dichotomous thinking of nuanced data, and that statistical significance has little to say with practical significance.

Chapter 5: On the future of psychology

”Change is of the essence”

By carrying out an extensive analysis on the nature of scientific progress, a thorough review on the prevalence of accessibility related biases, and lastly a comprehensive discussion on the fallacies of significance testing, this thesis argues in favour of the stated proposition that null hypothesis significance testing may indeed impede further progress within psychology. Finally, some potential methodological substitutes are discussed which might liberate the discipline from the current accessibility related biases, encourage further progress, and promote newfound credibility in psychology.

Table of Contents

1. On the critical attitude

"The critical attitude is characterized by readiness to change ones convictions"

1.1. Introduction

Karl Popper (1965) considered the critical attitude the most paramount virtue a scientist could endorse. The critical attitude may be identified as the scientific attitude, and it is characterized by questioning dogma, accepting the fallible nature of science, and a readiness to change ones convictions when confronted by contradicting evidence (ibid., p. 50). Popper's philosophical legacy conjures scientists to analyze and challenge influential theories and methodologies in need of critical revision (ibid., p. 56).

Pragmatically preserving ones most cherished ideas is a sign of the dogmatic attitude (ibid., p 49), distinguished from the critical attitude, by neglecting refutations even when defeat should be accepted. In the works of Popper, the scientific tradition constitute a culture of justified critique, striving towards improving on the inevitably fallible conjectures passed on from previous generations (ibid., p. 50). This thesis will apply this critical attitude towards perhaps the most persistent dogma in psychology today; the null hypothesis significance test.

1.2. Outline

1.2.1. Revitalising an old dogma

In 2005, Ioannidis published a paper in which he argued for "why most published research findings are false". According to Ioannidis, the biggest problem threatening the medical and behavioural sciences today are accessibility related biases (2005, p. 696). In the current thesis, accessibility related biases refers to the group of biases that inhibits a representative sample of all experimental findings being publicly accessible for professionals within any scientific field of study. In this group, several distinct biases may be identified: "positive results bias", which "occurs when authors are more likely to submit, or editors accept, positive than null results" (Song et al., 2010, p. 3); "selective reporting bias", occurring "when studies with multiple outcomes report only some of the outcomes" (Song et al., 2010, p. 21); "non-replication bias", which occurs when authors are less likely to submit, or editors accept, replication studies independent of study results (Cumming, 2014, p. 4); as well as several less influential biases.

Perhaps even more interesting, Ioannidis further argues that these biases seem to be a consequence of "claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance" (2005, p. 696). The concern that null hypothesis significance testing may distort the scientific literature is not new. Already in the sixties had this topic been discussed in the psychological literature (Bakan, 1966; Smart, 1964; Sterling, 1959), more famously coined the "file drawer problem" by Rosenthal (1979). By adopting a significance level of .05, as common within most of the behavioural sciences (Cohen, 1994, p. 999), one out of twenty studies will yield positive results even in the case of no actual effect. Rosenthal argued that there is a theoretical possibility that the "journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > .05$) results" (1979, p. 638).

Even if the situation is not as catastrophic as Rosenthal predicted, several prominent authors (Cumming, 2012, 2014; Kirk, 2003; Kline, 2004; Meehl, 1967, 1978) have argued that as null hypothesis significance testing might serve to uphold accessibility related biases, the current approach of analyzing and interpreting data may actually impede further progress within the behavioural sciences. Yet, significance testing persists in the current literature, and researchers in psychology are seemingly reluctant to accept the argument. Previous work on the topic has thoroughly explained several conjunctive fallacies of the significance test, for instance that *p*-values are commonly misinterpreted (Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Kirk, 2003; Nickerson, 2000), that it cultivates dichotomous thinking of nuanced data (Cumming, 2014; Rosenthal & Gaito, 1963, 1964), and the problem of practical significance (Bakan, 1966; Cohen, 1962; Johnson, 1999; Thompson, 1998; Tukey, 1991). It is thus clear that this methodological dogma is in need of critical revision, but it is not necessarily evident that significance testing may impede scientific progress.

This thesis argues that this reluctance to abandon reliance in significance testing is understandable, even justified, in the absence of an extensive analysis on the unresolved inferential propositions constituting the argument. Therefore, the aim in the current thesis is to provide some initial indications concerning the truth-values of the individual propositions of the argument, postulating a preliminary conclusion to the argument in the hope that it will contribute to further discussion on the topic.

1.2.2. Constructing a valid argument

In order to properly evaluate the argument stated previously, and thus to enable asserting the validity of the conclusion, the argument has in this thesis been constructed by the pattern of a hypothetical syllogism[1], facilitating the examination of each proposition individually:

(i) If *significance testing,* then *accessibility related biases* (if *p,* then *q*)
(ii) If *accessibility related biases,* then *scientific stagnation* (*if q,* then *r*)
(iii) *Significance testing* (*p*)
(iv) ∴ *scientific stagnation* (*r*)

Previous work on the topic has typically ignored the two inferential propositions [(i) and (ii)], and rather stated the flaws of significance testing and jumped to the conclusion (iv). Ioannidis one to other hand has thoroughly explained the first inferential proposition (i), but neglected the second (ii). By disregarding these inferential propositions, the argument is fundamentally flawed, and the premises do not entail the conclusion. Therefore, it is clearly understandable, even justified, that the argument has had little impact on the general consensus, and that researchers are reluctant to accept the conclusion.

In order to properly assert the validity of the entire argument, the truth-value of each inferential proposition must necessarily be accounted for [(if *p*, then *q*) and (if *q* then *r*)], as well as the declarative sentence (iii). Only by affirming each proposition may the premises necessarily entail the conclusion. Until these unresolved philosophical and methodological has been thoroughly accounted for, the argument is not valid. In the next section, several research questions are defined with the aim of asserting the validity of the argument.

1.3. Objectives

In light of the recent discussion on the inadequacies and potentially devastating corollaries of null hypothesis significance testing, the main objective throughout this thesis is to explore the unexamined conclusion at the heart of the argument, more specifically formulated in (1):

(1) *Can null hypothesis significance testing ensure further progress within psychology?*

As seen in the construction of the argument in the previous section, research question (1) depends on several inferential propositions as well as one declarative statement. Thus, in the

[1] Rules of inference by *hypothetical syllogism* ($p \rightarrow q, q \rightarrow r, p, \vdash r$) (Tomassi, 1999, p. 113)

extension of the main research question, some underlying questions must be examined:

(2) *Which principle(s) drive scientific progress forward?*
(3) *What may impede these principle(s) of being fulfilled?*

To provide an appropriate response to the former of these two questions, the most influential theories on scientific progression within the philosophy of science will be extensively covered in the following chapter. Regarding the latter research question (3), Ioannidis' assumption concerning accessibility related biases will be examined in an attempt to determine the effect of these biases within the realm of psychology in chapter 3. Furthermore, research question (4), dealt with in chapter 4, will help resolve whether accessibility related biases is indeed being upheld by null hypothesis significance testing or not:

(4) *Are the aforementioned principle(s) fulfilled in the psychological sciences?*

An extensive analysis of the three latter research questions (2) - (4), should provide sufficient information to determine to the validity of the argument, and thus the outcome of research question (1).

1.4. Organisation

The thesis is organized as follows.

Chapter 2 reviews the most influential theoretical models accounting for scientific progress from the philosophy of science. The main focus will be to identify the driving forces of progress, and which, if any, premise(s) must be met to ensure further progress within a scientific discipline.

In chapter 3, an examination of the file drawer problem and the accessibility related biases will be provided. Support of the existence of such biases within the scientific literature is accounted for, as well as the most common method of handling biases in meta-analyses.

Chapter 4 will introduce the null hypothesis significance test by investigating the basic assumptions and the historical development of the approach. Following is a thorough review of its alleged connection with accessibility related biases, before turning to the more general criticism.

Concluding this thesis is chapter 5, which briefly summarize the main arguments of the thesis and show how the research questions postulated in section 1.3 has been accounted for. Finally, some thoughts concerning the future of the psychological sciences are discussed.

2. On the nature of scientific progress

"Problem-solving is the driving force of scientific progress,

and problem-solving is contingent upon pluralism"

2.1. Introduction

This chapter will focus on the philosophy of science by reviewing the most influential theories on the nature of scientific progress. In particular, the goal is to identify the underlying principle(s) that drives progress forward, as stated previously in research question (2).

Several authors have argued that null hypothesis significance testing might essentially impede further progress within psychology (Cumming, 2012, 2014; Kirk, 2003; Kline, 2004; Meehl, 1967, 1978), yet with little impact on the general consensus. In the previous chapter it was argued that this may be due to the lack of any theoretical analysis of the topic, and this chapter will thus attempt to provide such an extensive analysis.

The chapter is organized as follows. First Toulmin's (1963) basic criteria for which any model on scientific progress must accommodate will be discussed in section 2.2.1. Popper's (1963) proposal of a new methodological practice is discussed in section 2.2.2. In section 2.2.3, Duhem's (1954) theses are presented, followed by a review of Kuhn's (1970) paradigms (section 2.2.4). In section 2.2.5, Lakatos' (1992) rational research programmes are discussed, before introducing the parsimonious principles offered by Laudan (1996) and Feyerabend (1993) in section 2.2.6. Concluding the chapter is section 2.3.

2.2. Theories on scientific progress

2.2.1. Evolution of the sciences

Stephen E. Toulmin (1963) was one of the first theorists that attempted to discover some general rules driving scientific progress forward. He proposed that there must be some principles that separate successful science from unsuccessful science, and that these principles in turn has to be able to explain the enormous success of the scientific enquiry (ibid., p. 14).

In order to provide an adequate model for scientific progress, Toulmin (1963, p. 15-16) proposed that three criteria had to be met: any satisfactory model need to acknowledge the complexity and diversity inhered in the scientific nature; it has to reflect in large part the

actual history of science; and, it must be parsimonious[2]. Perhaps inspired by the success and acknowledgment of Darwin's theory of natural selection, Toulmin claimed that «science as a whole - the activity, its aims, its methods and ideas - evolves by variation and selection» (ibid., p. 17).

Throughout the work of Toulmin, the metaphors are quite evidently linked to the theory of evolution. He claimed that new scientific theories arise spontaneously (1963, p. 111); usually in some form closely related to previous theories (ibid., p. 111); he proposed that theories became increasingly complex over time (ibid., p. 112); and he attempted to show how certain theories that seemed insufficient in its predictive powers at one time in history could be the theory with the largest predictive powers at a later time (ibid., p. 113).

In judging the explanatory power of different theories, Toulmin relied on inductive confirmation (1963, p. 112). Briefly explained, inductive confirmation infer general laws based on a finite set of observations (Copi, 1961, p. 361): if the investigator discovers that the first observed swan is white, and this characteristic is consistent after observing several swans, the investigator may infer that all swans are white[3]. Inductive confirmation implies that scientific theories become more reliable as the accumulation of evidence in its favour increases.

In short, Toulmin argued that science evolves by variation and selection. For every new generation of scientists, new theories and methodologies will become dogma while older theories and methodologies will be abandoned, giving rise to a slow, but progressive science (1963, p. 110). Toulmin viewed scientific progress as a survival of the fittest, building on previous accomplishments, though changing in character over time. The three criteria he set for an adequate model of scientific progression is in a sense intertwined: the actual history of science *is* the history of complexity and diversity, and it is this history that enables Toulmin's parsimonious principle of natural selection of theories over time to take place.

[2] The principle of parsimony, also known as Ockham's razor, states that "an explanation of the facts should be no more complicated than necessary" (Jefferys & Berger, 1992, p. 64). In dealing with competing theoretical ideas, one should prefer the theory with the least complexity.

[3] Rules of inductive confirmation: x is F and G, y is F and G, z is F and G, hence, all Fs are Gs (Copi, 1961, p. 361). In the example above: x, y, and z refers to instances observed, F indicate swan, and G indicate white.

2.2.2. From problems to problems

Karl R. Popper, by some called the "last of the great logicians" (Harre, 1994), found the reliance on inductive confirmation[4] by his contemporaries highly unsatisfactory (Popper, 1968, p. 27). As Hume (1912, p. 23) had already shown the paradoxical nature of inductive confirmation by arguing that inductive logic go beyond provable demonstration[5], Popper attempted to provide the scientific enterprise with a rational that could establish absolute certainty, which induction arguably cannot. Hume stated that as inductive confirmation rely on experience, "all our experimental conclusions proceed upon the supposition that the future will be conformable to the past" (ibid, p. 35). However, any justification of such a relationship must evidently be based on inductive confirmation, which would beg the question (Russell, 1959, p. 68).

Referring approvingly to Hume's argumentation, Popper (1965, p. 45) argued that the logic of science must rely on deductive refutation, "as only the falsity of the theory can be inferred form empirical evidence, and this inference is a purely deductive one" (ibid., p. 55). In order to fully understand Popper emphasis we need to take a brief look at propositional logic. In the case of confirmation versus refutation, both can be constructed as conditional statements, following the pattern of *if*... *then*... (Copi, 1961, p. 224). Regarding scientific testing, the "if-statement" is usually characterized by a specific hypothesis, and the "then-statement" contains the prediction or observation, hence, the goal of the test is to acquire proof of the "then-statement".

The problem with confirming a statement[6], as Popper (1965, p. 53) pointed out, is that it cannot establish absolute certainty, as it allows for multiple alternative hypotheses being confirmed by the same prediction. The hypotheses "if *the soda contains sugar*" and "if *the soda contains aspartame*", may both be supported by the observation that "*the soda tastes sweet*". Thus reasoning from confirming evidence is necessarily incomplete. By refuting a statement[7], however, one can ascertain absolute certainty, as a single instance of refutation

[4] The positivist movement relied heavily on induction (Delanty & Strydom, 2010, p. 15), reflected in numerous books and articles. Schlick (1979, ch. 3) stated this quite directly in his works, and is a good representative for the movement as such.

[5] Hume's distinction between knowledge provable by demonstration and knowledge that go beyond provable demonstration is comparable to the analytic/synthetic distinction. For the interested reader, Ayer (1946) might be a good place to start due to accessible language. Further reading might include Kant (1899), Frege (1950), and Carnap (1956), as they all endorse the distinction, while Quine (1951) might be read to counterbalance.

[6] Rules of inference for deductive confirmation by *affirming the consequent* ($p \rightarrow q$, q, $\vdash p$) (Tomassi, 1999, p. 75-76).

[7] Rules of inference for deductive refutation by *modus tollens* ($p \rightarrow q$, $\neg q$, $\vdash \neg p$) (Tomassi, 1999, p. 75-76).

can validly disconfirm the hypothesis tested (ibid, p. 55). If the prediction is that the soda tastes sweet, but it doesn't taste sweet at all, one can confidently reject the hypothesis "*the soda contains sugar*", and conclude "*the soda does not contain sugar*".

Applying falsifiability[8] as a rational for ascertaining absolute certainty implies that science progress not by verifying already accepted theories, as Toulmin and Schlick argued, rather by refuting false theories. Thus the aim for the investigator is to deduce a singular contingent[9] hypothesis from the theory in question, and expose the extracted hypothesis to a crucial experiment (Popper, 1965, p. 36). If the experiment succeed to refute the experimental hypothesis, the theory must necessarily be false, while if the experiment fails to refute the hypothesis, the theory is tentatively accepted.

The means by judging competing, tentatively accepted, theories according to Popper is by the falsifiability of a theory; theories "more exposed to refutation" (1965, p. 32) might be viewed as stronger theories. That is not to say they are more likely to be true than other theories, which would imply reliance on inductive principles[10], rather it indicates that highly corroborated theories must be easier to refute and thus more impressive if surviving several rigid attempts at refutation. The implications is that the greater the content of the theory, the more falsifiable it is, and thus, the lower the probability of it being true (ibid., p. 217). Popper provides an excellent example of this relationship: if hypothesis *a* is "it will rain on Saturday", hypothesis *b* "it will rain on Sunday", and hypothesis *ab* "it will rain on Saturday and Sunday", it is clear that the hypothesis *ab* contains more content, a lower probability of being true, and thus, if not refuted, indicates a stronger theory (ibid., p. 218)[11].

In terms of scientific progress, Popper argued that the aim of science, and accordingly the measure of scientific progress, is to produce theories with higher informative content and hence lower probability of being true than previously accepted theories (Popper, 1965, p.

[8] Falsifiability may be described as the innate possibility of a hypothesis to be disproven by means of deductive refutation (Popper, 1968, p. 86-87). It does not imply that the hypothesis *is* indeed false, rather it implies that a falsifiable hypothesis has certain restrictions, which, in the event of them occurring, may refute the hypothesis.

[9] Any deduced hypotheses must necessarily be contingent with the theory according to Popper (1965, p. 36). Consequently, if the hypothesis is refuted, the theory must be refuted as well, as any theory that yields false conclusions must indubitably be false as well.

[10] Salmon (1967) argued that "modus tollens without corroboration is empty; modus tollens with corroborations is induction" (ibid., p. 26), implying that Popper's principle of falsification relies on inductive inference. This would have been true if Popper claimed that highly corroborated theories are more likely to be true, but that is not what he postulated. Popper simply postulated that highly corroborated theories should be preferred, as they must have been scrutinized more severely (1965, p. 217).

[11] Let *C* denote *content*, and *p* denote *probability*, then Popper's idea could be presented as: $C(a) < C(ab) > C(b)$, implying that the conjunction of *ab* provides more content than either *a* or *b*, while $p(a) > p(ab) < C(b)$, illustrate that the conjunction of *ab* provides lower probability than either *a* or *b* (Popper, 1965, p. 218).

219). Furthermore, as the goal of the investigator is attempting falsification, theories must necessarily arise from problems with previously accepted theories. "The most lasting contribution to the growth of scientific knowledge… are the new problems which it raises, so that we are led back to the view of science and of the growth of knowledge as always starting from, and always ending with, problems" (ibid., p. 222). For Popper, then, science is a continuous sequence of bold conjectures and disappointing refutations, "a method of eliminating false theories" (ibid., p. 56), starting from, and ending with, problems.

2.2.3. A brief intermission

In his highly influential book *The Aim and Structure of Physical Theory*, Pierre M. M. Duhem (1954, p. 184) argued that science has come to a point in which theoretical system has become so intricate and interdependent, that any single hypothesis will always rely on several, often tacit, hypotheses (Laudan, 1965, p. 296)[12]. Provided that this is true, Popper's principle of falsification would be seemingly unattainable. Duhem's principle, called the separability thesis, inhibits the researcher to single out any particular hypothesis from the theoretical system from which it is deduced (Duhem, 1954, p. 187).

An obvious corollary of the separability thesis is that no hypothesis can be conclusively refuted by observation alone, termed the falsifiability thesis. If the observation contradicts the hypothesis, the only conclusion to be drawn "is that at least one of the hypotheses constituting this group is unacceptable and ought to be modified; but the experiment does not designate which one should be changed" (Duhem, 1954, p. 187). This holistic perspective of science might not hold true for all scientific experiments, but when a particular hypothesis "requires other interpretative statements in order to confront experience... then observation is not decisive by itself" (Ariew, 1984, p. 321).

Williard V. O. Quine (1951, p. 39) takes Duhem's theses to supports the notion of underdetermination, the idea that "there are in principle always an indefinite number of theories that fit the observed facts more or less adequately" (Hesse, 1980, p. viii). As there is

[12] Lakatos (1992) provides an excellent example on the interconnectivity of hypotheses:

> Galileo claimed that he could 'observe' mountains on the moon and spots on the sun and that these 'observations' refuted the time-honoured theory that celestial bodies are faultless crystal balls. But his 'observations' were not 'observational' in the sense of being observed by the – unaided – senses: their reliability depended on the reliability of his telescope – and of the optical theory of the telescope – which was violently questioned by his contemporaries (p. 14-15).

no way of separating hypotheses from one another, ad hoc hypotheses[13] will be of equal value as the initial hypothesis. For Quine, "any statement can be held true come what may, if we make drastic enough adjustments elsewhere in the system" (1951, p. 40). Ariew (1984, p. 315) have argued that Quine's underdetermination is historically false. In fact, history shows that scientists often have difficulties finding as much as one good theory in support of the seemingly unexplainable observation according to Ariew. Underdetermination, then, should rather be viewed as the "inability of empirical evidence to conclusively confirm or disconfirm a given theory" (ibid, p. 316).

In defending Popper's falsification, Adolf Grünbaum (1976a, p. 116) argued that the duhemian logic are fundamentally flawed, as it cannot be proven that all experiments indeed rely on auxiliary hypotheses. In fact, Duhem has to prove, according to Grünbaum (ibid., p. 118), that for every set of observations, there are some auxiliary hypotheses in conjunction with the deduced hypothesis, or his theses does not hold. Laudan, in turn, claimed that Grünbaum has misinterpreted Duhem's initial claim (Laudan, 1965, p. 296), as "Duhem is not asserting that every hypothesis can be saved, but only that unless one had proved that it cannot be saved, then it is not falsified" (ibid., p. 297). Thus, if one is to claim that a scientific hypothesis is refuted, the investigator who claims so has to bear the burden of proof (i.e., that no auxiliary hypotheses may save the refuted hypothesis) (Laudan, 1965, p. 298)[14].

2.2.4. Puzzle-solving

Following the theoretical tradition of Duhem and Quine, Thomas S. Kuhn argued that the history of science consists of numerous paradigms[15] (Kuhn, 1970, p. 10). For the purpose of this thesis, it should be sufficient to consider a paradigm as comprising of a set of indisputable theoretical assumptions, fundamental "puzzle-solving" strategies, and standards for scientific practice (ibid., p. 10-11). These standards of the paradigm are usually derived from one or more particularly successful theories, which at the time of inception were scientifically exceptional, while still "sufficiently open-ended to leave all sorts of problems for the…

[13] Ad hoc hypotheses may be thought of as theory modifications (Grünbaum, 1976b, p. 329). If an observation contradicts the expected outcome of an experiment, one may, instead of discarding the theory, add a second hypothesis that would modify the theory sufficiently to deal with the unexpected anomaly.

[14] Popper (1968, p. 40-42) argued that even though it might be possible to save any hypothesis by making adjustments elsewhere in the system, "the empirical method shall be characterized as a method that excludes precisely those ways of evading falsification" (ibid., p. 42).

[15] Kuhnian hermeneutics are rather ambiguous in Kuhn's original work, but for those interested; Hoyningen-Huene (1993) goes into great detail to explain the kuhnian terms in an orderly fashion.

practitioners to resolve" (ibid., p. 10). In broad terms, it might be said that every paradigm goes through certain phases from its beginning to its demise: normal science, characterized by solving the left-over puzzles; the rise of anomalies, marked by a growing amount of unsolvable puzzles; and revolutionary sciences, defined as a period of distrust in the current paradigm, inevitably leading to an overthrow and replacement of the paradigm.

As briefly mentioned, paradigms are commonly defined by successful theories and a promise of further growth of knowledge. Indeed, the main task in the normal sciences is to apply the fundamental puzzle-solving strategies in order to explore and expand the current paradigm further (Kuhn, 1970, p. 52). Kuhn argued that the normal sciences implicitly directs the scientific inquiry towards three classes of problems; "determination of significant fact, matching facts with theory, and articulation of theory" (ibid., p. 34). The first class of problems is concerned with applying the newfound knowledge of the paradigm for technological advancements (ibid., p. 26), e.g., using x-rays to improve the health technology. The second class of problems aim "to demonstrate agreement" (ibid., p. 27) between the theoretical foundations and the factual nature of the world. The third class of problems is resolving issues not yet dealt with, thus expanding the scope of the paradigm further (ibid., p. 27-28). As the paradigm decide the standards of scientific practice, it will encourage the investigators to seek puzzles the paradigm may provide adequate solutions to (ibid., p 39)[16].

In the quest of solving puzzles, every paradigm will experience certain observations contradicting its theoretical assumptions. Anomalies are bound to arise, but as the paradigm restricts the researchers theoretical perspective, the scientific community will usually consider such irregularities as acceptable (Kuhn, 1970, p. 52), and, in adhering to Quine's proposal, stipulate appropriate ad hoc solutions to deal with the anomalies (ibid., p. 53). Kuhn postulated that as anomalies goes against the expectations of the scientific community, it is fairly easy to understand why anomalies are suppressed and assumed to be inadequacies on the researchers side rather than the paradigm (ibid., p. 64). Nevertheless, as a growing number of anomalies arise, they will become increasingly difficult to disregard, and, over time, the anticipation of the researchers will gradually change (ibid., p. 64). At the point were the researchers expect to observe anomalies, rather than theoretically correct predictions, the demise of the current paradigm is inevitable and the time of revolutionary science brewing.

[16] This will also work to restrict the investigator from directing modus tollens towards the current paradigm (Kuhn, 1970, p. 40), and thus ensure stability within the paradigm.

The shift in confidence, from assuming the paradigm will provide satisfying predictions and explanations, provoke uncertainty among the members of the paradigm (Kuhn, 1970, p. 68). Kuhn argued that only when the scientific community acknowledge the shortcomings of the current paradigm, may novel theories emerge in order to solve the arisen anomalies (ibid., p. 75). Confirming to Quine's underdetermination, Kuhn emphasized that several novel theories would emerge simultaneously (ibid., p. 77), and the choice of the coming paradigm would be based on relativistic principles[17] rather than rational ones. By appealing to Duhem's theses, Kuhn argued that as observations cannot be understood separate from the theoretical background (ibid., p. 80)[18], it is impossible to compare competing theories by any rational principles (ibid., p. 111-112). Accordingly, paradigms are not chosen due to best fit: "a decision of that kind can only be made on faith" (ibid., p. 152).

In the works of Kuhn, progress will occur mainly in the normal sciences, characterized by puzzle-solving within the restricted framework set forth by the paradigm. As Kuhn maintain a relativistic approach in explaining the mechanisms of revolutionary science and shift between paradigms, progress through revolutions cannot assure that science comes closer to an objective truth, an idea shared with Popper (1965, p. 223-240). Kuhn, as Toumin, refers to the progress of science in terms of evolutionary processes. There are no teleological reasons for scientific change, as there is no goal for evolution (Kuhn, 1970, p. 172-173), there is only change, and change does not necessarily imply progress.

2.2.5. Protective modifications

For Imre Lakatos (1992), neither Kuhn nor Popper provided an accurate model of the actual history of scientific progress, though both postulated important ideas for further development of a model. Popper's methodological falsification had been challenged by Duhem's appeal to the interconnectivity of auxiliary hypotheses. Moreover, Kuhn had convincingly argued that researcher's rarely abandon their theories after a single instance of contrary evidence. Lakatos agreed with both Duhem and Kuhn. However, he couldn't except that science was irrational.

In the same manner as Duhem claimed that single *hypotheses* couldn't be understood in a vacuum, Lakatos began by arguing that single *theories* couldn't be understood in a vacuum (Lakatos, 1992, p. 32). Kuhn's view that only one paradigm existed at a time was

[17] A great response to Kuhn's relativistic perspective can be found in Laudan (1996, ch. 5).
[18] Hanson (1958, ch. 1) goes into great detail on the theory-ladenness in all observations for the interested reader.

historically false according to Lakatos. For instance, both the geocentric and heliocentric paradigm existed simultaneously until the heliocentric model superseded the former due to its explanatory powers (Lakatos & Zahar, 1975).

According to Lakatos, the history of science revealed several *research programmes* (Lakatos, 1992, p. 47), resembling that of Kuhn's paradigms, but with one major exception; several programmes coexisted at the same time. A research programme would consist of a *hard core* of indisputable facts, as well as a *protective belt* of auxiliary hypotheses (ibid., p 48). Similarly to Kuhn's paradigms, the researchers invested in the programme would adhere to certain rules: "some tell us what paths of research to avoid (*negative heuristics*), and others what paths to pursue (*positive heuristics*)" (ibid., p. 47, italics in original). The negative heuristics are the injunction to direct modus tollens towards the hard core of the programme (ibid., p. 48), while the positive heuristics serves to modify the protective belt in order to comply with observations opposing the hard core (ibid., p. 50).

A research programme, then, is in continuous change throughout its history, an idea Popper had argued in favour of previously. However, Lakatos' protective belt served another important feature that Popper did not acknowledge; it helped the researcher protect their most valuable ideas by adding auxiliary hypotheses (Lakatos, 1992, p. 50). Yet, not all alterations of the protective belt are equal. Following an alteration, research programmes can be viewed as either progressive or degenerative, depending on the outcome of the modification (ibid., p 70-72). Lakatos holds that if an alteration not only resolve the occurring anomaly, but also leads to novel predictions[19], then a programme ought to be perceived as progressive (ibid., p. 80). If, on the other hand, the alteration provides nothing more than ad hoc adjustments, it should be viewed as degenerative. The competing research programmes may be judged by their heuristic powers, and the abandonment of one programme in favour of a competing one is thus based on rational principles rather than relativistic ones (ibid., p. 19)

Lakatos' research programmes involve several features worth discussing. First, it challenges the paradigmatic nature of science set forth by Kuhn, by claiming that several programmes operate at the same time. Secondly, it brings back rationality to science by appealing to heuristic powers, instead of social psychological factors. Most importantly perhaps, is that scientific progression is viewed as a continuum, not only within the research programme, but also throughout the history of science. Progression within a programme

[19] The term novel prediction has two meanings according to Lakatos and Zahar (1975, p. 369); it can either be the correct prediction of a phenomena not yet observed, or it can be the correct prediction of a phenomenon that *is* observed, but which it was not initially intended to solve.

involves modifying the protective belt to solve the problems brought on by observations challenging the hard core, while progression outside of the programme involves pluralism by contrasting competing programmes.

2.2.6. Problem-solving and pluralism

Larry Laudan (1996, p. 78) began with a simple question: what is the aim of science? Without specifying a goal of the scientific enterprise, Laudan argued, one cannot judge whether or not science is progressing. Setting traditions aside, "the aim of science is to secure theories with a high problem-solving effectiveness" (ibid., p. 78). And furthermore, "science progresses just in case successive theories solve more problems than their predecessors" (ibid., p. 78).

From Laudan's perspective, two categories of problems may be identified: *empirical* and *conceptual* ones. Empirical problems can be further distinguished by *potential problems*, *solved problems*, and *anomalous problems* (Laudan, 1996, p. 79). Potential and solved problems account for what is conceived as unresolved and solved questions respectively. Anomalous problems can be described as problems that rival theories have resolved, and pose a direct threat to the particular research tradition. Conceptual problems are general theoretical problems: internal inconsistencies, assumptions not yet supported by empirical observations, failing to conform to more general theories, and so on (Laudan, 1996, p. 78).

It is in the weighing of the two categories of problems Laudan separates himself from his empirically focused predecessors; by claiming "the elimination of conceptual difficulties is as much constitutive of progress as increasing empirical support" (Laudan, 1996, p. 80). By indicating that theoretical problems are as important as empirical problems, Laudan questions a fairly rigid assumption held within the philosophy of science and opens up for alternatives. His alternative; it is all about continuous problem-solving, whether empirical or conceptual.

In a book-length essay, Paul Feyerabend (1993, p. 9) argued in a similar vein as Laudan, challenging the idea that every paradigm or research programme consists of a set of fixed rules followed blindly by the researchers invested. "Science is an essentially anarchic enterprise: theoretical anarchism is more… likely to encourage progress than its law-and-order alternatives" (ibid., p. 5). Any theoretical boundaries would not be able to create the diversity and complexity witnessed from the history of science (ibid., p 10)

Feyerabend had two convincing arguments for stressing the importance of "anarchistic science". First, science is about discovering the unknown, and even the theories we believe to be true may be false as had Popper argued (1993, p. 12). Settling for a truth indicates to stop

looking for alternatives, and in effect will impede further progress. Secondly, researchers are individuals with their own special ways of understanding and exploring the world they live in (ibid., p. 12). Restraining the researchers potentials can neither provide scientific development nor personal development, and progress would cease accordingly. The history of science does not entail monotonous thinking, Feyerabend assert, it is full of accidental discoveries. "The only principle that does not inhibit progress is: anything goes" (ibid., p. 14).

2.3. Conclusions

This chapter has provided an in depth discussion on the most influential theories on scientific progress. Two principles are found throughout the works of the different theorists: problem-solving and pluralism. These principles seem to have two distinct, though interconnected, functions for the progress of science: what essentially drives scientific progress forward is the problem-solving abilities of the respective framework, and these problem-solving abilities are contingent upon competing ideas and diversity, in other words, pluralism.

3. On the prevalence of accessibility related biases

"A representative sample of all experimental findings must be publicly accessible"

3.1. Introduction

As the previous chapter provided a comprehensive discussion concerning which principles drive scientific progress forward, the current chapter will focus on what may impede or inhibit these principles of being fulfilled, stated in research question (3).

Ioannidis (2005) argued that the biggest issues within the medical and behavioural sciences today are accessibility related biases (ibid. p. 696), a concern shared with many before him (Bakan, 1966; Rosenthal, 1979; Smart, 1964; Sterling, 1959) and after (Cumming, 2014; Song et al., 2010). More specifically, several distinct biases may be identified: positive results bias (Song et al., p. 3); selective reporting bias (Song et al., p. 21); non-replication bias (Cumming, 2014, p. 4); as well as some less influential biases (Song et al., p. 24-37). Though the biases will be treated individually in this chapter, it is essential to recognize that positive results bias provides incentives for the latter biases (Cumming, 2014, p. 2).

The chapter begins with a discussion concerning the file drawer problem in section 2. Section 3 introduced the accessibility related biases individually, focusing mainly on the three previously mentioned biases, before providing a brief review on some less influential biases. Section 4 examines the funnel plot, the most common method to assess the existence of biases in meta-analyses, while section 5 concludes the current chapter.

3.2. The file drawer problem

After Sterling (1959, p. 33) observed that studies that rejected the null hypothesis had a greater probability of being published, and that few, if any, journals published replication studies, he came to the conclusion that surely more experiments had been conducted than those being published. Rosenthal (1979) expanded this idea further, and presented the file drawer problem. Introduced as a worst case scenario, the file drawer problem is the theoretical possibility that the "journals are filled with the 5% of the studies that show Type I errors, while the file drawers back at the lab are filled with the 95% of the studies that show non-significant (e.g., $p > .05$) results" (1979, p. 638).

Though significance testing will be more thoroughly discussed in section 4.2.1. it is important to grasp the essentials before reading on. Significance levels are decided before the experiment is conducted, and is commonly set to .05 or .01 within the behavioural sciences (Cohen, 1994, p. 999), indicating the probability that an experiment will yield a false positive, a Type I error (Kline, 2003, p. 27). Given a significance level of .05, 1 out of 20 studies will yield statistically significant results even in the case of no true effect, implying that the file drawer problem is not an unthinkable dystopia; rather it's an uncomfortably realistic scenario.

Concerning scientific progress, the existence of a file drawer problem may essentially obstruct scientific heterogeneity, consequentially impeding pluralism and thus the problem-solving abilities of psychological research. As such, the inferential proposition (ii), between accessibility related biases and scientific stagnation (if *q*, then *r*) may accordingly be seen as supported given the existence of a file drawer problem within psychology.

3.3. Accessibility related biases

3.3.1. Positive results bias

The most common term of positive results bias stems from Dickersin (1990), defined as "the tendency on the parts of investigators, reviewers, and editors to submit or accept manuscripts for publication based on the direction or strength of the study findings" (ibid., abstract). A similar definition is found in Song et al. (2010) in which positive results bias "occurs when authors are more likely to submit, or editors accept, positive than null results" (ibid., p. 3). Both terms sufficiently accounts for positive results bias, however, whereas the former term provide a more broad definition, the latter clearly emphasizes the neglect of null results, and hence endorse more directly the subject matter.

One of the first systematic studies on positive results bias within the psychological sciences was conducted by Sterling (1959). He carefully reviewed all articles published in 1955-56 from four randomly selected journals in four major areas of psychology, and found that over 97% (n=362) of the articles that used significance testing found positive results, that is, they rejected the null hypothesis at $p \leq .05$ (ibid., p. 31). Sterling suggested, as Rosenthal (1979), that the published results surely were not representative of all the experimental results from the numerous psychology labs (Sterling, 1959, p. 34). Smart (1964) wanted to explore further on Sterling's findings by comparing the percentage of positive results in published papers to those found in abstracts presented at APA annual meetings and PhD dissertations.

He found that while close to 91% (*n*=309) of the published studies were positive, only around 75% (*n*=169) of the APA abstracts and PhD dissertations were positive (ibid., p. 226). Smart emphasized that "unless the scientist is aware of *all* experimental tests performed for a certain hypothesis then a rational decision as to warrantability of the hypothesis cannot be reached" (ibid., p. 228). 30 years after Sterling's first study, Sterling, Rosenbaum, and Weinkam (1995) repeated the study in hope of finding some improvements. Reviewing nine academic journals within the same four areas of psychology, Sterling et al. could confirm that little had changed: still close to 96% (*n*=597) of the articles yielded positive results. Several other studies within psychology corroborate the previous findings, as seen in Table 1.

Table 1
Proportion of published studies with significant results

Study	Sources	% 'positive' results (*n*)
Sterling (1959, p. 31)	*Experimental psychology (1955); Comparative and Physiological Psychology (1956); Clinical Psychology (1955); Social Psychology (1955)*	97% (362)
Smart (1964, p. 34)	*Experimental Psychology (1962, Vol. 63); Comparative and Physiological Psychology (1962, Nos. 1, 2, 3); Clinical Psychology (1961); Social Psychology (1962)*	91% (309)
Smart (1964, p. 34)	*APA Annual Meeting (1962); Dissertation Abstracts (1962)*	75% (169)
Bozart & Roberts (1972, p. 775)	*Journal of Consulting and Clinical Psychology (Jan.1967-Aug. 1970); Journal of Counseling Psychology (Jan.1967-Aug. 1970); Personnel and Guidance Journal (Jan.1967-Aug. 1970).*	94% (1334)
Greenwald (1975, p. 11)	*Journal of Personality and Social Psychology (1972)*	88% (199)
Sterling, Rosenbaum, & Weinkam (1995, p. 109)	*General Psychology (1987); Learning, Memory and Cognition (1987); Human Perception and Performance (1987); Animal Behavior Processes (1986); Behavioral Neuroscience (1987); Comparative Psychology (1987); Journal of Consulting and Clinical Psychology (1987); Journal of Personality and Social Psychology (1987).*	96% (597)
Fanelli (2010, p. 3)	Randomly selected articles	91% (141)

The brief review of the positive publication bias should be disturbing to read for most psychologists. Only one study found that less than 90% of the articles showed positive results, and most were well above. When discussing scientific methodology, it is easy to forget that

psychology is also a branch of the medical sciences. Most of the studied included journals from clinical and consulting psychology, and it's frightening that in the struggle of statistical significance, the prevalence of distortions may "results in an overestimation of treatment effects or an underestimation of adverse effects" (Song et al., 2010, p. 81).

3.3.2. Selective reporting bias

Selective reporting bias "occurs when studies with multiple outcomes report only some of the outcomes measured and the selection of an outcome for reporting is associated with the statistical significance or importance of the results" (Song et al., 2010, p. 21). Phillips (2004) identifies three distinct ways selective reporting bias commonly occurs: in the selection of outcome variables; in the selection of representation of variables; and in the selection of assessments tools.

In one of the most direct studies on selective reporting bias, Chan, Krleža-Jerić, Schmid, and Altman (2004, p. 735) investigated the prevalence of selective reporting bias, the correlation between selective reporting bias and statistical significance, and the consistency between specified trials protocols and specified trials in 68 published articles in Canada. The results supported the assumptions: 88% of the studies had a minimum of one unreported outcome; statistically significant results were more often published; and 40% of the studies consisted of major discrepancies between the protocols and the articles (ibid., p. 737-738). Similar results were found in Denmark (Chan, Hróbjartsson, Haahr, Gøtzsche, & Altman, 2007, p. 2460-2462), were 71% had omitted at least on outcome, the prevalence of positive results bias were confirmed, and 62% consisted of major discrepancies (n=122).

Cumming and others (Cumming, 2014; De Angelis et al., 2004; Ioannidis, 2005) have argued that one way of dealing with selective reporting bias is prespecified research designs. By registering "the procedure, selection of participants, sample sizes, measures, and statistical analyses" (Cumming, 2014, p. 4) in advance, all deviations from the original proposal must necessarily be documented and made clear in the final publication. In the event that a study does not get published, De Angelis et al. (2004, p. 607) argues that access to the research material is still essential and, though not published, may be included in future meta-analyses. However, selective reporting bias is still a consequence of positive results bias, and as such should not be confused as a resolved issue.

3.3.3. Non-replication bias

In this thesis, non-replication bias occurs when authors are less likely to submit, or editors accept, replication studies independent of experimental results (Cumming, 2014, p. 4). There are at least two important roles of replication studies within modern scientific research. First, as Sterling (1959) and Rosenthal (1979) stressed, replications ensure that over time false positive studies are refuted, and thus protect science from maintaining null-fields of study, i.e., "fields with absolutely no yield of true scientific information" (Ioannidis, 2005, p. 700). An equally important task is that a single study rarely provides definitive answers (Cumming, 2014, p. 4; Ioannidis, 2005, p. 696), and only in the presence of several replications, or close replications, may a meta-analysis provide more precise conclusions.

In 2011, Bem (2011, p. 421) published a series of experiments providing statistical significant evidence for precognition in a high impact journal. Precognition as Bem explained it is the "conscious cognitive awareness… of a future event that could not otherwise be anticipated through any known inferential process" (ibid., p. 407). Obviously articles that provide support of precognition are reviewed thoroughly by any editorial board before being accepted, but the article was excellent by most methodological standards, and thus attracted massive media coverage (Aldhous, 2010; Macrae, 2010). However, when Ritchie, Wiseman, and French (2012) conducted a study replicating one of Bem's most impressive findings, they failed to achieve statistical significance in three independent experiments identical to that of Bem. Eager to report the replication study, an article was written and sent to the same journal that presented Bem's original paper (Aldhous, 2011). Yet, to the researchers surprise, the editor of the journal declined the submitted paper and wrote back: "This journal does not publish replication studies, whether successful or unsuccessful". Luckily, their article ended up being published in *PLoS ONE* and received the deserved attention.

Perhaps the most efficient way of dealing with the non-replication bias is through an open-access database of replication findings (Schooler, 2001, p. 437), or journals dedicated at publishing pure replications. *Perspectives on Psychological Science* has just confirmed that they will begin publication of replication studies, in a process actively involving the authors of the original studies (Association for Psychological Science, n.d.). By doing so, the aim is to provide incentives to conduct replication studies, as well as making sure that the outcome of a replication study should be taken seriously.

3.3.4. Less influential biases

In their extensive review of accessibility related biases, Song et al. (2010) listed in all nine more biases that they conclude have a limited effect on the overall distortion of scientific research findings. These biases will not be covered in detail, but as they are in conjunction to the positive results bias, they will be briefly described before moving on to funnel plots.

According to Song et al. (2010, p. 24), time lag bias constitutes two similar biases: when the time to publication is associated with the study results, as well as the changes in effect size over time. Perhaps the most direct support of the former time lag bias is found by Simes (1987, p. 20), were he compared the time and place of 38 published trials. He found that all of the studies yielding statistical significant results were published in high impact journals within 5 years of trial closure, while studies with non-significant results on the other hand, were published in less prestigious journals, and some never got published at all (ibid., p. 21). Examples of changes in effect size over time can be found throughout the literature, for instance Gehr, Weiss, and Porzsolt (2006) found that the effect size reported for three out of the four medical therapies they studied decreased over time.

Grey literature bias occurs when published results substantially differ from "those presented in reports, working papers, dissertations or conference abstracts" (Song, Eastwood, Gilbody, Duley, & Sutton, 2000, p. 3). The effects of including grey literature may have great influence on meta-analyses, as McAuley, Pham, Tugwell, and Moher (2000, p. 1229) found when including grey literature in 41 studies, the overall treatment effect fluctuated ± 10%.

Egger et al. (1997, p. 328) found support of a language bias when comparing RCTs published in English- and German-language journals. Articles published in English were almost twice as likely to report statistical significant results than those published in German. Similar results were reported in a study comparing trials within neuroscience, were 33% of the German studies as to 57% of the English studies showed statistically significant results (Heres, Wagenpfeil, Hamann, Kissling, and Leucht, 2004, p. 231).

Citation bias "occurs when the probability that a study will be cited is associated with the study results" (Song et al., p. 31). In support of the existence of citation bias, Kjaergard and Gluud (2002, p. 408) found that the strongest predictor of citation frequency was a statistical significant result. Furthermore, Nieminen, Rucker, Miettunen, Carpenter, and Schumacher (2007, p. 941) found that over a ten-year period, articles with significant results were over twice as often cited as those with non-significant results.

Song et al. (2010, p. 3) included duplicate bias, place of publication bias, country bias, indexing bias, and media attention bias as well. However, as they concluded that "limited evidence" in the literature supported these biases, these will not be dealt with any further.

3.4. Funnel plots

Several authors (Duval & Tweedie, 2000; Song et al., 2010; Sterne & Egger, 2001) have argued that accessibility related biases may be dealt with when computing meta-analyses by appealing to funnel plots, as introduced by Light and Pillemer (1984) and popularised by Egger, Smith, Schneider, and Minder (1997). Briefly described, the funnel plot assigns the individual studies in a scatter plot (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006, p. 597), typically with treatment effects on the horizontal axis and for instance sample size on the vertical axis. The assumption is that "if all studies come from a single underlying population, this graph should look like a funnel, with the effect sizes homing in on the true underlying value as *n* increases" (Light & Pillemer, 1984, p. 63, italics added). In the occurrence of accessibility related biases, a slice of the funnel should be missing. However, Lau et al. (2006, p. 598) argues that the funnel plots accuracy is strongly affected by several factors, such as coding of outcome, choice of metric, and choice of weight on the vertical axis. Furthermore, as funnel plots cannot provide an accurate picture of accessibility related biases, the authors argue that preventing these biases of occurring in the first place should rather be the issue.

3.5. Conclusions

Having established the severity of potential accessibility related biases, perhaps most clearly by the file drawer problem, as well as finding support of its prevalence within psychological literature, it may be asserted that the inferential proposition between accessibility related biases and scientific stagnation seem indeed to be true.

4. On the fallacies of significance testing

"Significance testing is the underlying cause of accessibility related biases"

4.1. Introduction

Having established that the principle of scientific progress may be impeded by accessibility related biases, this chapter will provide an extensive analysis in order to determine whether the aforementioned principles are fulfilled in the psychological sciences, as stated in research question (4). The outcome of the analysis should provide sufficient information to conclude the main research question (1).

Ioannidis (2005) claim that null hypothesis significance testing is the underlying cause of the current accessibility related biases is perhaps the most serious allegation against significance testing. In addition, this methodological dogma aids several conjunctive fallacies, such as misinterpretation of the *p*-value (Carver, 1978; Cohen, 1994; Falk & Greenbaum, 1995; Kirk, 2003; Nickerson, 2000), cultivation of dichotomous thinking of nuanced data (Cumming, 2014; Rosenthal & Gaito, 1963, 1964), and the problem of practical significance (Bakan, 1966; Cohen, 1962; Johnson, 1999; Thompson, 1998; Tukey, 1991). In the light these arguments, null hypothesis significance testing may perhaps pose the most serious challenge towards ensuring further progress within psychology today.

The chapter is organized as follows. Section 2 discusses the basics of null hypothesis significance testing as well as the historical development and the current practice within the psychological sciences. Following in section 3 is a discussion concerning significance testing and accessibility related biases. In addition, in section 4, the main conjunctive fallacies are the examined. Concluding the chapter is section 5.

4.2. Null hypothesis significance testing

4.2.1. Null hypothesis significance testing

In the psychological literature, as well as in the medical and other behavioural sciences, null hypothesis significance testing is assessed as a technique of comparing groups. Significance testing helps examine whether some independent variable has an effect on the dependent variable by comparing the observed difference between the populations under the different

conditions (Loftus, 1996, p. 161). Commonly, the hypothesis the investigator set out to explore, called the alternative hypothesis (H_1 or H_A), entails that there *is* an effect and thus the population means should differ ($\mu_1 \neq \mu_2 \neq \ldots \mu_n$)[20], while the null hypothesis (H_0) on the contrary, entails that there are *no* effect and the population means will not differ ($\mu_1 = \mu_2 = \ldots \mu_n$) (Cumming, 2012, p. 21)[21]. By means of deductive refutation, significance testing aims at refutation of the H_0 and thereby confirming the alternative hypothesis, H_1.

After conducting the experiment, the investigator will calculate the *p*-value, indicating the probability of getting the observed effect or more extreme given the null hypothesis is true (Cumming, 2012, p. 27). Thus, the lower the *p*-value, the more extreme the data has to be in order to represent the true null hypothesis. The next step is to compare the calculated *p*-value to the significance level. Decided in advance of the experiment, the significance level (1-α) indicates at which point the investigator should accept or reject the null hypothesis. In psychology, the significance level is usually set to .05 or .01 (Cohen, 1994, p. 999), implying that given H_0 is true, only 5% or less of these results can be even more inconsistent than the observed results (Kline, 2004, p. 63). If the *p*-value is less or equal to the significance level ($p \leq .05$), the null hypothesis is rejected, and the alternative hypothesis is accepted.

4.2.2. Historical development

The approach to the null hypothesis significance testing as assessed within psychology today is in fact a hybrid of two competing approaches developed individually by Ronald Fisher (1925) on the one hand and Jerzy Neyman and Egon S. Pearson (1933) on the other (Kline, 2004, p. 6; Sterne & Smith, 2001, p. 226).

Fisher's model is sometimes called the *p*-value approach (Kline, 2004, p. 7), because it introduced the *p*-value as measure to establish "the strength of evidence against the null hypothesis" (Sterne & Smith, 2001, p. 226). As Popper had argued, the only way to ascertain true knowledge is by deductive refutation, which captures the essence of the *p*-value approach. As such, Fisher never advocated for an alternative hypothesis (Kline, 2004, p. 7), as

[20] The alternative hypothesis may be nondirectional or directional. Nondirectional hypothesis does not specify the direction of the effect, positive or negative, while a directional hypothesis specify the excepted direction of the effect, for instance $\mu_1 < \mu_2$ (Kline, 2004, p. 37).

[21] Cohen (1994, p. 1000) separated two types of H_0: the nil hypothesis and the non-nil hypothesis. The former is the one described above which assumes $\mu_1 = \mu_2 = \ldots \mu_n$. Non-nil hypotheses on the other hand assumes that there is a difference between the populations, and may be given a value not equal to cero, for instance $\mu_1 - \mu_2 = 10$. As nil hypotheses is the most common is psychology (Cohen, 1994, p. 1002; Kline, 2004, p. 37), non-nil hypotheses will no be dealt with further.

the falsity of the null hypothesis does not entail the truth of the alternative. The standard of rejecting the null hypothesis at $p < .05$ comes from Fisher's work, though he emphasised that judging the *p*-value should be the task of the researcher (ibid., p. 7), implying that every experiment should be judged individually.

The Neyman-Pearson model, called by some the fixed-*p* approach (Kline, 2004, p. 7), attempted to rid of the subjective character of the significance test (Sterne & Smith, 2001, p. 227). The perhaps greatest advantage of their model was the introduction of statistical power. In the Fisher approach, only Type I error were accounted for, while the Neyman-Pearson model included adjustments for Type II error as well (ibid., p. 227). Proposing an alternative hypothesis, prespecified rejection regions, and a fixed level of confidence across studies, the Neyman-Pearson model provided a seemingly more objective approach to significance testing than Fisher's initially did (Kline, 2004, p. 7).

However, as Sterne and Smith (2001, p. 227) argues, researchers rarely specify a concise alternative hypothesis or even perform the appropriate power calculation to find the number of participants needed in advance. Rather they rely on what Gigerenzer (1993, p. 312) termed the "hybrid logic" of statistical inference, claiming conclusive findings based on the *p*-value with little consideration of the Type II error rate (Sterne & Smith, 2001, p. 227). This hybrid logic has become an unfortunate dogma within psychological research today, and has several unfortunate consequences discussed in section 4.3.

4.2.3. Statistical tests in psychological research

Three of the most widely used statistical tests within psychology today are the *t* test, the *F* test, and the chi-square (Kline, 2004, p. 40). Though the tests are applied for different means, the outcome is the same; a summarized result accompanied a sample statistic. "The difference between the statistic and the value of the corresponding population parameter specified in the null hypothesis is compared against an estimate of sampling error" (ibid., p. 40), and then converted into probabilities. The next step is to compare the *p*-value to the prespecified α, and dependent on the value of *p*, the null hypothesis will either be rejected or accepted.

4.3. Significance testing and accessibility related biases

Having established in section 4.2 that the most common statistical tests within psychology are all dependent on significance testing, the current section is concerned with Ioannidis' alleged inferential proposition (i) concerning significance testing and accessibility related biases.

Ioannidis (2005, p. 696) argued that the file drawer problem presented by Rosenthal (1979) might in actuality be even more serious. Introduced previously as a scenario were only the 5% studies showing Type I errors are published, Ioannidis call attention to the fact that the file drawer problem only takes statistical significance into account. However, the "probability that a research finding is indeed true depends on the prior probability of it being true, the statistical power of the study, and the level of significance" (Ioannidis, 2005, p. 696).

More precisely, it depends on the prior probability of a relationship being true (R/R-1), were R is the ratio of "actual relationship" to "no relationships", the probability of finding a true relationship (1-β), and the probability of falsely claiming a relationship (α) (Ioannidis, 2005, p. 696). Given that a certain amount (c) of relationships are being examined in a field, the post-study probability that a research finding is indeed true after achieving statistical significance is the false positive report probability (Wacholder, Chanock, Garcia-Closas, El ghormli, & Rothman, 2004). Ioannidis provide the following 2x2 table of the expected values (Table 2), which indicate that the false positive report probability is $(1 - \beta)R/(R - \beta R + \alpha)$, and that a research finding is more likely true than false, if $(1 - \beta)R > \alpha$.

Table 2
Research findings and true relationships (Ioannidis, 2005, p. 697)

Research finding	True relationships		
	Yes	*No*	Total
Positive	$c(1 - \beta)R/(R + 1)$	$c\alpha/(R + 1)$	$c(R + \alpha - \beta R)/(R + 1)$
Negative	$c\beta R/(R + 1)$	$c(1 - \alpha)/(R + 1)$	$c(1 - \alpha + \beta R)/(R + 1)$
Total	$cR/(R + 1)$	$c/(R + 1)$	c

Admittedly, this looks rather complicated, but a simple illustration can make it far more comprehensible. As stated previously, there are really only three factors that need to be determined; the prior probability, the statistical power, and the significance level. For instance if 1000 hypotheses are being tested, and the prior probability is .1, calculations would indicate that 100 hypotheses are indeed true. With a statistical power of .5 as well as a .05 significance level, both common levels in psychological research, 50 of the true relationship will be falsely

rejected while 45 false relationships will be accepted, as seen in Table 3. Calculating the false positive report probability yields 0.526, and being greater than the significance level (.05), a statistically significant research finding is more likely to be true than false.

Table 3
Research findings and true relationships with α (.05) and β (.5)

Research finding	True relationships		
	Yes	*No*	Total
Yes	50	45	95
No	50	855	905
Total	100	900	1000

Furthermore, Ioannidis introduces equations accounting for the proportion of research findings caused by accessibility related biases (*u*), as these biases may lower the false positive report probability substantially (2005, p. 697). These calculations will not be discussed in detail, as it should be evident from the current analysis, that adding yet another dimension of error into these equations will make the false positive report probability plummet.

However, it is worth noticing that Ioannidis argues that the underlying cause of the current problem is the "convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance" (2005, p. 696). Null hypothesis significance testing seem to set the stage for accessibility related biases, by providing an approach for analyzing and interpreting data which presents seemingly objective, clear-cut answers in the form of significant or non-significant.

4.4. General criticism

4.4.1. Misinterpretation of the *p*-value

A common mistake is to confuse scientific inference and null hypothesis significance testing (Carver, 1978, p. 382; Kirk, 2003, p. 85). With scientific inference, the intention is to provide the probability (p) that the null hypothesis (H_0) is true given a set of data (D), i.e., $p\ (H_0|D)$ (Cohen, 1994, p. 998). Applying significance testing, however, provides the probability of collecting these data or more extreme given the null hypothesis is true, i.e., $p\ (H_0|D)$ (ibid., p. 998). Misunderstanding this very important distinction, has several conjunctive fallacies, most

commonly the belief that a low *p*-value indicates that the null hypothesis is false, famously called the "illusion of probabilistic proof by contradiction" by Falk and Greenbaum (1995)[22].

Nickerson (2000) distinguished several other common fallacies based on the illusion of probabilistic proof by contradiction, such as the belief that the *p*-value is the probability that H_0 is true (ibid., p 246); that 1-*p* is the probability that the H_A is true (ibid., p. 246); that a small *p*-value indicates that results are replicable (ibid., p. 256); that a small *p*-value is indicative of larger effects (ibid., p. 257); as well as the belief that statistical significance means practical significance (ibid., p. 257). Nickerson (2000, p. 289) concludes that null hypothesis significance testing has more severe adverse than positive effects on scientific conclusions, and that the best solution is to avoid significance testing all together.

4.4.2. Making dichotomous conclusions from nuanced data

In a series of studies exploring the way psychologists interpret *p*-values, Rosenthal and Gaito (1963, 1964) found that the psychologists dramatically lost confidence in the research findings when the *p*-values exceeded .05. At the centre of this "cliff effect" is the arbitrarily predetermined significance level of .05, turning a rich and nuanced collection of data into a dichotomous reject-nonreject decision (Cumming, 2014, p. 5; Kirk, 2003, p. 87), commonly only reported as a statistically significant or non-significant result. By realizing that *p*-values barely above the significance threshold are treated similarly to those of much greater value should certainly provide convincing support that data is not properly represented by null hypothesis significance testing (Kirk, 2003, p. 87).

4.4.3. Statistical significance and practical significance

Cohen (1962, p. 145) were concerned that psychologists were only attentive to the Type I errors (i.e., falsely rejecting the null hypothesis), while seemingly uninterested in the Type II errors. The issue, as Tukey (1991, p. 100) argued, is that all null hypotheses are necessarily

[22] Carver (1978, p. 383) wrote a clever analogy on the "illusion of probabilistic proof by contradiction". Here it is have shortened it down slightly, but the essence remains the same:

> What is the probability of obtaining a dead person (*D*) given that the person was hanged (*H*); i.e., what is *p* (*D*|*H*)? Obviously it will be very high, perhaps 0.97 or higher. Now, let us reverse the question. What is the probability that a person has been hanged (*H*), given that the person is dead (*D*); that is, what is *p* (*H*|*D*)? This time the probability will undoubtedly be very low, perhaps 0.01 or lower. No one would be likely to make the mistake of substituting the first estimate (0.97) for the second (0.01); that is to accept 0.97 as the probability that a person has been hanged given that the person is dead.

false on a certain level. It is close to impossible even imagining a situation were a population parameter do not differ, that is where $\mu_1 = \mu_2$. An obvious corollary is that Type I errors frequently occur (Bakan, 1966, p. 425; Kirk, 2003, p. 86), while Type II errors rarely happens. In *The Insignificance of Statistical Significance Testing*, Johnson (1999, p. 765) explained that the determining factors of the *p*-value, and thus the determination of statistical significance, are a function of the difference between the population parameter and the null hypothesis, as well as the sample size. As the population parameter is never null, increasing the sample size sufficiently, thus decreasing the standard deviation, will eventually provide the researcher with a sample parameter that yield statistical significant results (Thompson, 1998, p. 799). Cohen (1962, p. 147) argued further that due to this inconvenient relationship, statistical significance is practically useless, and cannot provide any indications on questions concerning practical significance.

4.4.4. Flawed support of significance testing

Schmidt and Hunter (1997) provided a list of some common, but according to them, flawed arguments in support of significance testing. These arguments, and the counterarguments that follow with, will be discussed in this section, as they often come up in discussion on the topic.

The argument that only significance testing can judge whether a finding is real or due to chance, is perhaps the most common objection against disposing of significance testing (Schmidt & Hunter, 1997, p. 39). This argument is based on a misunderstanding of statistical power and relates to the Type II errors. The common statistical power in psychological research centres somewhere around .50 (Cohen, 1962, p. 148, 1990, p. 1304; Schmidt, Hunter, & Urry, 1976, p. 474), which indicates that theoretically half of the studies will be non-significant even in the case of a true effect. Thus, the statistical power only reflects that half of these studies will yield false conclusions over time. Conversely, if there is no true effect, the significance level will equal the alpha level (Schmidt & Hunter, 1997, p. 40), indicating a relatively low error rate. However, as argued in section 4.*, all null hypotheses are usually false, indicating that significance testing can in fact never conclusively determine the truth of any single experimental finding.

Another argument holds that hypothesis testing is impossible without significance testing (Schmidt & Hunter, 1997, p. 42), indicating that no proper alternative to the null hypothesis exists. By comparing the statistical approaches in the “hard sciences” with those of the “soft sciences”, Hedges (1987, p. 452) found that the “hard sciences” had achieved greater

success mainly due to meta-analytical approaches, comparable to those of point estimates and confidence intervals within the behavioural sciences (Schmidt & Hunter, 1997, p. 43).

A frequent argument in support of significance testing is that the problem is not the inadequacies of significance testing; it is the lack of replication (Schmidt & Hunter, 1997, p. 44). Again, the problem is connected to statistical power, as reproducibility with a power of .50, still would imply that only 50% of all replications will be successful. In a series of studies, the statistical power must necessarily be multiplied, implying that the probability of finding significant results when an effect truly exists in only three consecutive studies is as low as 12.5% (ibid., p. 45). It is also worth noting that this is argument usually refers only the findings yielding significant results. However, replicating non-significant results is generally as difficult, as the probability of replicating a non-significant result is commonly 1 - statistical power (ibid., p. 45), which, as earlier mentioned, would imply 1 - .50, giving the same probability as in the previous example.

Schmidt and Hunter (1997, p. 56) also mention the argument that the problem is not in significance testing; rather it is the misuse of them. This argument may hold some truth, but as Rosenthal and Gaito (1963, 1964) argued, even trained professionals make simple mistakes when dealing with significance testing. In contrast, argues Schmidt and Hunter further, "point estimates of effect sizes and their associated confidence intervals are much easier to… understand" (1997, p. 57). Coulson, Healey, Fidler, and Cumming (2010, p. 5) extended the support of Schmidt and Hunter, when they found far more accurate interpretations of study results presented as confidence intervals, than that of significance testing.

4.5. Conclusions

The most common approach for analyzing and interpreting data within psychology is null hypothesis significance testing, an approach seemingly overwhelmed by controversies and criticism. Ioannidis' claim that significance testing serve to uphold accessibility related biases seem by further investigation to hold true, and in the light of the several conjunctive fallacies of significance testing, the most appropriate response may be to abandon this methodological dogma altogether.

5. On the future of psychology

"Change is of the essence"

5.1. A brief summary

The aim of this thesis has been to explore the recent discussion concerning the inadequacies of the statistical models to analyze and interpret data in the medical and behavioural sciences, in particular the suitability of null hypothesis significance testing to ensure further progress in psychology. More specifically, four research questions were formulated in section 1.3:

(1) *Can null hypothesis significance testing ensure further progress within psychology?*
(2) *Which principle(s) drive scientific progress forward?*
(3) *What may impede or inhibit these principle(s) of being fulfilled?*
(4) *Are the aforementioned principle(s) fulfilled in the psychological sciences?*

Indications from the philosophy of science concerning the nature of scientific progress, led to the proposal that continuous problem-solving is the driving force of progress (section 2.3), and that problem-solving is contingent upon pluralism, i.e., diversity of competing ideas (section 2.3). Furthermore, to ensure pluralism it was suggested that a representative sample of all experimental findings must be publicly available (section 3.2), and that the prevalence of accessibility related biases may potentially obstruct theoretical and empirical diversity, ultimately restraining scientific progress (section 3.4). Having established that null hypothesis significance testing is the most common approach to analyze and interpret data in psychology (section 4.2), an extensive review of Ioannidis' claim concerning significance testing and the accessibility related biases was provided (section 4.3). In addition, some conjunctive fallacies challenged the suitability of the approach, for instance that *p*-values are misinterpreted, that it cultivates dichotomous thinking of nuanced data, and that statistical significance has little practical value (section 4.4).

5.2. Evaluating the validity of the argument

It was claimed in section 1.2.2, that the validity of the often-stated argument that significance testing may impede further progress was dependent on several propositions, only some of which have actually been examined. Furthermore, it was argued that until the philosophical

foundation of the argument has been extensively analyzed and sufficient support may decide its validity, reluctance to accept the proposition is understandable.

The argument was constructed by the pattern of hypothetical syllogism:

(i) If *significance testing,* then *accessibility related biases* (if *p,* then *q*)
(ii) If *accessibility related biases,* then *scientific stagnation* (*if q,* then *r*)
(iii) *Significance testing* (*p*)
(iv) ∴ *scientific stagnation* (*r*)

In order to evaluate the validity of the argument, the truth-value of each of the inferential propositions [(i) and (ii)] and the declarative sentence (iii) need to be determined, and only by affirming each of the propositions may the premises entail the conclusion. In light of the information from research question (2) and (3), it seems evident that the second proposition is supported, as accessibility related biases may inhibit pluralism and thus problem-solving. By exploring the current situation within psychology, significance testing is clearly one of the most common approaches to analyzing and interpreting data, lending support to sentence (iii). Ioannidis perhaps best explained that significance testing indeed support accessibility related by introducing the false positive report probability, and thus affirming the inferential sentence (i). The current analysis can obviously only provide preliminary indications, but in the event that it should be correct, the premises entail the conclusion and the argument is hence valid. Considering the main research question (1), it does not seem that null hypothesis significance testing can ensure further progress within psychology, and minding the conjunctive fallacies of significance testing, this author argues for the abandonment of this methodological dogma in favour of competing methodological substitutes.

5.3. Effect sizes and confidence intervals

This thesis does not focus on alternatives to the significance testing, as providing a thorough review of competing methodological substitutes is beyond the scope of the thesis. However, to illustrate how easily adaptable effect sizes and confidence intervals really are, the next three paragraphs will explain these concepts briefly.

An effect sizes is basically any point of statistical interest, for instance means, frequencies, correlations and so on (Cumming, 2014, p. 9). The sample effect sizes calculated from the collected data, is commonly the point estimate of the population effect size. A point

estimate is thus essentially a descriptive statistic providing the exact location of a sample parameter, indicating the representation of the population parameter (Kirk, 2003, p. 89).

By taking sampling error into account, a confidence interval is added to estimate the level of uncertainty involved with the point estimate (Kline, 2004, p. 27). Steiger and Fouladi (1997, p. 230) provided a quite precise definition of a confidence interval; A 1 - α confidence interval for a certain parameter will, if measured multiple times, include the parameter with a probability of 1 - α. For instance, if α = .05, then the confidence interval will include the given parameter 95% of the times. In other words, the 1 - α confidence interval contains all the values for which H_0 would not be rejected at the α significance level, while all values outside of the confidence intervals would be cause rejection (Kirk, 2003, p. 89).

As Cumming (2014, p. 5) argued, one of the main reasons why effect sizes and confidence intervals supersedes significance testing, is that it provides a precise, nuanced, and easily understandable picture of the collected data. More importantly, by disposing of the significant/non-significant dichotomy, the psychological sciences may be liberated from the current accessibility related biases. Lastly, by adopting effect sizes and confidence intervals, Kirk (2003, p. 100) argues that the transformation from single studies to meta-analyses is far more convenient, and may consequentially lead to a meta-analytic renaissance.

5.4. Change is of the essence

Kirk (2003) might have put it best as he quite aptly wrote: «it is evident that the current practice of focusing exclusively on a dichotomous reject-nonreject decision strategy of null hypothesis [significance] testing can actually impede scientific progress» (ibid., p. 100).

The current methodological dogma of analyzing and interpreting data in the medical and behavioural sciences by means of null hypothesis significance testing is in need of critical revision. The psychological sciences, a discipline pride of its methodological traditions, have a chance to stand in the forefront of a methodological shift. Doing so, psychology will exhibit that it encourages theoretical novelty, advocate continuous progress, and may hence revitalize newfound credibility in the psychological sciences.

References

Aldhous, P. (2010, November 11). Is this evidence we can see the future. *New Scientist*. Retrieved from: http://www.newscientist.com/article/dn19712-is-this-evidence-that-we-can-see-the-future.html#.VAOOhktqcjp.

Aldhous, P. (2011, May 5). Journal rejects studies contradicting precognition. *New Scientist*. Retrieved from: http://www.newscientist.com/article/dn20447-journal-rejects-studies-contradicting-precognition.html#.VAAapUtqcjo.

American Psychological Association (2001). *Publication Manual of the American Psychological Association* (5th edition). Washington, DC: American Psychological Association.

Ariew, R. (1984). The Duhem thesis. *British Journal for the Philosophy of Science*, *35*(4), 313-325.

Registered Replication Reports (n.d.). *Association for Psychological Science.* Retried from: http://www.psychologicalscience.org/index.php/replication.

Ayer, A. J. (1946). *Language, truth and logic*. London, UK: Victor Gollancz Ltd.

Bakan, D. (1966). The test of significance in psychological research. *Psychological Bulletin, 66*(6), 423-437.

Begg, C. B., & Berlin, J. A. (1988). Publication bias: A problem in interpreting medical data. *Journal of the Royal Statistical Society, 151*(3), 419-463.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*(3), 407-425.

Bozarth, J. D., & Roberts, R. R. (1972). Signifying significant significance. *American Psychologist, 27*(8), 774-775.

Brown, S. D., & Tinsley, H. E. A. (2000). *Handbook of applied multivariate statistics and mathematical modeling*. San Diego, CA: Academic Press.

Carnap, R. (1956). *Meaning and necessity: A study in semantics and modal logic*. Chicago, IL: University of Chicago Press.

Carver, R. (1978). The case against statistical significance testing. *Harvard Educational Review, 48*(3), 378-399.

Chan, A. W., Krleža-Jerić, K., Schmid, I., & Altman, D. G. (2004). Outcome reporting bias in randomized trials funded by the Canadian Institutes of Health Research. *Canadian*

Medical Association Journal, 171(7), 735-740.

Chan, A. W., Hróbjartsson, A., Haahr, M. T., Gøtzsche, P. C., & Altman, D. G. (2004). Empirical evidence for selective reporting of outcomes in randomized trials: Comparison of protocols to published articles. *Jama, 291*(20), 2457-2465.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology, 65*(3), 145-153.

Cohen, J. (1990). Things I have learned (so far). *American psychologist, 45*(12), 1304-1312.

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

Copi, I. M. (1961). *Introduction to logic*. New York, NY: Macmillan.

Coulson, M., Healey, M., Fidler, F., & Cumming, G. (2010). Confidence intervals permit, but don't guarantee, better inference than statistical significance testing. *Frontiers in Psychology*, *1*, 26.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York, NY: Routledge.

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

De Angelis, C., Drazen, J. M., Frizelle, F. A., Haug, C., Hoey, J., Horton, R., Kotzin, S., Laine, C., Marusic, A., Overbeke, A. J. P. M., Schroeder, T. V., Sox, H. C., & Weyden, M. B. V. D. (2004). Clinical trial registration: A statement from the International Committee of Medical Journal Editors. *New England Journal of Medicine, 351*(12), 1250-1251.

Delanty, G., & Strydom, P. (2010). *Philosophies of social science: The classic and contemporary readings*. Maidenhead, UK: Open University Press.

Dickersin, K. (1990). The existence of publication bias and risk factors for its occurrence. *Jama, 263*(10), 1385-1389.

Duhem, P. M. M. (1954). *The aim and structure of physical theory*. Princeton, NJ: Princeton University Press.

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal, 315*(7109), 629-634.

Egger, M., Zellweger-Zähner, T., Schneider, M., Junker, C., Lengeler, C., & Antes, G. (1997). Language bias in randomised controlled trials published in English and German. *The Lancet, 350*(9074), 326-329.

Falk, R., & Greenbaum, C. W. (1995). Significance tests die hard: The amazing persistence of

a probabilistic misconception. *Theory and Psychology, 5*(1), 75-98

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

Feyerabend, P. (1993). *Against method: outline of an anarchistic theory of knowledge*. London, UK: New Left Books.

Fisher, R. (1925). *Statistical methods for research workers.* Edinburgh, UK: Oliver and Boyd.

Frege, G. (1950). *The foundations of arithmetic: A logico-mathematical enquiry into the concept of number*. New York, NY: Philosophical Library.

Frick, R. W. (1996). The appropriate use of null hypothesis significance testing. *Psychological Methods, 1*(4), 379-390.

Gehr, B. T., Weiss, C., & Porzsolt, F. (2006). The fading of reported effectiveness. A meta-analysis of randomised controlled trials. *BMC Medical Research Methodology, 6*(25).

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin, 82*(1), 1-20.

Grünbaum, A. (1976a). The duhemian argument. In S. G. Harding (Ed.), *Can theories be refuted?* (116-131). Dordrecht, Holland: D. Reidel Publishing Company.

Grünbaum, A. (1976b). Ad hoc auxiliary hypotheses and falsificationism. *The British Journal for the Philosophy of Science*, *27*(4), 329-362.

Hanson, N. R. (1958). *Patterns of discovery: An inquiry into the conceptual foundations of science*. Cambridge, MA: University Press.

Harre, R. (1994, September 19). Obituary: Professor Sir Karl Popper. *The Independent*. Retrieved from: http://www.independent.co.uk/news/people/obituary-professor-sir-karl-popper-1449760.html.

Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist, 42*(5), 443-455.

Heres, S., Wagenpfeil, S., Hamann, J., Kissling, W., & Leucht, S. (2004). Language bias in neuroscience - Is the Tower of Babel located in Germany?. *European psychiatry, 19*(4), 230-232.

Hesse, M. (1980). *Revolutions and Reconstructions in the Philosophy of Science*. Brighton, UK: Harvester Press.

Hoyningen-Huene, P. (1993). *Reconstructing scientific revolutions: Thomas S. Kuhn's philosophy of science*. Chicago, IL: University of Chicago Press.

Hume, D. (1912). *An enquiry concerning human understanding, and selections from: A treatise of human nature*. Chicago, IL: Open Court Publishing Co.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine, 2*(8), e124.

Jefferys, W. H., & Berger, J. O. (1992). Ockham's razor and bayesian analysis. *American Scientist, 80*(1), 64-72.

Johnson, D. H. (1999). The insignificance of statistical significance testing. *The Journal of Wildlife Management*, *63*(3), 763-772.

Kant, I. (1899). *Critique of pure reason*. New York, NY: Willey Book Co.

Kirk, R. E. (2003). The importance of effect magnitude. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (83-105). Malden, MA: Blackwell Publishing.

Kjaergard, L. L., & Gluud, C. (2002). Citation bias of hepato-biliary randomized clinical trials. *Journal of Clinical Epidemiology, 55*(4), 407-410.

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

Kuhn, T. S. (1970). *The structure of scientific revolutions*. Chicago, IL: University of Chicago Press.

Lakatos, I. (1992). Falsification and the methodology of scientific research programmes. In J. Worrall & G. Currie (Eds.), *The methodology of scientific research programmes* (8-101). Cambridge, UK: University of Cambridge Press.

Lakatos, I., & Zahar, E. (1975). Why did Copernicus' research program supersede Ptolemy's?. In R. S. Westman (Ed.), *The Copernican Achievement* (354-383). Los Angeles, CA: University of California Press.

Lau, J., Ioannidis, J., Terrin, N., Schmid, C. H., & Olkin, I. (2006). The case of the misleading funnel plot. *British Medical Journal, 333*(7568), 597-600.

Laudan, L. (1965). Grünbaum on "the duhemian argument". *Philosophy of Science*, *32*(3/4), 295-299.

Laudan, L. (1996). *Beyond positivism and relativism*. Boulder, CO: Westview Press.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research.* Cambridge, MA: Harvard University Press.

Loftus, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, 5(6), 161-171.

Macrae, F. (2010, November 18). Are humans psychic? Startling new study 'proves' that we can see the future. *Daily Mail*. Retrieved from: http://www.dailymail.co.uk/sciencetech/article-1330596/Humans-psychic-powers-New-

study-proves-future.html.

McAuley, L., Pham, B., Tugwell, P., & Moher, D. (2000). Does the inclusion of grey literature influence estimates of intervention effectiveness reported in meta-analyses?. *The Lancet, 356*(9237), 1228-1231.

Meehl, P. E. (1967). Theory testing in psychology and physics: A methodological paradox. *Psychological Methods, 34*(2), 103-115.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806-834.

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of The Royal Society of London, Series A, 231*(694-706), 289-337.

Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods, 5*(2), 241-301.

Nieminen, P., Rucker, G., Miettunen, J., Carpenter, J., & Schumacher, M. (2007). Statistically significant papers in psychiatry were cited more often than others. *Journal of Clinical Epidemiology, 60*(9), 939-946.

Phillips, C. V. (2004). Publication bias in situ. *BMC Medical Research Methodology, 4*(20).

Popper, K. R. (1965). *Conjectures and refutations: The growth of scientific knowledge*. New York, NY: Harper & Row.

Popper, K. R. (1968). *The logic of scientific discovery*. London, UK: Hutchinson & Co.

Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, *60*(1), 20-42.

Ritchie, S. J., Wiseman R., & French, C. C. (2012). Failing the future: Three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS ONE, 7*(3): e33423.

Rosenthal, R. (1979). The «file drawer problem» and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641.

Rosenthal, R, & Gaito, J. (1963). The interpretation of levels of significance by psychological researchers. *Journal of Psychology, 55*(1), 33-38.

Rosenthal, R, & Gaito, J. (1964). Further evidence for the cliff effect in interpretation of levels of significance. *Psychological Reports, 15*(2), 570.

Russell, B. (1959). *The problems of philosophy*. New York, NY: Oxford University Press.

Salmon, W. C. (1967). *The Foundations of Scientific Inference*. Pittsburgh, PA: University of Pittsburgh Press.

Schlick, M. (1979). The nature of truth in modern logic. In H. L. Mulder, & B. F. B. van de Velde-Schlick (Eds.), *Philosophical Papers: Volume 1* (41-103). Dordrecht, Holland: D. Reidel Publishing Company.

Schmidt, F. L., & Hunter, J. E. (1997). Eight common but false objections to the discontinuation of significance testing in the analysis of research data. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (37-64). Mahwah, NJ: Erlbaum.

Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validation studies. *Journal of Applied Psychology, 61*(4), 473-485.

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature, 470*(7335), 437.

Simes, R. J. (1987). Confronting publication bias: A cohort design for meta-analysis. *Statistics in Medicine*, *6*(1), 11-29.

Smart, R. G. (1964). The importance of negative results in psychological research. *Canadian Psychologist, 5a*(4), 225-232.

Song, F., Parekh, S., Hooper, L., Loke, Y. K., Ryder, J., Sutton, A. J., Hing, C., Kwok, C. S., Pang, C., & Harvey, I. (2010). Dissemination and publication of research findings: An updated review of related biases. *Health Technology Assessment, 14*(8), 1-24.

Steiger, J. H., & Fouladi, R. T. (1997). Noncentrality interval estimation and the evaluation of statistical models. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests?* (221-257). Mahwah, NJ: Erlbaum.

Sterling, T. D. (1959). Publication decisions and their possible effects on inferences drawn from tests of significance - or vice versa. *Journal of the American Statistical Association, 54*(285), 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *The American Statistician, 49*(1), 108-112.

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology, 54*(10), 1046-1055.

Sterne, J. A., Smith, G. D., & Cox, D. R. (2001). Sifting the evidence - What's wrong with significance tests? Another comment on the role of statistical methods. *British Medical Journal, 322*(7280), 226-231.

Thompson, B. (1998). In praise of brilliance: Where that praise really belongs. *American Psychologist, 53*(7), 799-800.

Tomassi, P. (1999). *Logic*. New York, NY: Routledge.

Toulmin, S. E. (1963). *Foresight and Understanding: An Enquiry into the Aims of Science*. New York, NY: Harper & Row.

Tukey, J. W. (1991). The philosophy of multiple comparisons. *Statistical Science, 6*(1), 100-116.

Wacholder, S., Chanock, S., Garcia-Closas, M., El ghormli, L., & Rothman, N. (2004). Assessing the probability that a positive report is false: An approach for molecular epidemiology studies. *Journal of the National Cancer Institute, 96*(6), 434-442.